

2

Regression Analysis

回归分析

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



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

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



2.1 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 章结尾给出过图 1 这个表格。

图 1 这个表格汇总某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。

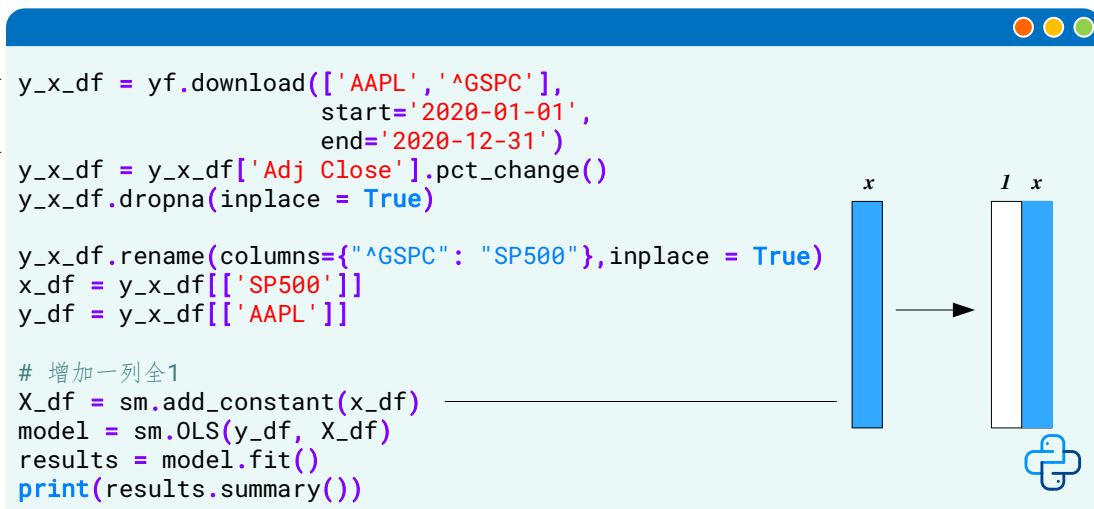
```

OLS Regression Results
=====
Dep. Variable:          AAPL      R-squared:                0.689
Model:                  OLS      Adj. R-squared:           0.687
Method:                 Least Squares      F-statistic:            550.5
Date:                   Mon, 01 Jan 2024    Prob (F-statistic):      5.16e-65
Time:                   07:03:51           Log-Likelihood:          675.37
No. Observations:       251              AIC:                   -1347.
Df Residuals:           249              BIC:                   -1340.
Df Model:                1
Covariance Type:        nonrobust
=====
                    coef      std err          t      P>|t|      [0.025      0.975]
-----
const              0.0019      0.001        1.819      0.070      -0.000      0.004
SP500              1.1234      0.048       23.462      0.000        1.029      1.218
=====
Omnibus:                52.109      Durbin-Watson:           1.871
Prob(Omnibus):           0.000      Jarque-Bera (JB):        210.792
Skew:                   0.772      Prob(JB):                1.69e-46
Kurtosis:                7.216      Cond. No.:               46.0
=====

```

图 1. 一元线性回归结果 |  Bk7_Ch02_01.ipynb


Bk7_Ch02_01.ipynb 绘制本节图像。下面，让我们一起简单聊聊其中关键语句。



```

a y_x_df = yf.download(['AAPL', '^GSPC'],
b                       start='2020-01-01',
c                       end='2020-12-31')
d y_x_df = y_x_df[['Adj Close']].pct_change()
e y_x_df.dropna(inplace = True)
f y_x_df.rename(columns={'^GSPC': "SP500"}, inplace = True)
g x_df = y_x_df[['SP500']]
h y_df = y_x_df[['AAPL']]
i
j # 增加一列全1
X_df = sm.add_constant(x_df)
model = sm.OLS(y_df, X_df)
results = model.fit()
print(results.summary())

```

代码 1. 下载、处理数据并完成回归运算 |  Bk7_Ch02_01.ipynb

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

代码 1 下载并处理数据并完成回归运算。

a 用 `yfinance.download`, 简作 `yf.download`, 下载金融数据。大家首先需要用 `pip install yfinance` 安装该库。

本例中, 我们下载了苹果公司 (AAPL) 和标准普尔 500 指数 (^GSPC) 在 2020 年 1 月 1 日到 2020 年 12 月 31 日期间的股价数据。下载的数据被存储在数据帧 `y_x_df` 的数据帧中。

此外, 如果大家下载数据遇到困难, 我们还用 `to_csv()` 和 `to_pickle()` 将数据保存下来。大家可以用 `pandas.read_csv()` 或 `pandas.read_pickle()` 直接读取数据。

b 提取 'Adj Close', 即调整后收盘价, 即考虑了股票分红和拆股等因素后的收盘价。然后, 用方法 `pct_change()` 计算日收益率。

c 删除数据框中包含缺失值的行。`inplace=True` 会直接在原始数据帧上进行修改, 而不是返回一个新的数据帧。

d 用 `rename()` 方法将 "^GSPC" 列名设置为 "SP500"。同样, `inplace=True` 直接在原始数据帧上进行修改, 而不返回一个新的数据帧。

e 和 **f** 分别提取两列作为回归分析中的因变量和因变量散点数据。

g 用 `statsmodels.api`, 简作 `sm`, 中的 `add_constant()` 给数据帧 `x_df` 增加一列全 1 列。目的是为了使线性回归模型能够拟合常数项, 即截距项。如果没有这一列全 1 列, 我们得到的便是无截距线性回归模型。

h 创建了一个最小二乘线性回归模型对象。在这个函数中, `y_df` 是因变量数据, `X_df` 是自变量数据。**i** 完成回归模型拟合。

j 打印线性回归分析结果。

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 (股指) 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, 股指可以部分“解释”同一个市场上股票的涨跌。图 1 是利用 `statsmodels.api.OLS()` 函数构造的线性模型结果。图 3 所示为用 `seaborn.jointplot()` 绘制回归图, 并且绘制边际分布。

特别是从散点图中, 我们明显能够看到很强的正相关性, 下面让我们量化这种相关性。

⚠ 再次强调, 线性回归不代表“因果关系”。

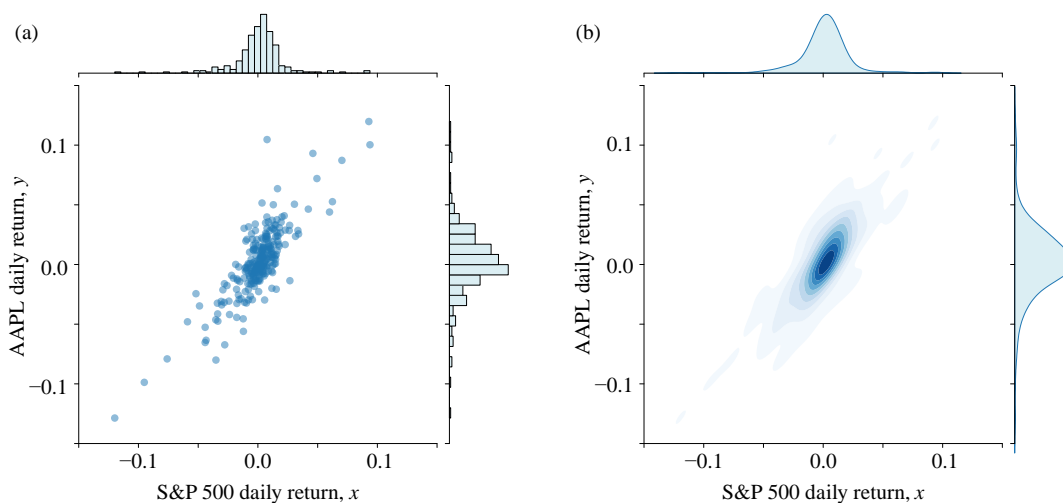


图 2. 日收益率数据关系 | Bk7_Ch02_01.ipynb

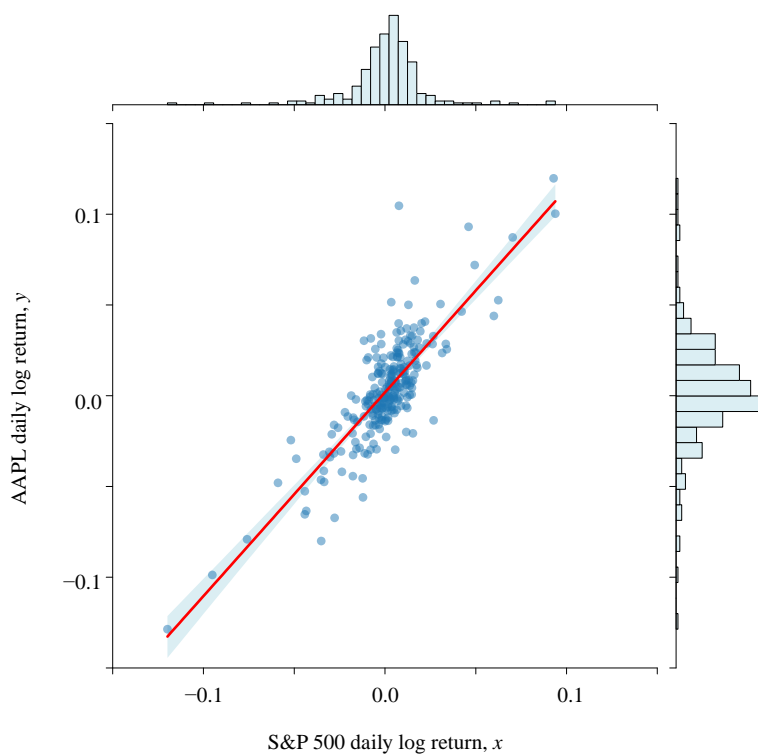


图 3. 用 seaborn.jointplot() 绘制回归直线 | Bk7_Ch02_01.ipynb

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。

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 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。

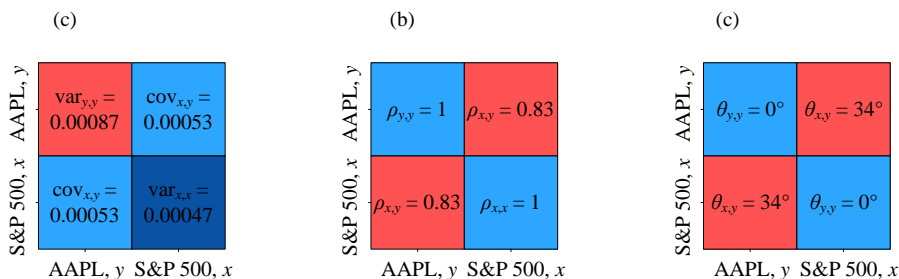


图 4. $[y, x]$ 数据的协方差矩阵、相关性和夹角热图 | Bk7_Ch02_01.ipynb

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。大家有实际需要时再回头查阅本章其余内容。

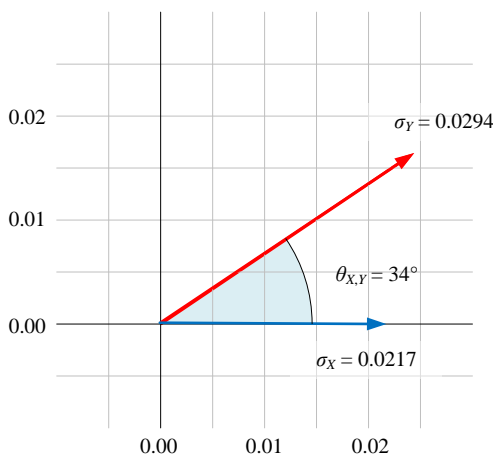


图 5. 标准差向量空间角度解释夹角 | Bk7_Ch02_01.ipynb

2.2 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)。方差分析是一种用于确定线性回归模型中不同变量对目标变量解释程度的统计技术。方差分析通过比较模型中不同的变量的平均方差，来确定哪些变量对目标变量的解释程度更高。

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

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

代码 2 完成 ANOVA 分析，下面聊聊其中关键语句。

```
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm

a data = pd.DataFrame({'x': x_df['SP500'], 'y': y_df['AAPL']})
b model_V2 = ols("y ~ x", data).fit()
c anova_results = anova_lm(model_V2, typ=1)

d print(anova_results)
```

代码 2. ANOVA 分析 |  Bk7_Ch02_02.ipynb

a 用 `pandas.DataFrame()` 构造一个数据帧，列名 'x' 中为 SP500 数据，列名 'y' 中为 AAPL 数据。

前文用 `from statsmodels.formula.api import ols` 从 `Statsmodels` 中的 `formula.api` 模块导入了 `ols` 函数。这个函数允许使用公式字符串来指定线性回归模型。**b** 创建了一个线性回归模型。"y ~ x" 是一个公式字符串，它表示因变量 y 与自变量 x 之间的线性关系。`data` 是包含数据的数据数据帧。

前文先用 `from statsmodels.stats.anova import anova_lm` 从 `Statsmodels` 中的 `stats.anova` 模块导入了 `anova_lm` 函数。**c** 使用 `anova_lm` 函数进行方差分析。`model_V2` 是线性回归模型的对象，而 `typ=1` 表示使用 "Type I" 方差分析。Type I 方差分析逐步添加每个自变量，检验每个自变量的贡献。方差分析通常用于确定模型中是否有显著的变量，以及这些变量对因变量的贡献程度。

d 打印方差分析的结果，具体如图 6。

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 列为方差的三个来源；
- ▶ 第 2 列 df 代表**自由度** (degrees of freedom)；自由度是指在计算统计量时可以随意变化的独立数据点的数量。
- ▶ 第 3 列 SS 代表**平方和** (Sum of Squares)；平方和通常用于描述数据的变异程度，即它们偏离平均值的程度。
- ▶ 第 4 列 MS 代表**均方和** (Mean Sum of Squares)；在统计学中，均方和是一种平均值的度量，其计算方法是将平方和除以自由度。
- ▶ 第 5 列 F 代表 *F*-test 统计量。*F* 检验是一种基于方差比较的统计检验方法，用于确定两个或多个样本之间是否存在显著性差异。

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

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 $k = D + 1$ (+1 代表常数项参数)。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	$MSR = SSR/DFR$	$F = MSR/MSE$	$p\text{-value of } F\text{-test}$
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)。总离差平方和 SST 描述所有观测值与总体均值之间差异的平方和，用来评整个数据集的离散程度。
- ◀ **残差平方和** (Sum of Squares for Error, SSE)，也称 RSS (residual sum of squares)。残差平方和 SSE 反映了因变量中无法通过自变量预测的部分，也称为误差项，可以用于检查回归模型的拟合程度和判断是否存在异常值。在回归分析中，常用通过最小化残差平方和来确定最佳的回归系数。
- ◀ **回归平方和** (Sum of Squares for Regression, SSR)，也称 ESS (explained sum of squares)。回归平方和 SSR 反映了回归模型所解释的数据变异量的大小，用于评估回归模型的拟合程度以及自变量对因变量的影响程度。

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。

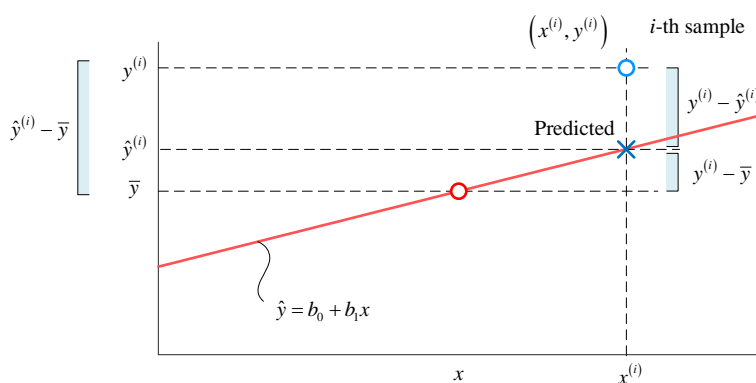


图 7. 通过一元线性回归模型分解因变量的变化

表 2. 三个平方和的定义

平方和	定义	图像
-----	----	----

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

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

等式关系

对于线性回归来说，方差分析实际上就是把 SST 分解成残差平方和 SSE、回归平方和 SSR：

$$SST = SSR + SSE \quad (1)$$

即：

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

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。此外，本章还会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。

2.3 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：

$$SST = \sum_{i=1}^n (y^{(i)} - \bar{y})^2 \quad (3)$$

其中，期望值 \bar{y} 为：

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

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 $|y^{(i)} - \bar{y}|$ 。图 8 中这些正方形的一条边都在期望值 \bar{y} 这个高度上。

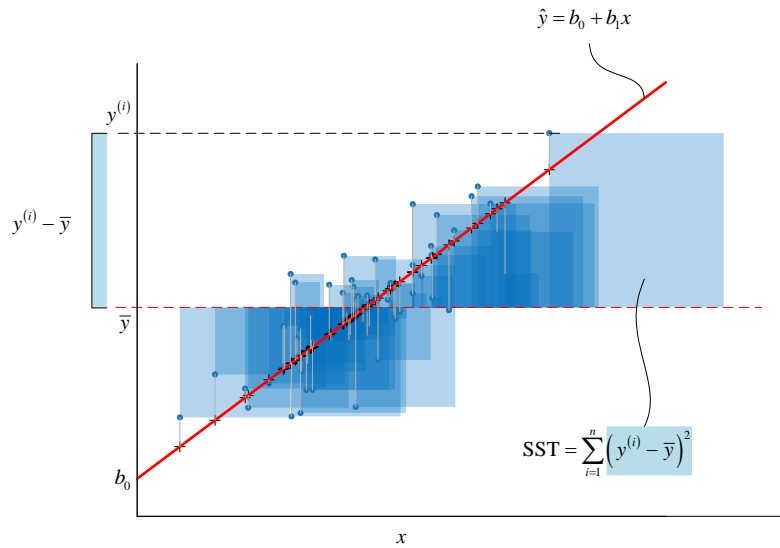


图 8. 总离差平方和 SST

总离差自由度 DFT

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

$$DFT = n - 1 \quad (5)$$

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：

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

k 是回归模型的参数，其中包括截距项。因此，

$$k = D + 1 \quad (7)$$

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：

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

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

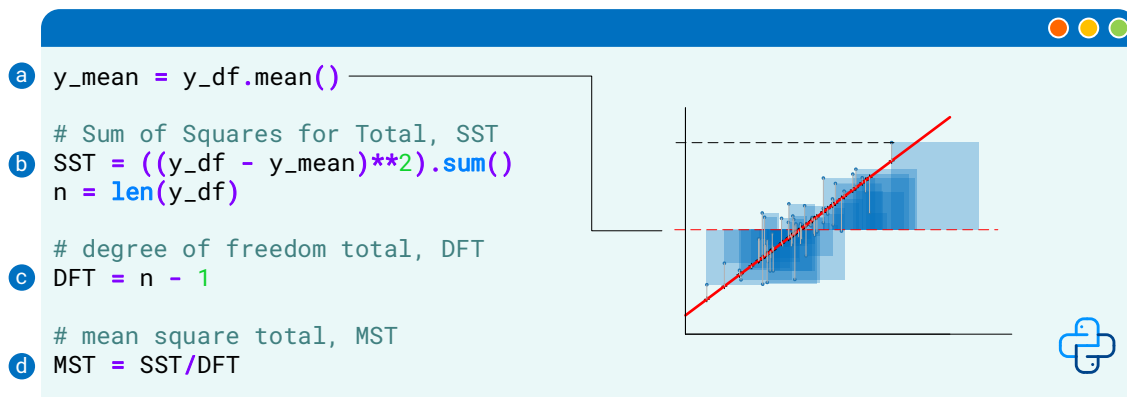
平均总离差 MST

平均总离差 (mean square total, MST) 的定义为：

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

实际上，平均总离差 MST 便是因变量 Y 样本数据方差。

Bk7_Ch02_02.ipynb 还复刻了上述 SST 和 MST 结果，下面聊聊代码 3。



```

a y_mean = y_df.mean()
b # Sum of Squares for Total, SST
  SST = ((y_df - y_mean)**2).sum()
  n = len(y_df)
c # degree of freedom total, DFT
  DFT = n - 1
d # mean square total, MST
  MST = SST/DFT
  
```

代码 3. 计算 SST 和 MST | Bk7_Ch02_02.ipynb

- a 计算期望值 \bar{y} 。
- b 计算总离差平方和 $\text{SST} = \sum_{i=1}^n (y^{(i)} - \bar{y})^2$ 。
- c 计算总离差自由度 $\text{DFT} = n - 1$ 。其中 n 为参与拟合的样本数。
- d 计算平均总离差 $\text{MST} = \text{SST}/\text{DFT}$ 。

2.4 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：

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

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 $|\hat{y}^{(i)} - \bar{y}|$ 。

⚠ 注意，图中所有正方形的一个顶点都在回归直线上。

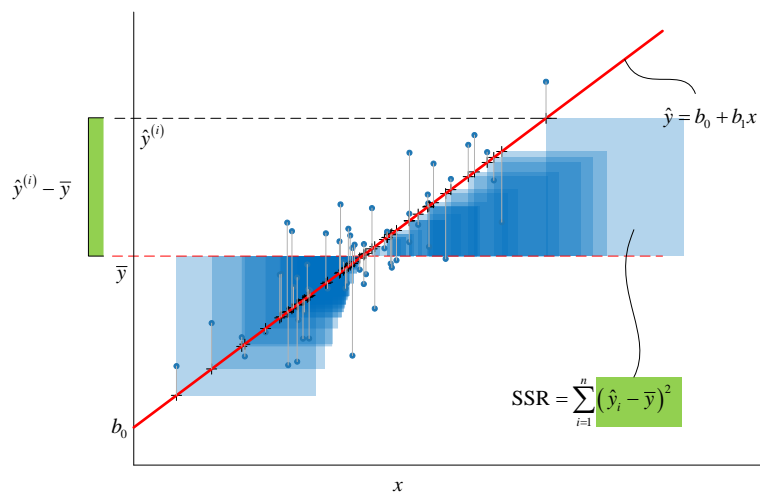


图 9. 回归平方和

回归自由度 DFR

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

$$\text{DFR} = k - 1 = D \quad (11)$$

本例中， $D = 1$ 。

平均回归平方 MSR

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

$$\text{MSR} = \frac{\text{SSR}}{\text{DFR}} = \frac{\text{SSR}}{k - 1} = \frac{\text{SSR}}{D} \quad (12)$$

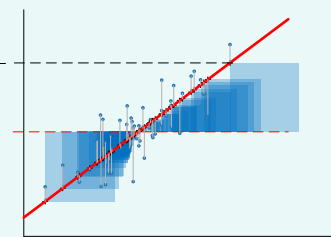
代码 4 计算 SSR 和 MSR。

```
# predicted
a y_hat = results.fittedvalues
  y_hat = y_hat.to_frame()
  y_hat = y_hat.rename(columns={0: 'AAPL'})

# Sum of Squares for Regression, SSR
b SSR = ((y_hat - y_mean)**2).sum()

# degrees of freedom for regression model
c DFR = 1

# MSR: mean square regression
d MSR = SSR/DFR
```



代码 4. 计算 SSR 和 MSR | Bk7_Ch02_02.ipynb

- a 从 `results` 中获取预测值 $\hat{y}^{(i)}$ 。预测值 $\hat{y}^{(i)}$ 在图中红线上。
- b 计算回归平方和 $SSR = \sum_{i=1}^n (\hat{y}^{(i)} - \bar{y})^2$ 。
- c 设定回归自由度 DFR (因变量数量)，本例中 $DFR = 1$ 。
- d 计算平均回归平方 $MSR = SSR/DFR$ 。

2.5 残差平方和 SSE

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

$$SSE = \sum_{i=1}^n (\varepsilon^{(i)})^2 = \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2 \quad (13)$$

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 $|y^{(i)} - \hat{y}^{(i)}|$ 。对于 OLS 一元线性回归，我们期待图中蓝色正方形面积之和最小。

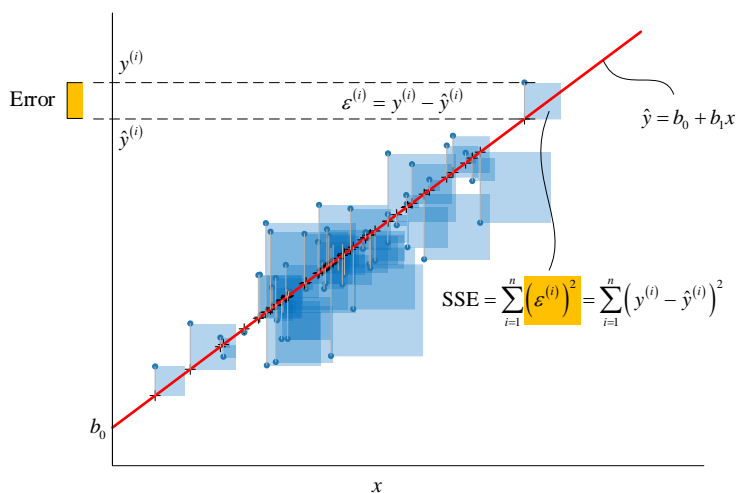


图 10. 残差平方和 SSE

残差自由度 DFE

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

$$DFE = n - k = n - D - 1 \quad (14)$$

残差平均值 MSE

残差平均值 (mean squared error, MSE) 为：

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

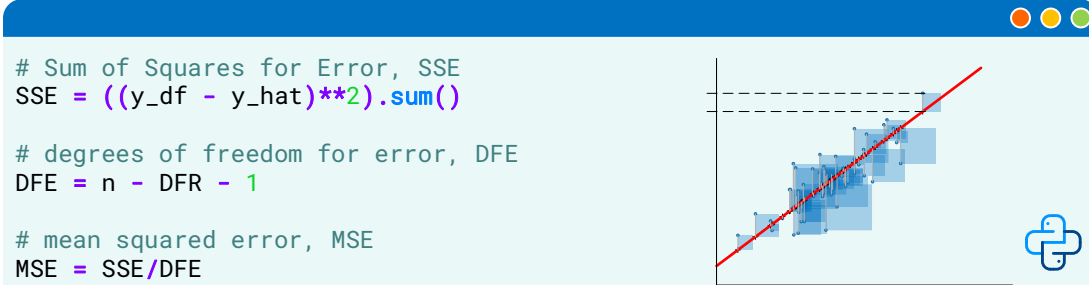
$$\text{MSE} = \frac{\text{SSE}}{\text{DFE}} = \frac{\text{SSE}}{n-k} = \frac{\text{SSE}}{n-D-1} \quad (15)$$

均方根残差 RMSE

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

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{\text{DFE}}} = \sqrt{\frac{\text{SSE}}{n-p}} = \sqrt{\frac{\text{SSE}}{n-D-1}} \quad (16)$$

代码 5 计算 SSE 和 MSE，请大家自行分析这几句。



```

a # Sum of Squares for Error, SSE
SSE = ((y_df - y_hat)**2).sum()

b # degrees of freedom for error, DFE
DFE = n - DFR - 1

c # mean squared error, MSE
MSE = SSE/DFE
  
```

代码 5. 计算 SSE 和 MSE | Bk7_Ch02_02.ipynb

2.6 几何视角：勾股定理

大家别忘了《矩阵力量》反复提到的线性回归几何视角！

一个直角三角形

看到 (2) 中三个求和，我们下面用向量范数算式完成三个求和运算：

$$\begin{aligned} \text{SST} &= \sum_{i=1}^n (y^{(i)} - \bar{y})^2 = \|\mathbf{y} - \bar{y}\mathbf{I}\|_2^2 \\ \text{SSR} &= \sum_{i=1}^n (\hat{y}^{(i)} - \bar{y})^2 = \|\hat{\mathbf{y}} - \bar{y}\mathbf{I}\|_2^2 \\ \text{SSE} &= \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 \end{aligned} \quad (17)$$

根据 (2)，我们可以得到如下等式：

$$\underbrace{\|\mathbf{y} - \bar{y}\mathbf{I}\|_2^2}_{\text{SST}} = \underbrace{\|\hat{\mathbf{y}} - \bar{y}\mathbf{I}\|_2^2}_{\text{SSR}} + \underbrace{\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2}_{\text{SSE}} \quad (18)$$

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, $y - \bar{y}I$ 就是斜边对应的向量, 斜边长度为 $\|y - \bar{y}I\|$ 。 $\hat{y} - \bar{y}I$ 为第一条直角边, $\hat{y} - \bar{y}I$ 代表回归模型解释的部分。 $y - \hat{y}$ 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 $y - \bar{y}I$ 和 $\hat{y} - \bar{y}I$ 的起点为 $\bar{y}I$ 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:

$$(\sqrt{SST})^2 = (\sqrt{SSR})^2 + (\sqrt{SSE})^2 \quad (19)$$

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

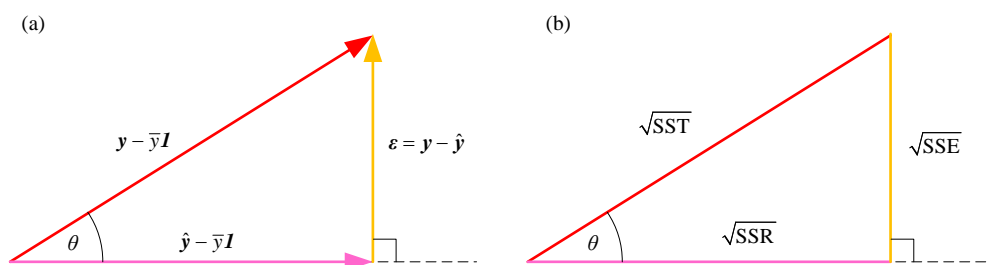


图 11. 几何角度看三个平方和

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 (灰色底色)。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。

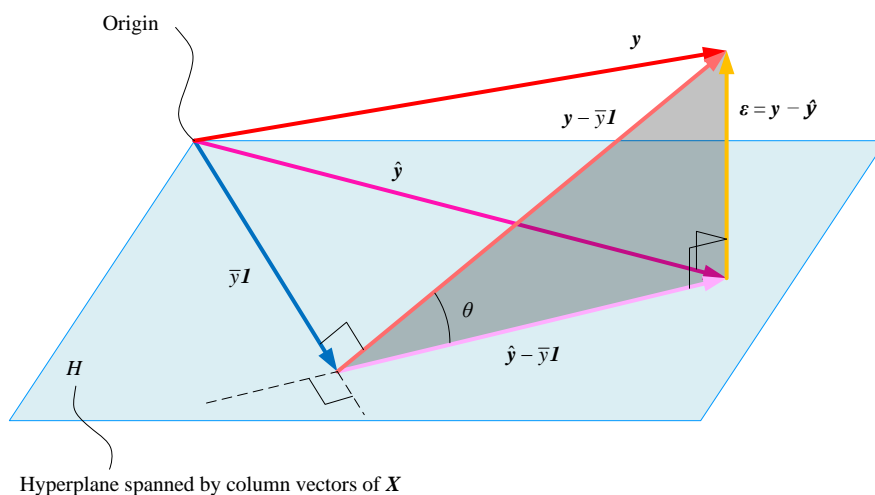


图 12. 四面体的四个面都是直角三角形

2.7 拟合优度：评价拟合程度

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

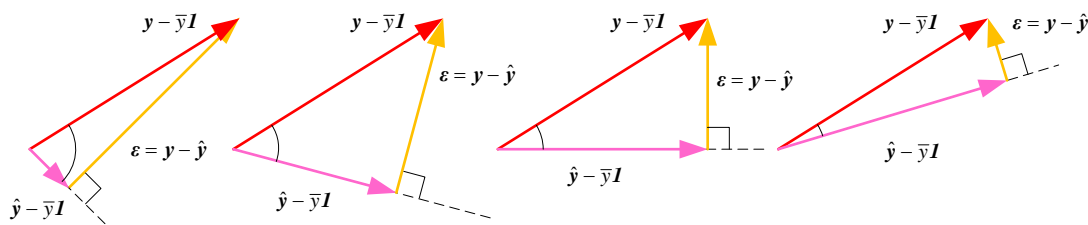


图 13. 因变量向量和预测值向量夹角从大到小

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)。拟合优度是指一个统计模型与观测数据之间的拟合程度，即模型能够多好地解释数据。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) 是量化反映模型拟合优度的统计量。从几何角度来看， R^2 是图 12 中 θ 余弦值 $\cos\theta$ 的平方：

$$R^2 = \cos^2(\theta) \quad (20)$$

利用图 11 (b) 直角三角形三边之间的关系， R^2 可以整理为：

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (21)$$

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。拟合优度可以帮助评估回归模型的可靠性和预测能力，并对模型进行改进和优化。

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。

$$R^2 = \rho_{x,y}^2 = \left(b_1 \frac{\sigma_x}{\sigma_y} \right)^2 \quad (22)$$

其中，

$$b_1 = \rho_{x,y} \frac{\sigma_y}{\sigma_x} \quad (23)$$

也就是说，在一元线性回归中， R^2 的平方根等于线性相关系数的绝对值。也就是说，当 ρ 等于 1 或 -1 时， R^2 为 1，表示因变量完全由自变量解释；当 ρ 等于 0 时， R^2 为 0，表示自变量对因变量没有任何

解释能力。因此， R^2 越接近 1，表示自变量对因变量的解释能力越强，线性相关系数 ρ 的绝对值也越大，反之亦然。

因此，线性相关系数 ρ 和决定系数 R^2 都是衡量变量之间线性关系强弱的重要指标，它们可以帮助我们理解自变量对因变量的解释能力，评估模型的拟合优度，以及选择最佳的回归模型。

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)。简单来说，修正决定系数考虑到自变量的数目对决定系数的影响，避免了当自变量数量增加时决定系数的人为提高。修正决定系数的具体定义为：

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

修正决定系数的作用在于，当模型中自变量的数量增加时，它能够惩罚**过拟合** (overfitting)，并避免了决定系数因为自变量个数增加而提高的问题。因此，在比较不同模型的拟合优度时，使用修正决定系数会更加准确，能够更好地刻画模型的解释能力。

过拟合是指一个模型在训练数据上表现良好，但在测试数据上表现较差的现象。在过拟合的情况下，模型过度地学习了训练数据的特征和噪声，导致其在测试数据上的预测能力下降。

过拟合通常发生在模型复杂度过高或者训练数据太少的情况下。例如，在一元线性回归中，如果使用高次多项式来拟合数据，就容易出现过拟合的情况。在这种情况下，模型会过度拟合训练数据，导致其在新数据上的预测能力下降。

为了避免过拟合，可以采取以下方法：增加训练数据量、降低模型复杂度、采用**正则化** (regularization) 技术等。



本书第 5 章将讲解正则化回归。


接着前文代码，代码 6 计算决定系数和修正决定系数。再次强调，虽然 Statsmodels 的回归函数已经帮助我们计算得到了这些回归分析结果；但是，仍然强烈建议大家知道这些结果背后的数学工具。而且本书还格外建议大家利用多视角 (比如，几何、数据、优化等等) 来理解这些算法。

```

# 计算决定系数
a R2 = SSR/SST

# 计算修正决定系数
b R2_adj = 1 - MSE/MST

```

代码 6. 计算决定系数和修正决定系数 |  Bk7_Ch02_02.ipynb

2.8 F 检验：模型参数不全为 0

在线性回归中， F 检验用于检验线性回归模型参数是否显著。它通过比较回归平方和和残差平方和的大小来判断模型是否具有显著的解释能力。

统计量

F 检验的统计量为：

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

代码 7 展示如何计算 F 检验统计量，并验证 Statsmodels 回归分析结果。

```

# 计算F检验的统计量
a F_test = MSR/MSE
print(F_test)

# 验算F检验的统计量
b N = results.nobs
c k = results.df_model+1
d dfm, dfe = k-1, N - k
e F = results.mse_model / results.mse_resid
print(F)

```

代码 7. F 检验的统计量 |  Bk7_Ch02_02.ipynb

原假设、备择假设

假设检验 (hypothesis testing) 是统计学中常用的一种方法，用于根据样本数据推断总体参数是否符合某种假设。假设检验通常包括两个假设：原假设和备择假设。

原假设 (null hypothesis) 是指在实验或调查中假设成立的一个假设，通常认为其成立。

备择假设 (alternative hypothesis) 是指当原假设不成立时，我们希望成立的另一个假设。

通过收集样本数据，并根据统计学原理计算出样本统计量的概率分布，我们可以计算出拒绝原假设的概率。如果这个概率小于预设的显著性水平（比如 0.05），就可以拒绝原假设，认为备择假设成立。反之，如果这个概率大于预设的显著性水平，就不能拒绝原假设。

F 检验是单尾检验，原假设 H_0 、备择假设 H_1 分别为：

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

具体来说， F 检验的零假设是模型的所有回归系数都等于零，即自变量对因变量没有显著的影响。如果 F 检验的 p 值小于设定的显著性水平，就可以拒绝零假设，认为模型是显著的，即自变量对因变量有显著的影响。

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 $n-k$) 以及显著性水平 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，

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

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，

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

如果 p 值小于 α ，则可以拒绝零假设 H_0 。



Bk7_Ch02_02.ipynb 计算图 6 所示方差分析表格中统计量。

2.9 t 检验：某个回归系数是否为 0

在线性回归中， t 检验主要用于检验线性回归模型中某个特定自变量的系数是否显著。具体地， t 检验的零假设是特定回归系数等于零，即自变量对因变量没有显著的影响。如果 t 检验的 p 值小于设定的显著性水平，就可以拒绝零假设，认为该自变量的系数是显著不为零的，即自变量对因变量有显著的影响。

需要注意的是， t 检验一般用来检验一个特定自变量的系数是否显著，而不能判断模型整体是否显著。如果需要判断模型整体的显著性，可以使用前文介绍的 F 检验。

原假设、备择假设

对于一元线性回归， t 检验原假设和备择假设分别为：

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

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。

斜率系数

b_1 的 t 检验统计量：

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

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

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

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 (除去 NaN)。标准误越大，回归系数的估计值越不可靠。

代码 8 计算 b_1 的 t 检验统计量。请大家对照 (30) 和 (31)，逐句分析代码。

```

a MSE = SSE/DFE
MSE = MSE.values
# 计算MSE

b b1 = p.SP500
# 斜率系数

c SSD_x = np.sum((x_df.values - x_mean)**2)

d SE_b1 = np.sqrt(MSE/SSD_x)
# 标准误

e T_b1 = (b1 - 0)/SE_b1
# b1的t检验统计量

```

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

$$t_{b1} = \frac{\hat{b}_1 - b_{1,0}}{SE(\hat{b}_1)}$$

代码 8. b_1 的 t 检验统计量 | Bk7_Ch02_03.ipynb

临界值

如果下式成立，接受零假设 H_0 ：

$$-t_{1-\alpha/2, n-2} < T < t_{1-\alpha/2, n-2} \quad (32)$$

否则，则拒绝零假设 H_0 。

特别地，如果原假设和备择假设为：

$$\begin{cases} H_0 : b_1 = 0 \\ H_1 : b_1 \neq 0 \end{cases} \quad (33)$$

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。

截距项系数

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

$$t_{b0} = \frac{\hat{b}_0 - b_{0,0}}{SE(\hat{b}_0)} \quad (34)$$

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

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

请大家对照 (34) 和 (35)，逐句学习代码 9。

```

a b0 = p.const
  # 截距系数

b SE_b0 = np.sqrt(MSE*(1/n + x_mean**2/SSD_x))
  # 标准误

c T_b0 = (b0 - 0)/SE_b0
  # b0的t检验统计量

```

代码 9. b_0 的 t 检验统计量 | Bk7_Ch02_03.ipynb

举个例子

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。在统计学中，双尾假设检验是指在假设检验过程中，假设被拒绝的区域位于一个统计量分布的两个尾端，即研究者对于一个参数或者统计量是否等于某一特定值，不确定其比该值大或小，而是存在两种可能性，因此需要在两个尾端进行检验。

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：

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

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：

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

如图 1 所示， $t_{b1} = 23.446$ ，因此：

$$t_{b1} > t_{0.975, 250} \quad (38)$$

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：

$$t_{b0} < t_{0.975, 250} \quad (39)$$

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：

$$\hat{b}_1 \pm t_{1-\alpha/2, n-2} \cdot \text{SE}(\hat{b}_1) \quad (40)$$

这个置信区间的含义是，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：

$$\hat{b}_0 \pm t_{1-\alpha/2, n-2} \cdot \text{SE}(\hat{b}_0) \quad (41)$$

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

同理，真实 b_0 在以上区间的概率为 $1 - \alpha$ 。

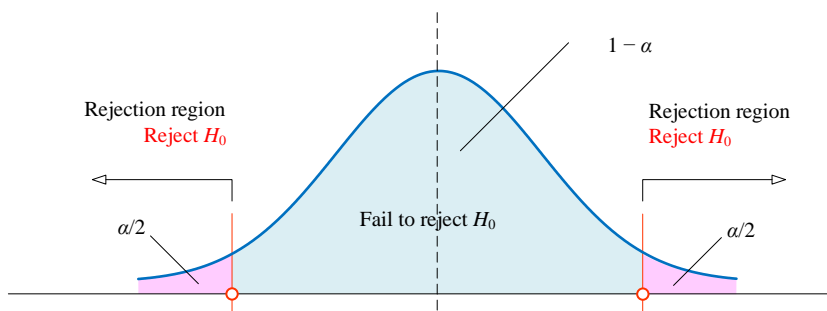


图 14. 双尾检验

请大家对照 (40) 和 (41) 查看代码 10。注意，**b** 中 `stats.t.ppf()` 是 SciPy 库中的一个函数，用于计算 t 分布的百分点函数 (percent point function)。其中， $1 - \alpha/2$ 为置信水累积分布函数 CDF，DFE 为自由度。因此，这句的目的是 t 分布中给定置信水平 ($1 - \alpha$) 和自由度条件下的双尾置信区间的边界值。

请大家对比代码 10 结果和图 1。

```

a alpha = 0.05
# 显著水平
b t_95 = stats.t.ppf(1 - alpha/2, DFE)
# t值

# 系数b1的1 - alpha 置信区间
c b1_upper_95 = b1 + t_95*SE_b1
  print(b1_upper_95)
  b1_lower_95 = b1 - t_95*SE_b1
  print(b1_lower_95)

# 系数b0的1 - alpha 置信区间
d b0_upper_95 = b0 + t_95*SE_b0
  print(b0_upper_95)
  b0_lower_95 = b0 - t_95*SE_b0
  print(b0_lower_95)

```

代码 10. b_0 和 b_1 的置信区间 | Bk7_Ch02_03.ipynb

2.10 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：

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

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

置信区间的宽度为：

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

随着 $|x^{(i)} - \bar{x}|$ 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。在回归分析中，预测值置信区间用于评估回归模型的预测能力。通常，预测值的置信区间越窄，说明模型预测的精度越高。

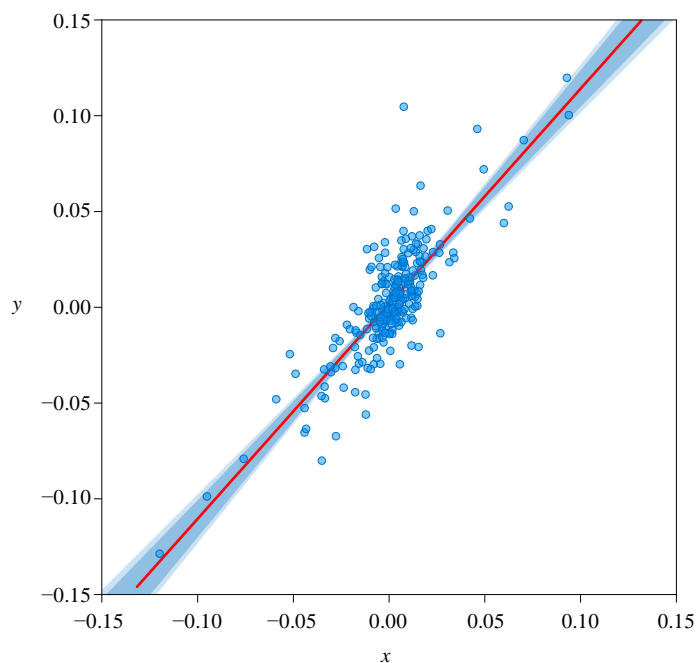



图 15. 一元线性回归线置信区间 (95% 和 99%) |  Bk7_Ch02_03.ipynb

2.11 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：

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

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

与预测值的置信区间不同，预测区间同时考虑了预测的误差和未来观测值的随机性。

预测区间包含两个方面的误差：回归方程中的估计误差和对未来观测值的随机误差。与预测值的置信区间不同，预测区间考虑了未来观测值的随机性，因此通常比置信区间更宽。

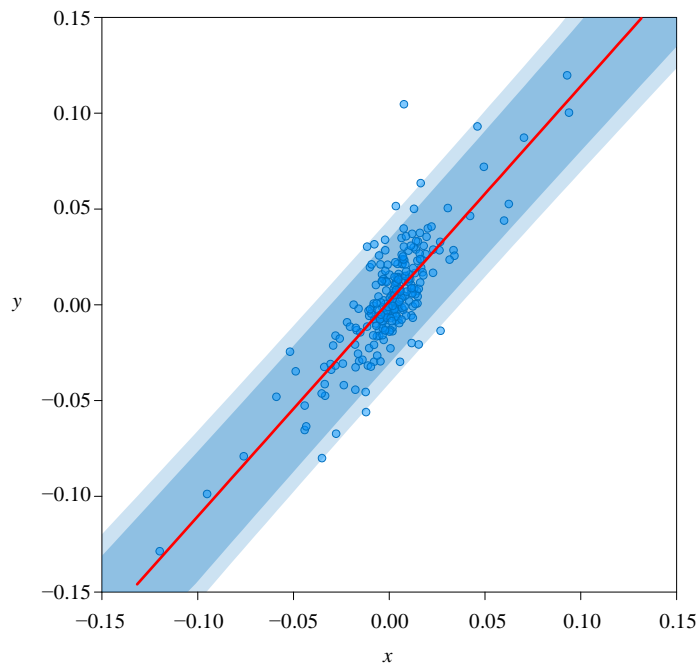


图 16. 一元线性回归线预测区间 |  Bk7_Ch02_03.ipynb

2.12 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：

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

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

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

似然函数为：

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

常用对数似然 $\ln(L)$ ：

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

$$\ln(L) = -\frac{n}{2} \cdot \ln(2\pi\sigma^2) - \frac{\text{SSE}}{2\sigma^2} \quad (48)$$

注意，MLE 中的 σ 为：

$$\sigma^2 = \frac{\text{SSE}}{n} \quad (49)$$

这样 $\ln(L)$ 可以写成：

$$\ln(L) = -\frac{n}{2} \cdot \ln(2\pi\sigma^2) - \frac{n}{2} \quad (50)$$



有关似然函数和对数似然函数，请大家回顾《统计至简》第 16、24 章。

2.13 信息准则：选择模型的标准

AIC 和 BIC 是线性回归模型选择中常用的信息准则，用于在多个模型中选择最优模型。

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

$$\text{AIC} = 2k - 2\ln(L) \quad (51)$$

Penalty

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。

$$\text{BIC} = k \cdot \ln(n) - 2\ln(L) \quad (52)$$

Penalty

其中， n 为样本数据数量。BIC 的惩罚项比 AIC 大。

在使用 AIC 和 BIC 进行模型选择时，应该选择具有最小 AIC 或 BIC 值的模型。这意味着，较小的 AIC 或 BIC 值表示更好的模型拟合和更小的模型复杂度。

需要注意的是，AIC 和 BIC 都是用来选择模型的工具，但并不保证选择的模型就是最优模型。在实际应用中，应该将 AIC 和 BIC 作为指导，结合领域知识和经验来选择最优模型。同时，还需要对模型的假设和限制进行检验，以确保模型的可靠性和实用性。

2.14 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。残差分析的基本思想是，如果回归模型能够很好地拟合数据，那么残差应该是随机分布的，没有明显的模式或趋势。因此，对残差的分布进行检查可以提供关于模型拟合优度的信息。

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

残差分析通常包括以下步骤：

- ▶ 绘制残差图。残差图是观测值的残差与预测值之间的散点图。如果残差呈现出随机分布、没有明显的模式或趋势，那么模型可能具有较好的拟合优度。
- ▶ 检查残差分布。通过绘制残差直方图或核密度图来检查残差分布是否呈现出正态分布或近似正态分布。如果残差分布不是正态分布，那么可能需要采取转换或其他措施来改善模型的拟合。
- ▶ 检查残差对自变量的函数形式。通过绘制残差与自变量之间的散点图或回归曲线，来检查残差是否随自变量的变化而呈现出系统性变化。如果存在这种关系，那么可能需要考虑增加自变量、采取变量转换等方法来改善模型的拟合。

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，本节利用 Omnibus 正态检验。

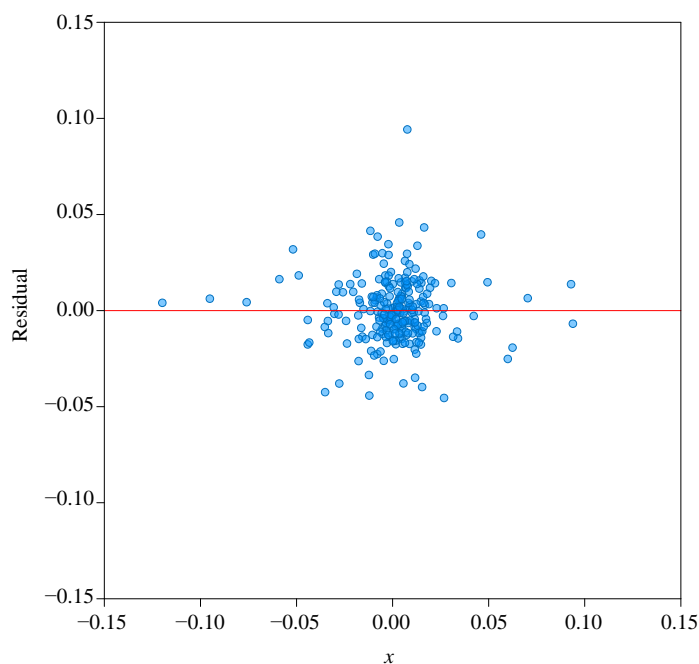


图 17. 残差散点图

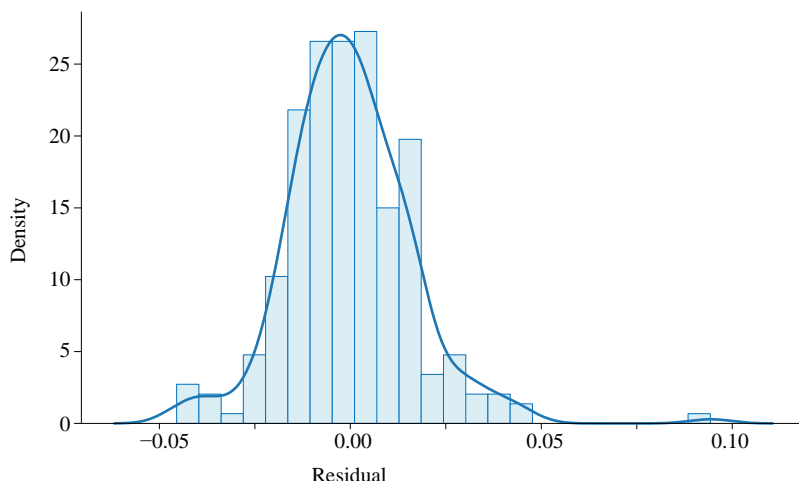


图 18. 残差分布直方图

Omnibus 正态检验 (Omnibus test for normality) 用于检验线性回归中残差是否服从正态分布。Omnibus 正态检验利用残差的偏度 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 χ^2 检验 (Chi-squared test)。

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



《统计至简》第 2 章讲过偏度、峰度，请大家回顾。

此外，**加权最小二乘法** (Weighted Least Squares, WLS) 是最小二乘法 OLS 的一种扩展形式。加权最小二乘法引入了权重因子，用于调整每个数据点的相对重要性。

使用加权最小二乘法的场景包括异方差性 (heteroscedasticity)。当数据的方差不是恒定的时候，可以使用加权最小二乘法 WLS 来降低方差不稳定性的影响。

对于可能是异常值的数据点，可以通过 WLS 降低其权重来减少其对拟合的影响。

总的来说，加权最小二乘法是在最小二乘法的基础上考虑了不同数据点的权重，使得拟合更加灵活。Statsmodels 中有专门处理加权最小二乘法的工具，请大家参考。

https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.WLS.html

2.15 自相关检测：Durbin-Watson

Durbin-Watson 用于检验序列的自相关。在线性回归中，**自相关** (autocorrelation) 用来分析模型中的残差与其在时间上的延迟版本之间的相关性。当模型中存在自相关时，它可能表明模型中遗漏了某些重要的变量，或者模型中的时间序列数据未被正确处理。

自相关可以通过检查残差图来诊断。如果残差图表现出明显的模式，例如残差值之间存在周期性关系或呈现出聚集在某个区域的情况，那么就可能存在自相关。在这种情况下，可以通过引入更多的自变量或使用时间序列分析方法来修正模型。图 19 所示为残差的自相关图。

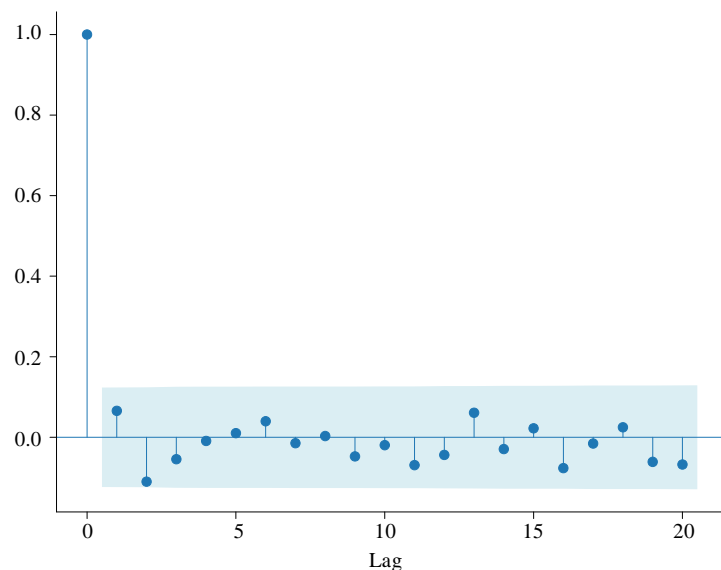


图 19. 残差自相关

Durbin-Watson 检测的统计量为：

$$DW = \frac{\sum_{i=2}^n \left((y^{(i)} - \hat{y}^{(i)}) - (y^{(i-1)} - \hat{y}^{(i-1)}) \right)^2}{\sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2} \quad (53)$$

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 $0 \sim 4$ 。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。

有关，请大家参考：

https://www.statsmodels.org/devel/generated/statsmodels.stats.stattools.durbin_watson.html

2.16 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 $\mathbf{X}_{k \times k}$ 是否存在**多重共线性** (multicollinearity)。

多重共线性是指在多元回归模型中，独立变量之间存在高度相关或线性关系的情况。多重共线性会导致回归系数的估计不稳定，使得模型的解释能力降低，甚至导致模型的预测精度下降。

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：

$$\text{condition number} = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} \quad (54)$$

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

在实际应用中，如果 $X^T X$ 的条件数过大，可以考虑采用特征缩放或正则化来改善。

下一章讲到多元回归分析时，条件数的作用更明显。



Bk7_Ch02_03.ipynb 代码复刻图 1 中除 ANOVA 以外的其他统计量值。



线性回归是一种用于研究自变量与因变量之间关系的统计模型。方差分析可以评估模型的整体拟合优度，其中的 F 检验可以用来线性模型参数整体显著性， t 检验可以评估单个系数的显著性。拟合优度指模型能够解释数据变异的比列，常用 R^2 来度量。AIC 和 BIC 用于模型选择，可以在模型拟合度相似的情况下，选出最简单和最有解释力的模型。自相关指误差项之间的相关性，可以使用 Durbin-Watson 检验进行检测。条件数是用于评估多重共线性的指标，如果条件数过大，可能存在严重的多重共线性问题。

综上，这些概念是线性回归分析中非常重要的指标，可以帮助我们评估模型的拟合程度、系数显著性、预测能力和多重共线性等问题。这一章的内容很有难度，现在不要求大家掌握所有的知识点。



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