10 回归分析

Regression Analysis

线性回归结果不能拿来就用

真理太复杂了，除了近似，我们别无他法。

Truth is much too complicated to allow anything but approximations.

—— 约翰·冯·诺伊曼 (John von Neumann) | 美国籍数学家 | 1903 ~ 1957

- 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.anova.anova_lm 获得 ANOVA 表格
- scipy.stats.t.sf() 求解 t 分布的互补累积分布函数 CCDF = 1 - CDF
- scipy.stats.t.ppf() 求解 t 分布的逆累积分布函数
- seaborn.regplot() 绘制回归图像
- 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
- scipy.stats.skew() 计算偏度
- scipy.stats.kurtosis() 计算峰度
- scipy.stats.normaltest() Omnibus 正态检验
- seaborn.distplot() 绘制直方图，叠合 KDE 曲线
- statsmodels.graphics.tsaplots.plot_acf() 绘制自相关结果

10.1 线性回归：一个表格、一条直线

一个表格

大家是否还记得我们在《统计力量》第 24 章给出过图 1 这个表格。这个表格总结的是一个线性回归分析的结果。本章的主要目的就是和大家理解这个表格各项数值的含义。下面首先介绍这个表格具体来自哪个线性回归。

```
                            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 1. 一元线性回归结果

一条直线

图 2 所示为这个一元 OLS 线性回归的自变量、因变量散点数据以及分布特征。自变量为一段时间内标普 500 股票指数日收益率，因变量为某只特定股票的同期日收益率。观察散点图，我们可以发现明显的"线性"关系。

从金融角度，股指可以"解释"同一个市场上股票的涨跌。再次强调，线性回归不代表"因果关系"。图 1 是利用 statsmodels.api.OLS() 函数构造的线性模型结果。

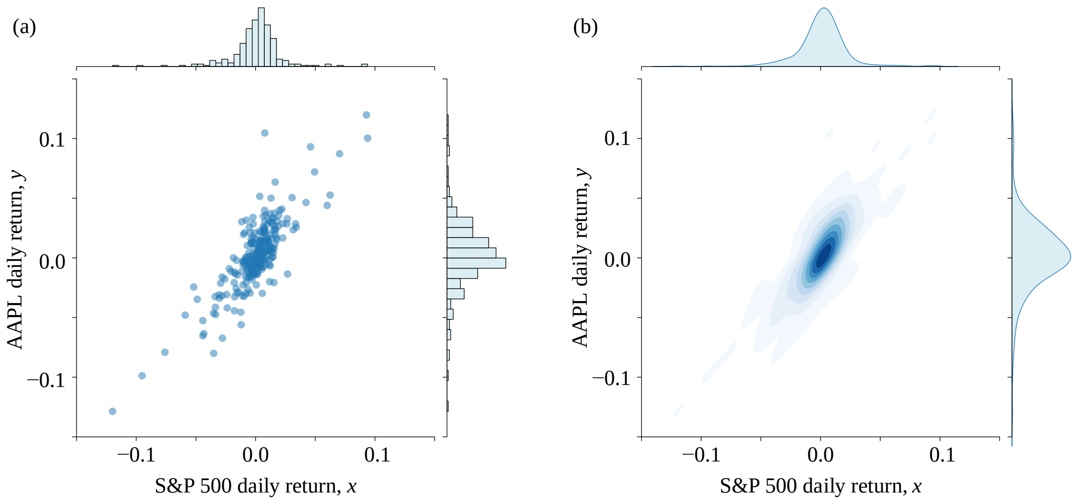


图 2. 日收益率数据关系

图 3 所示为用用 seaborn.jointplot() 绘制回归图，并且绘制边际分布。

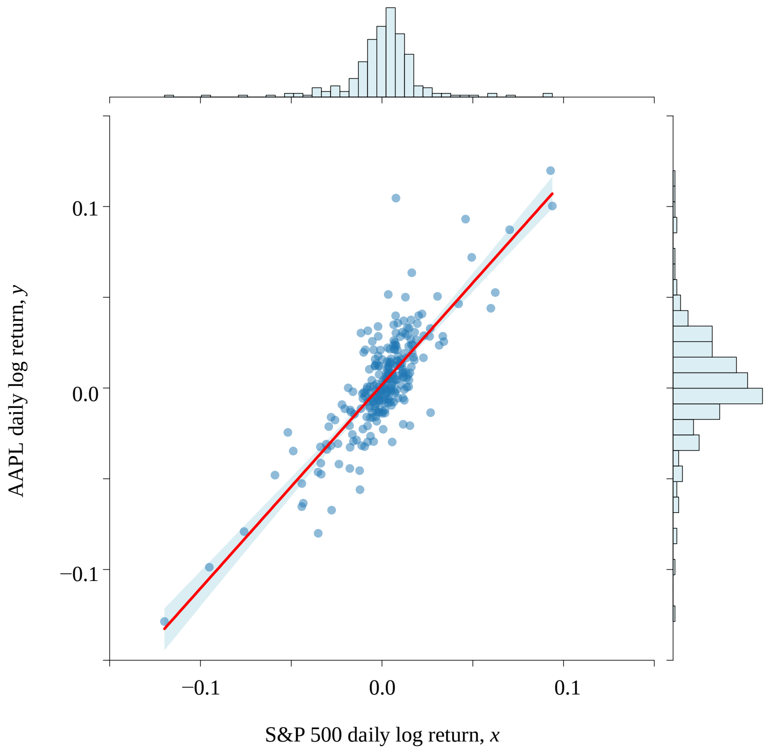


图 3. 用 seaborn.jointplot() 绘制回归直线

统计特征

图 4 (a) 所示为数据的协方差矩阵。《统计至简》第 12、24 章介绍过如何从条件概率角度理解线性回归。假设 X 和 Y 的均值为 0，请大家根据这个协方差矩阵写出线性回归解析式。

图 4 (b) 所示为相关性系数矩阵热图。《矩阵力量》第 23 章介绍过相关性系数可以看成是“标准差向量”之间夹角，具体如图 4 (c) 所示。

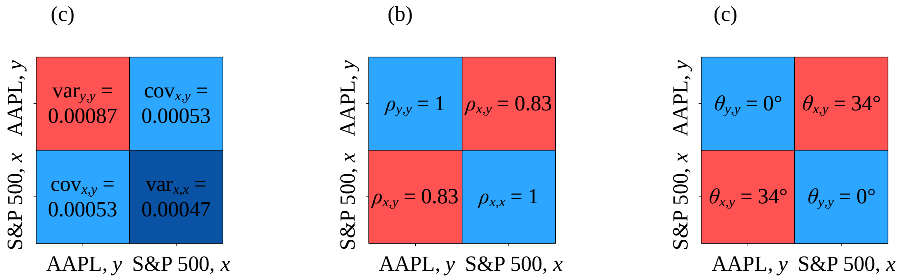


图 4. [**y**, **x**] 数据的协方差矩阵、相关性和夹角热图

图 5 所示为两个标准差向量的箭头图。夹角越小，说明因变量向量 **y** 和自变量向量 **x** 越相近。也就是说，夹角越小，自变量向量 **x** 能更充分解释因变量向量 **y**。本章后文还会利用这个几何视角解释回归分析结果。

本章内容相对比较枯燥，建议大家主要理解 ANOVA。本章其余内容，大家有实际需要时再回头查阅。

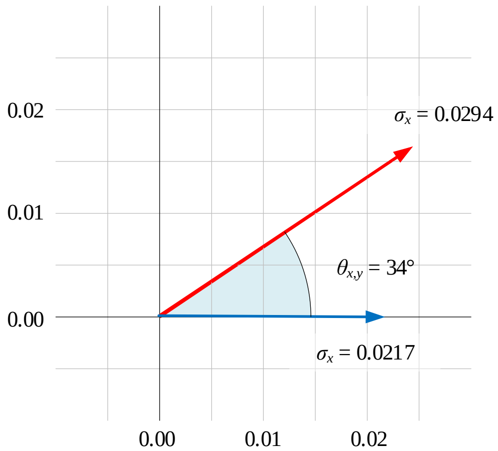


图 5. 标准差向量空间角度解释夹角

Bk6_Ch10_01.py 绘制本节图像。

10.2 方差分析 ANOVA

本节开始先介绍如何理解图 6 所示的 ANOVA 表格结果。ANOVA 的含义是方差分析 (Analysis of Variance)。ANOVA 是图 1 的重要组成部分之一。

```
             df    sum_sq   mean_sq           F        PR(>F)
x           1.0  0.149314  0.149314  549.729877  4.547141e-65
Residual  250.0  0.067903  0.000272         NaN           NaN
```

图 6. 一元线性回归 ANOVA 表格，来自本书第 6 章

表 1 所示为标准 ANOVA 表格对应的统计量。标准 ANOVA 表格比图 6 多一行。表 1 有五列：

第 1 列为计算方差的三个来源；

第 2 列 df 代表自由度 (degrees of freedom)；

第 3 列 SS 代表 Sum of Squares；

第 4 列 MS 代表 Mean Sum of Squares；

第 5 列 F 代表 F-test 统计量。

表中 n 代表参与回归的非 NaN 样本数量。k 代表回归模型参数数量，包括截距项。D 代表因变量的数量，因此 $k = D + 1$。下面将逐个解密表 1 中的每一个值的含义，以及它们和线性回归的关系。

表 1. ANOVA 表格

Source	df	SS	MS	F	Significance
Regressor	DFR = $D = k - 1$	SSR	MSR = SSR/DFR	F = MSR/MSE	p-value of F-test
Residuals	DFE = $n - D - 1 = n - k$	SSE	MSE = SSE/DFE		
Total	DFT = $n - 1$	SST			

三个平方和

为了理解 ANOVA 表格，我们首先要了解三个平方和：

◄ **总离差平方和** (Sum of Squares for Total, SST)，也称 TSS (total sum of squares)；

◄ **残差平方和** (Sum of Squares for Error, SSE)，也称 RSS (residual sum of squares)；

◀ **回归平方和** (Sum of Squares for Regression, SSR)，也称 ESS (explained sum of squares)。

图 7 给出计算三个平方和所需的数值。表 2 总结了三个平方和的定义。

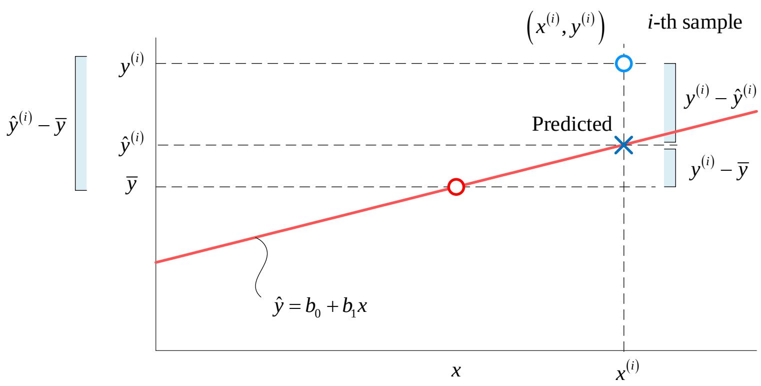
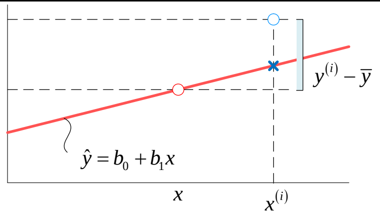
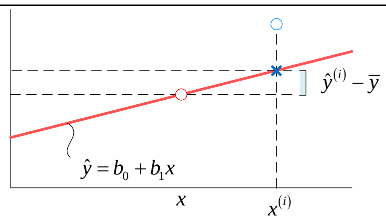
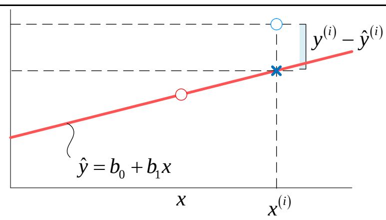


图 7. 通过一元线性回归模型分解因变量的变化

表 2. 三个平方和的定义

平方和	定义	图像
总离差平方和 (Sum of Squares for Total, SST)	$\text{SST} = \sum_{i=1}^{n}\left(y^{(i)} - \bar{y}\right)^2$	
回归平方和 (Sum of Squares for Regression, SSR)	$\text{SSR} = \sum_{i=1}^{n}\left(\hat{y}^{(i)} - \bar{y}\right)^2$	
残差平方和 (Sum of Squares for Error, SSE)	$\text{SSE} = \sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2$	

等式关系

对于线性回归来说，方差分析 (analysis of variance) 实际上就是分解总离差平方和 SST。把 SST 分解成残差平方和 SSE、回归平方和 SSR：

$$\text{SST} = \text{SSR} + \text{SSE} \tag{1}$$

即：

$$\underbrace{\sum_{i=1}^{n}\left(y^{(i)}-\bar{y}\right)^2}_{\text{SST}} = \underbrace{\sum_{i=1}^{n}\left(\hat{y}^{(i)}-\bar{y}\right)^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^{n}\left(y^{(i)}-\hat{y}^{(i)}\right)^2}_{\text{SSE}} \tag{2}$$

上式的证明并不难，本节不做展开讲解，本章后续会用向量几何视角解释以上等式关系。

本章后续将介绍由这三个平方和引出的一些列有关回归的统计量，特别是 R-squared 和 Adj. R-squared。

10.3 总离差平方和 SST

总离差平方和 (Sum of Squares for Total, SST) 代表因变量 y 所有样本点与期望值 $\bar{y}$ 的差异：

$$\text{SST} = \sum_{i=1}^{n}\left(y^{(i)}-\bar{y}\right)^2 \tag{3}$$

其中，期望值 $\bar{y}$ 为：

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y^{(i)} \tag{4}$$

如图 8 所示，SST 可以看做一系列正方形面积之和。这些正方形的边长为 $\left|y^{(i)}-\bar{y}\right|$。图 8 中这些正方形的一条边都在期望值 $\bar{y}$ 这个高度上。

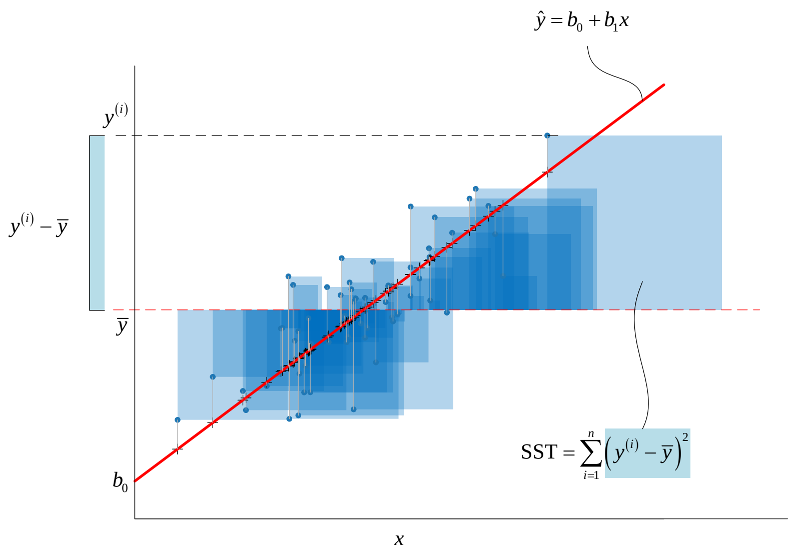


图 8. 总离差平方和 SST

总离差自由度 DFT

总离差自由度 (degree of freedom total, DFT) 的定义为：

$$\mathrm{DFT} = n-1 \tag{5}$$

n 是样本数据的数量 (NaN 除外)。

三个自由度关系

总离差自由度 DFT、回归自由度 DFR、残差自由度 DFE 三者关系为：

$$\mathrm{DFT} = n-1 = \mathrm{DFR} + \mathrm{DFE} = \underbrace{(k-1)}_{\mathrm{DFR}} + \underbrace{(n-k)}_{\mathrm{DFE}} = \underbrace{(D)}_{\mathrm{DFR}} + \underbrace{(n-D-1)}_{\mathrm{DFE}} \tag{6}$$

k 是回归模型的参数，其中包括截距项。因此，

$$k = D+1 \tag{7}$$

D 为参与回归模型的特征数，也就是因变量的数量。

举个例子，对于一元线性回归，$D = 1$，$k = 2$。如果参与建模的样本数据为 $n = 252$，几个自由度分别为：

$$\begin{cases} \mathrm{DFT} = 252-1 = 251 \\ k = D+1 = 2 \\ \mathrm{DFR} = k-1 = D = 1 \\ \mathrm{DFE} = n-k = n-D-1 = 252-2 = 250 \end{cases} \tag{8}$$

平均总离差 MST

平均总离差 (mean square total, MST) 的定义为：

$$\mathrm{MST} = \mathrm{var}(Y) = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1} = \frac{\mathrm{SST}}{\mathrm{DFT}} \tag{9}$$

实际上，总离差 MST 便是因变量 Y 样本数据方差。看到这里，大家应该理解为什么本章的内容叫“方差分析”了。

10.4 回归平方和 SSR

回归平方和 (Sum of Squares for Regression, SSR) 代表回归方程计算得到的预测值 $\hat{y}^{(i)}$ 和期望值 $\bar{y}$ 之间的差异：

$$\text{SSR} = \sum_{i=1}^{n} \left(\hat{y}^{(i)} - \bar{y} \right)^2 \tag{10}$$

图 9 所示为回归平方和 SSR 的几何意义。图 9 中的每个正方形边长为 $\left| \hat{y}^{(i)} - \bar{y} \right|$。

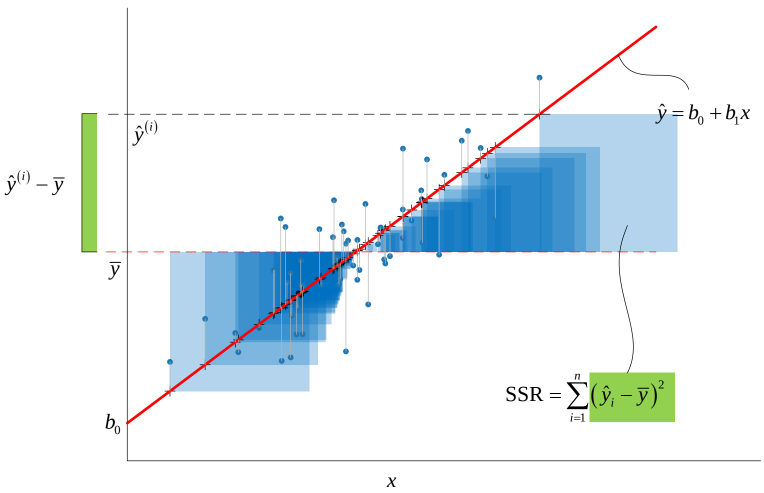


图 9. 回归平方和

回归自由度 DFR

回归自由度 (degrees of freedom for regression model, DFR) 为：

$$\text{DFR} = k - 1 = D \tag{11}$$

平均回归平方 MSR

平均回归平方 (mean square regression, MSR) 为：

$$\text{MSR} = \frac{\text{SSR}}{\text{DFR}} = \frac{\text{SSR}}{k-1} = \frac{\text{SSR}}{D} \tag{12}$$

10.5 残差平方和 SSE

残差平方和 (Sum of Squares for Error, SSE) 定义如下：

$$\text{SSE} = \sum_{i=1}^{n} \left(\varepsilon^{(i)}\right)^2 = \sum_{i=1}^{n} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 \tag{13}$$

相信大家对残差平方和 SSE 已经很熟悉。比如，在最小二乘法中，我们通过最小化残差平方和 SSE 优化回归参数。

图 10 所示为残差平方和 SSE 的示意图。图中每个正方形的边长为 $\left|y^{(i)} - \hat{y}^{(i)}\right|$。

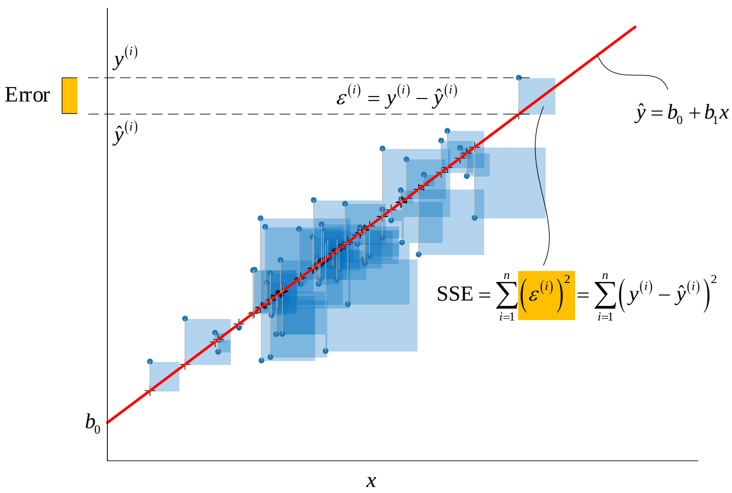


图 10. 残差平方和 SSE

残差自由度 DFE

残差自由度 (degrees of freedom for error, DFE) 为：

$$\text{DFE} = n - k = n - D - 1 \tag{14}$$

残差平均值 MSE

残差平均值 (mean squared error, MSE) 为：

$$\text{MSE} = \frac{\text{SSE}}{\text{DFE}} = \frac{\text{SSE}}{n-k} = \frac{\text{SSE}}{n-D-1} \tag{15}$$

均方根残差 RMSE

均方根残差 (Root mean square error, RMSE) 为 MSE 的平方根：

$$\mathrm{RMSE}=\sqrt{\mathrm{MSE}}=\sqrt{\frac{\mathrm{SSE}}{\mathrm{DFE}}}=\sqrt{\frac{\mathrm{SSE}}{n-p}}=\sqrt{\frac{\mathrm{SSE}}{n-D-1}} \tag{16}$$

10.6 几何视角：勾股定理

大家别忘了《矩阵力量》反复提到的几何视角！

一个直角三角形

看到 (2) 中三个求和，我们下面用向量方法完成三个求和运算：

$$\begin{aligned}\mathrm{SST}&=\sum_{i=1}^{n}\left(y^{(i)}-\bar{y}\right)^2=\left\|\boldsymbol{y}-\bar{y}\boldsymbol{1}\right\|_2^2\\\mathrm{SSR}&=\sum_{i=1}^{n}\left(\hat{y}^{(i)}-\bar{y}\right)^2=\left\|\hat{\boldsymbol{y}}-\bar{y}\boldsymbol{1}\right\|_2^2\\\mathrm{SSE}&=\sum_{i=1}^{n}\left(y^{(i)}-\hat{y}^{(i)}\right)^2=\left\|\boldsymbol{y}-\hat{\boldsymbol{y}}\right\|_2^2\end{aligned} \tag{17}$$

根据 (2)，我们可以得到如下等式：

$$\underbrace{\left\|\boldsymbol{y}-\bar{y}\boldsymbol{1}\right\|_2^2}_{\mathrm{SST}}=\underbrace{\left\|\hat{\boldsymbol{y}}-\bar{y}\boldsymbol{1}\right\|_2^2}_{\mathrm{SSR}}+\underbrace{\left\|\boldsymbol{y}-\hat{\boldsymbol{y}}\right\|_2^2}_{\mathrm{SSE}} \tag{18}$$

相信大家一眼就会看出来，(18) 代表着直角三角形勾股定理！

如图 11 (a) 所示，$\boldsymbol{y}-\bar{y}\boldsymbol{1}$ 就是斜边对应的向量，斜边长度为 $\left\|\boldsymbol{y}-\bar{y}\boldsymbol{1}\right\|$。$\hat{\boldsymbol{y}}-\bar{y}\boldsymbol{1}$ 为第一条直角边，$\hat{\boldsymbol{y}}-\bar{y}\boldsymbol{1}$ 代表回归模型解释的部分。$\boldsymbol{y}-\hat{\boldsymbol{y}}$ 为第二条直角边，代表残差项，也就是回归模型不能解释的部分。

注意，图 11 中 $\boldsymbol{y}-\bar{y}\boldsymbol{1}$ 和 $\hat{\boldsymbol{y}}-\bar{y}\boldsymbol{1}$ 的起点为 $\bar{y}\boldsymbol{1}$ 的终点，这相当于去均值。

如图 11 (b) 所示，这个勾股定理还可以写成：

$$\left(\sqrt{\mathrm{SST}}\right)^2=\left(\sqrt{\mathrm{SSR}}\right)^2+\left(\sqrt{\mathrm{SSE}}\right)^2 \tag{19}$$

此外，请大家注意图中 θ，θ 是向量 $\boldsymbol{y}-\bar{y}\boldsymbol{1}$ 和向量 $\hat{\boldsymbol{y}}-\bar{y}\boldsymbol{1}$ 的夹角，下一节会用到它。

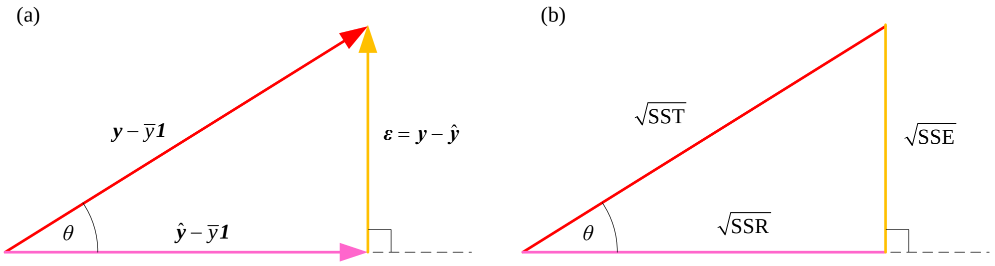


图 11. 几何角度看三个平方和

四个直角三角形

图 11 的直角三角形是图 12 这个四面体的一个面。而图 12 这个四面体的四个面都是直角三角形！

现在请大家自己试着理解这个四面体和四个直角三角形的含义，下一章会深入分析。

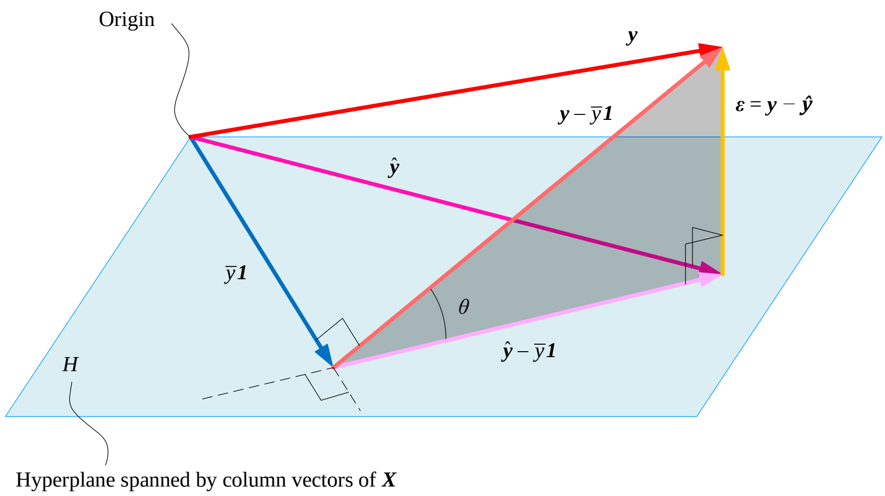


图 12. 四面体的四个面都是直角三角形

10.7 拟合优度：评价拟合程度

如图 13 所示，向量 $y - \bar{y}1$ 和向量 $\hat{y} - \bar{y}1$ 之间夹角 θ 越小，说明误差越小，代表拟合效果越好。

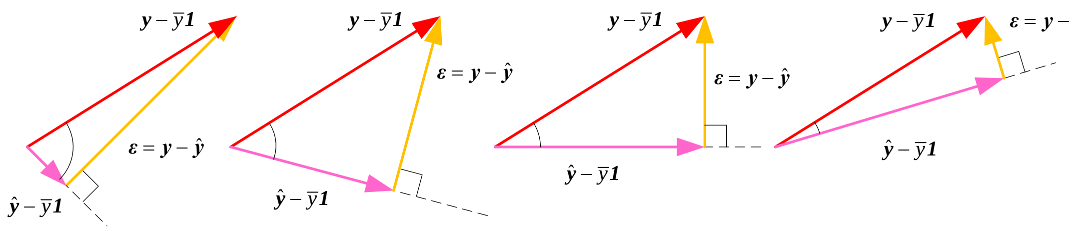


图 13. 因变量向量和预测值向量夹角从大到小

在回归模型创建之后，很自然就要考虑这个模型是否能够很好地解释数据，即考察这条回归线对观察值的拟合程度，也就是所谓的拟合优度 (goodness of fit)。简单地说，拟合优度是回归分析中考察样本数据点对于回归线的贴合程度。

决定系数 (coefficient of determination，R^2) 是定量化反映模型拟合优度的统计量。从几何角度，R^2 是图 12 中 θ 余弦值 $\cos\theta$ 的平方：

$$R^2 = \cos(\theta)^2 \tag{20}$$

利用图 11 (b) 直角三角形三边之间的关系，R^2 可以整理为：

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \tag{21}$$

当预测值越接近样本值，R^2 越接近 1；相反，若拟合效果越差，R^2 越接近 0。

一元线性回归

特别地，对于一元线性回归，决定系数是因变量与自变量的相关系数的平方，与模型系数 b_1 也有直接关系。

$$R^2 = \rho_{X,Y}^2 = \left(b_1 \frac{\sigma_X}{\sigma_Y}\right)^2 \tag{22}$$

其中，

$$b_1 = \rho_{X,Y} \frac{\sigma_Y}{\sigma_X} \tag{23}$$

修正决定系数

但是，仅仅使用 R^2 是不够的。对于多元线性模型，不断增加解释变量个数 D 时，R^2 将不断增大。我们可以利用修正决定系数 (adjusted R squared)，

$$
\begin{aligned}
R_{\text{adj}}^{2} &= 1-\frac{\text{MSE}}{\text{MST}} \\
&= 1-\frac{\text{SSE}/(n-k)}{\text{SST}/(n-1)} \\
&= 1-\left(\frac{n-1}{n-k}\right)\frac{\text{SSE}}{\text{SST}} \\
&= 1-\left(\frac{n-1}{n-k}\right)\left(1-R^{2}\right) \\
&= 1-\left(\frac{n-1}{n-D-1}\right)\frac{\text{SSE}}{\text{SST}}
\end{aligned} \tag{24}
$$

10.8 *F* 检验：模型参数不全为 0

统计量

F 检验的统计量为：

$$
\begin{aligned}
F &= \frac{\text{MSR}}{\text{MSE}} = \frac{\dfrac{\text{SSR}}{k-1}}{\dfrac{\text{SSE}}{n-k}} = \frac{\text{SSR}(n-k)}{\text{SSE}(k-1)} \\
&= \frac{\dfrac{\text{SSR}}{D}}{\dfrac{\text{SSE}}{n-D-1}} = \frac{\text{SSR}\cdot(n-D-1)}{\text{SSE}\cdot(D)} \sim F(k-1, n-k)
\end{aligned} \tag{25}
$$

原假设、备择假设

F 检验是单尾检验，原假设 H_0、备择假设 H_1 分别为：

$$
\begin{aligned}
&H_0:\ b_1 = b_2 = \cdots = b_D = 0 \\
&H_1:\ b_j \neq 0 \ \text{ for at least one } j
\end{aligned} \tag{26}
$$

如果是显著的，因变量与自变量之间存在线性关系。如果不显著，因变量与自变量之间不存在线性关系。

临界值

(25) 得到的 F 值和临界值 F_α 进行比较。临界值 F_α 可根据两个自由度 ($k-1$ 和 $n-k$) 以及置信水平 α 查表获得。$1-\alpha$ 为置信度或置信水平，通常取 $\alpha = 0.05$ 或 $\alpha = 0.01$。这表明，当作出接受原假设的决定时，其正确的可能性为 95%或 99%。

如果，

$$F > F_{1-\alpha}\left(k-1, n-k\right) \tag{27}$$

在该置信水平上拒绝零假设 H_0，不认为自变量系数同时具备非显著性，即所有系数不太可能同时为零。

否则，接受 H_0，自变量系数同时具有非显著性，即所有系数很可能同时为零。

举个例子

给定条件 $\alpha = 0.01$，$F_{1-\alpha}(1, 250) = 6.7373$。图 6 结果告诉我们，$F = 549.7 > 6.7373$，表明可以显著地拒绝 H_0。

也可以用图 6 中 p 值，

$$p\text{-value} = \mathrm{P}\left(F < F_{\alpha}\left(k-1, n-k\right)\right) \tag{28}$$

如果 p 值小于 α，则可以拒绝零假设 H_0。

Bk6_Ch10_02.py 计算图 6 所示方差分析表格中统计量。

10.9 t 检验：某个回归系数是否为 0

对于一元线性回归，t 检验原假设和备择假设分别为：

$$\begin{cases} H_0: \ b_1 = b_{1,0} \\ H_1: \ b_1 \neq b_{1,0} \end{cases} \tag{29}$$

一般 $b_{1,0}$ 取 0，也就是检验回归系数是否为 0。当然，$b_{1,0}$ 也可以取其他值。

b_1 的 t 检验统计值：

$$t_{b1} = \frac{\hat{b}_1 - b_{1,0}}{\mathrm{SE}\left(\hat{b}_1\right)} \tag{30}$$

$\hat{b}_1$ 为最小二乘法 OLS 线性回归估算得到的系数，$\mathrm{SE}\left(\hat{b}_1\right)$ 为其标准误：

$$\mathrm{SE}\left(\hat{b}_1\right)=\sqrt{\frac{\mathrm{MSE}}{\sum_{i=1}^{n}\left(x^{(i)}-\bar{x}\right)^2}}=\sqrt{\frac{\frac{\sum_{i=1}^{n}\left(\varepsilon^{(i)}\right)^2}{n-2}}{\sum_{i=1}^{n}\left(x^{(i)}-\bar{x}\right)^2}} \tag{31}$$

上式中，MSE 为本章前文介绍的残差平均值 (mean squared error)，n 是样本数据的数量 (除 NaN)。标准误差越大，回归系数的估计值越不可靠。

如果下式成立，接受零假设 H_0：

$$-t_{1-\alpha/2,\ n-2}<T<t_{1-\alpha/2,\ n-2} \tag{32}$$

否则，则拒绝零假设 H_0。

特别地，如果原假设和备择假设为：

$$\begin{cases} H_0:\ b_1=0 \\ H_1:\ b_1\neq 0 \end{cases} \tag{33}$$

如果 (32) 成立，接受零假设 H_0，即回归系数不具有显著统计性；白话说，也就是 $b_1 = 0$，意味着自变量和因变量不存在线性关系。否则，则拒绝零假设 H_0，即回归系数具有显著统计性。

对于一元线性回归，对截距项系数 b0 的假设检验程序和上述类似。b_0 的 t 检验统计值：

$$t_{b0}=\frac{\hat{b}_0-b_{0,0}}{\mathrm{SE}\left(\hat{b}_0\right)} \tag{34}$$

$\hat{b}_0$ 为最小二乘法 OLS 线性回归估算得到的系数，$\mathrm{SE}\left(\hat{b}_0\right)$ 为其标准误：

$$\mathrm{SE}\left(\hat{b}_0\right)=\sqrt{\frac{\sum_{i=1}^{n}\left(\varepsilon^{(i)}\right)^2}{n-2}\left[\frac{1}{n}+\frac{\bar{x}^2}{\sum_{i=1}^{n}\left(x^{(i)}-\bar{x}\right)^2}\right]} \tag{35}$$

t 检验统计值 T 服从自由度为 $n-2$ 的 t 分布。本节采用的 t 检验是双尾检测。比如给定显著性水平 $\alpha = 0.05$ 和自由度 $n - 2 = 252 - 2 = 250$，可以查表得到 t 值，即：

$$t_{1-\alpha/2,\ n-2}=t_{0.975,\ 250}=1.969498 \tag{36}$$

Python 中，可以用 stats.t.ppf(1 - alpha/2, DFE) 计算上式两值。

由于学生 t-分布对称，所以：

$$t_{\alpha/2,\ n-2}=t_{0.025,\ 250}=-1.969498 \tag{37}$$

如图 1 所示，$t_{b1} = 23.446$，因此：

$$t_{b1}>t_{0.975,\ 250} \tag{38}$$

表明参数 b_1 的 t 检验在 $\alpha = 0.05$ 水平下是显著的，也就是可以显著地拒绝 H_0: $b_1 = 0$，从而接受 H_1: $b_1 \neq 0$。回归系数的标准误差越大，回归系数的估计值越不可靠。

而 $t_{b0} = 1.759$，因此：

$$t_{b0} < t_{0.975,\ 250} \tag{39}$$

则表明参数 b_0 的 t 检验在 $\alpha = 0.05$ 水平下是不显著的，也就是不能显著地拒绝 H_0: $b_0 = 0$。尽管模型含有截距项，但若该项的出现是统计上不显著的 (即统计上等于零)，则从任何实际方面考虑，都可认为这个结果是一个过原点回归模型。

因此，系数 b_1 的 $1 - \alpha$ 置信区间为：

$$\hat{b}_1 \pm t_{1-\alpha/2,\ n-2} \cdot \mathrm{SE}\left(\hat{b}_1\right) \tag{40}$$

这个置信区间的的含义是，真实 b_1 在以上区间的概率为 $1 - \alpha$。

系数 b_0 的 $1 - \alpha$ 置信区间为：

$$\hat{b}_0 \pm t_{1-\alpha/2,\ n-2} \cdot \mathrm{SE}\left(\hat{b}_0\right) \tag{41}$$

同理，真实 b_0 在以上区间的概率为 $1 - \alpha$。

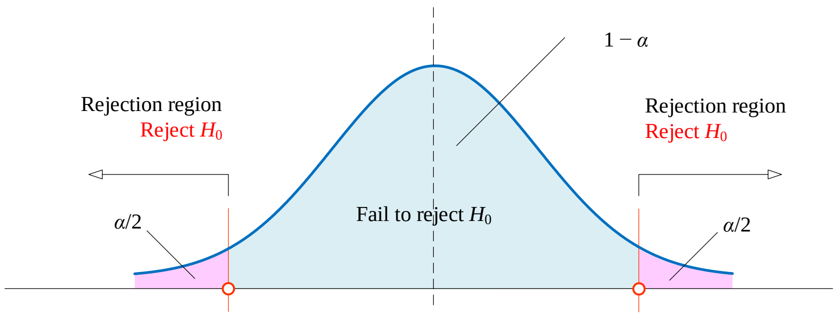


图 14. 双尾检验

10.10 置信区间：因变量均值的区间

本书前文在介绍一元线性回归中，大家都应该见过类似图 15 的图像。图中的带宽代表预测值的置信区间。

预测值 $\hat{y}^{(i)}$，的 $1 - \alpha$ 置信区间：

$$\hat{y}^{(i)} \pm t_{1-\alpha/2,\ n-2} \cdot \sqrt{\mathrm{MSE}} \cdot \sqrt{\frac{1}{n} + \frac{\left(x^{(i)} - \bar{x}\right)^2}{\sum_{k=1}^{n}\left(x^{(k)} - \bar{x}\right)^2}} \tag{42}$$

置信区间的宽度为：

$$2 \times \left\{ t_{1-\alpha/2,\, n-2} \cdot \sqrt{\text{MSE}} \cdot \sqrt{\frac{1}{n} + \frac{\left(x^{(i)} - \bar{x}\right)^2}{\sum_{k=1}^{n} \left(x^{(k)} - \bar{x}\right)^2}} \right\} \tag{43}$$

随着 $\left|x^{(i)} - \bar{x}\right|$ 不断增大，置信区间宽度不断增大。当 $x^{(i)} = \bar{x}$ 时，置信区间宽度最窄。随着 MSE (mean square error) 减小，置信区间宽度减小。

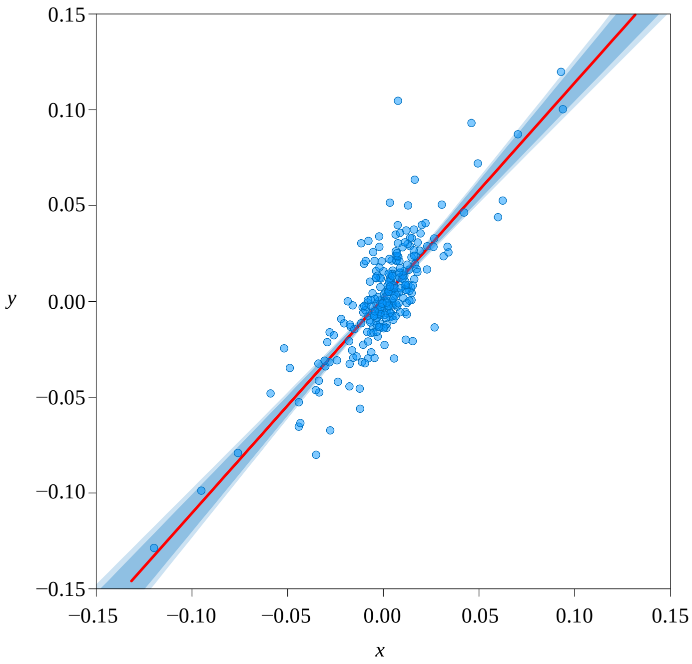


图 15. 一元线性回归线置信区间

10.11 预测区间：因变量特定值的区间

预测区间 (prediction interval) 是指回归模型估计时，对于自变量给定的某个值 x_p，求出因变量 y_p 的个别值的估计区间：

$$\hat{y}_p \pm t_{1-\alpha/2,\, n-2} \cdot \sqrt{\text{MSE}} \cdot \sqrt{1 + \frac{1}{n} + \frac{\left(x_p - \bar{x}\right)^2}{\sum_{k=1}^{n} \left(x^{(k)} - \bar{x}\right)^2}} \tag{44}$$

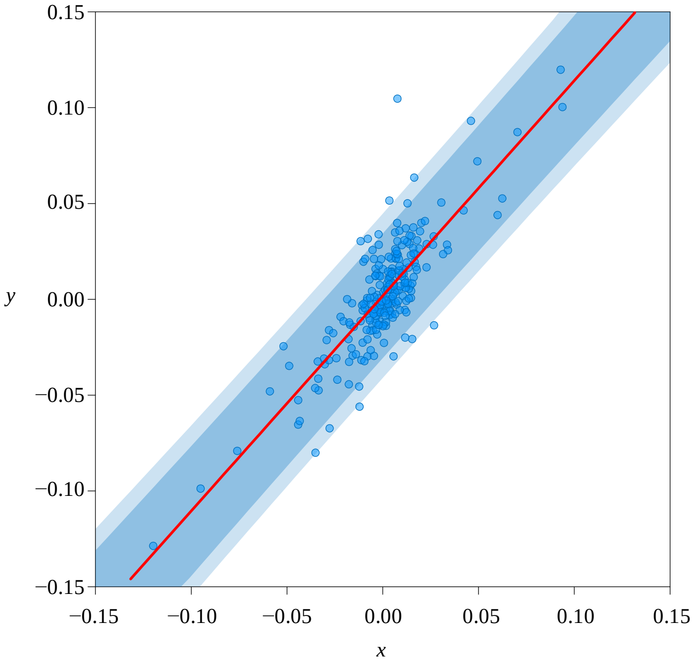


图 16. 一元线性回归线预测区间

10.12 对数似然函数：用在最大似然估计 MLE

似然函数是一种关于统计模型中的参数的函数，表示模型参数中的似然性。

残差的定义为：

$$\varepsilon^{(i)} = y^{(i)} - \hat{y}^{(i)} \tag{45}$$

在 OLS 线性回归中，假设残差服从正态分布 $N(0, \sigma^2)$，因此：

$$\Pr\left(\varepsilon^{(i)}\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(y^{(i)} - \hat{y}^{(i)}\right)^2}{2\sigma^2}\right) \tag{46}$$

似然函数为：

$$L = \prod_{i=1}^{n} \mathrm{P}\left(\varepsilon^{(i)}\right) = \prod_{i=1}^{n} \left\{\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{\left(y^{(i)} - \hat{y}^{(i)}\right)^2}{2\sigma^2}\right)\right\} \tag{47}$$

常用对数似然 $\ln(L)$：

$$\ln\left(L\right) = \prod_{i=1}^{n} \mathrm{P}\left(\varepsilon^{(i)}\right) = -n \cdot \ln\left(\sigma\sqrt{2\pi}\right) - \frac{\text{SSE}}{2\sigma^2} \tag{48}$$

注意，MLE 中的 σ 为：

$$\sigma^2 = \frac{\text{SSE}}{n} \tag{49}$$

这样 ln(*L*) 可以写成：

$$\ln(L) = \prod_{i=1}^{n} \mathrm{P}\left(\varepsilon^{(i)}\right) = -n \cdot \ln\left(\sigma\sqrt{2\pi}\right) - \frac{n}{2} \tag{50}$$

10.13 信息准则：选择模型的标准

AIC 为**赤池信息量准则** (Akaike information criterion, AIC)，定义如下：

$$\text{AIC} = \underbrace{2k}_{\text{Penalty}} - 2\ln(L) \tag{51}$$

其中，$k = D + 1$；L 是似然函数。

AIC 鼓励数据拟合的优良性；但是，尽量避免出现过度拟合。(51) 中 $2k$ 项为**惩罚项** (penalty)。

贝叶斯信息准则 (Bayesian Information Criterion, BIC) 也称**施瓦茨信息准则** (Schwarz information criterion, SIC)，定义如下。

$$\text{BIC} = \underbrace{k \cdot \ln(n)}_{\text{Penalty}} - 2\ln(L) \tag{52}$$

其中，n 为样本数据数量。BIC 的惩罚项比 AIC 大。

10.14 残差分析：假设残差服从均值为 0 正态分布

残差分析 (residual analysis) 通过残差所提供的信息，对回归模型进行评估，分析数据是否存在可能的干扰。

图 17 所示为残差的散点图。图 18 所示为残差分布的直方图。理想情况下，我们希望残差为均值为 0 的正态分布。为了检测残差的正态性，我们常用 Omnibus 正态检验、Jarque-Bera 检验。

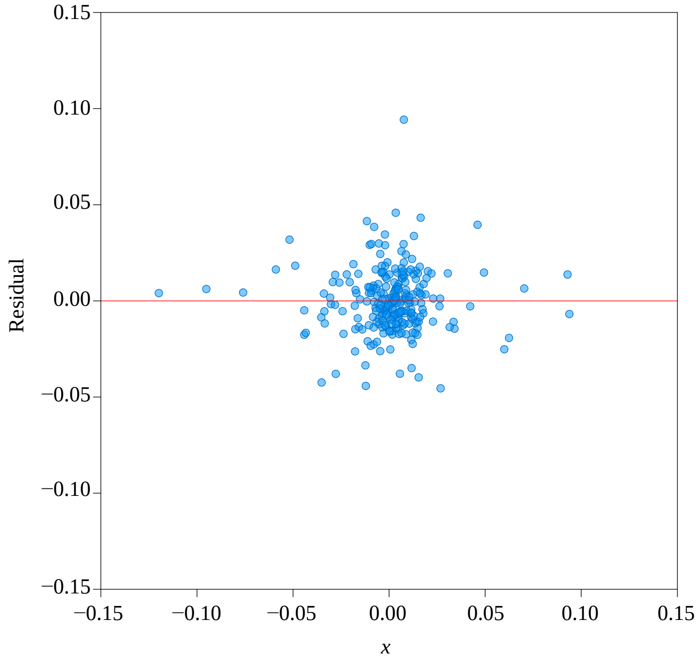


图 17. 残差散点图

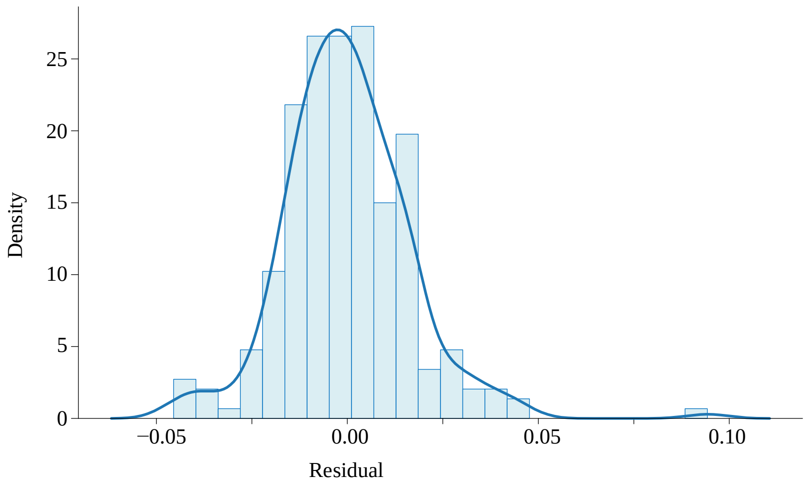


图 18. 残差分布直方图

Omnibus 正态检验 (Omnibus test for normality) 利用偏度 S 和峰度 K，检验残差分布为正态分布的原假设。Omnibus 正态检验的统计值为偏度平方、超值峰度平方两者之和。Omnibus 正态检验利用 $\chi 2$ 检验 (Chi-squared test)。

代码中我们利用 scipy.stats.normaltest() 复现了本章前文的 Omnibus 正态检验统计量值。

残差偏度 (skewness) S 的定义为：

$$S = \text{skewness} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2\right)^{\frac{3}{2}}} \tag{53}$$

残差峰度 (kurtosis) K 的定义：

$$K = \text{kurtosis} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} - \hat{y}^{(i)}\right)^2\right)^{2}} \tag{54}$$

K 减 3 就是超值峰度 (excess kurtosis)。

Jarque-Bera 检验也是用偏度 S 和峰度 K 来检验残差分布为正态分布的原假设：

$$JB = \frac{n}{6}\left(S^2 + \frac{1}{4}(K-3)^2\right) \tag{55}$$

Jarque-Bera 检验也利用 $\chi2$ 检验。

10.15 自相关检测：Durbin-Watson

Durbin-Watson 用于检验序列的自相关。图 19 所示为残差的自相关图。

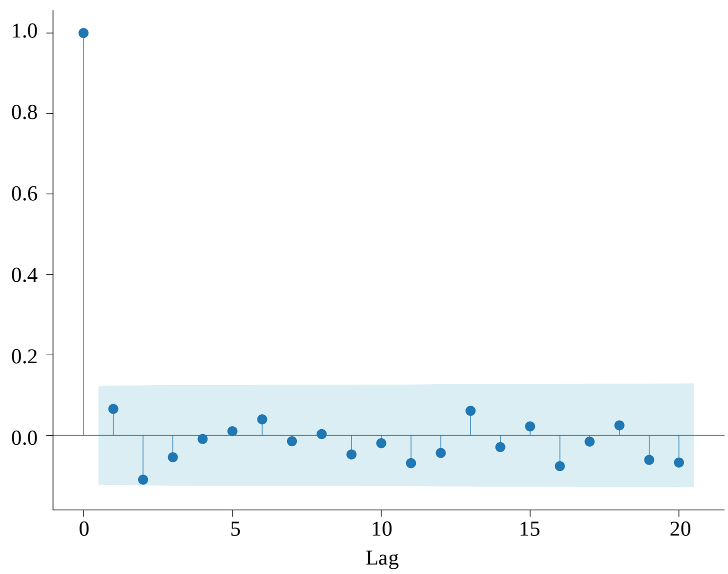


图 19. 残差自相关

Durbin-Watson 检测的统计量为：

$$DW = \frac{\sum_{i=2}^{n}\left(\left(y^{(i)}-\hat{y}^{(i)}\right)-\left(y^{(i-1)}-\hat{y}^{(i-1)}\right)\right)^2}{\sum_{i=1}^{n}\left(y^{(i)}-\hat{y}^{(i)}\right)^2} \tag{56}$$

上式本质上检测残差序列与残差的滞后一期序列之间的差异大小。DW 值的取值区间为 0 ~ 4。当 DW 值很小时 ($DW < 1$)，表明序列可能存在正自相关。当 DW 值很大时 ($DW > 3$) 表明序列可能存在负自相关。当 DW 值在 2 附近时 ($1.5 < DW < 2.5$)，表明序列无自相关。其余的取值区间表明无法确定序列是否存在自相关。

10.16 条件数：多重共线性

在线性回归中，条件数 (condition number) 常用来检验设计矩阵 $\boldsymbol{X}_{k\times k}$ 是否存在多重共线性。

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：

$$\text{condition number} = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \tag{57}$$

条件数小于 30，可以不必担心多重共线性。

下一章讲到多元回归分析时，条件数的作用更明显。

Bk6_Ch10_03.py 代码复现图 1 中除 ANOVA 以外的其他统计量值。

Scikit-learn 也提供线性回归分析工具，请大家参考如下网页：

https://scikit-learn.org/stable/auto_examples/inspection/plot_linear_model_coefficient_interpretation.html