

02

Regression Analysis

回归分析

线性回归结果不能拿来就用



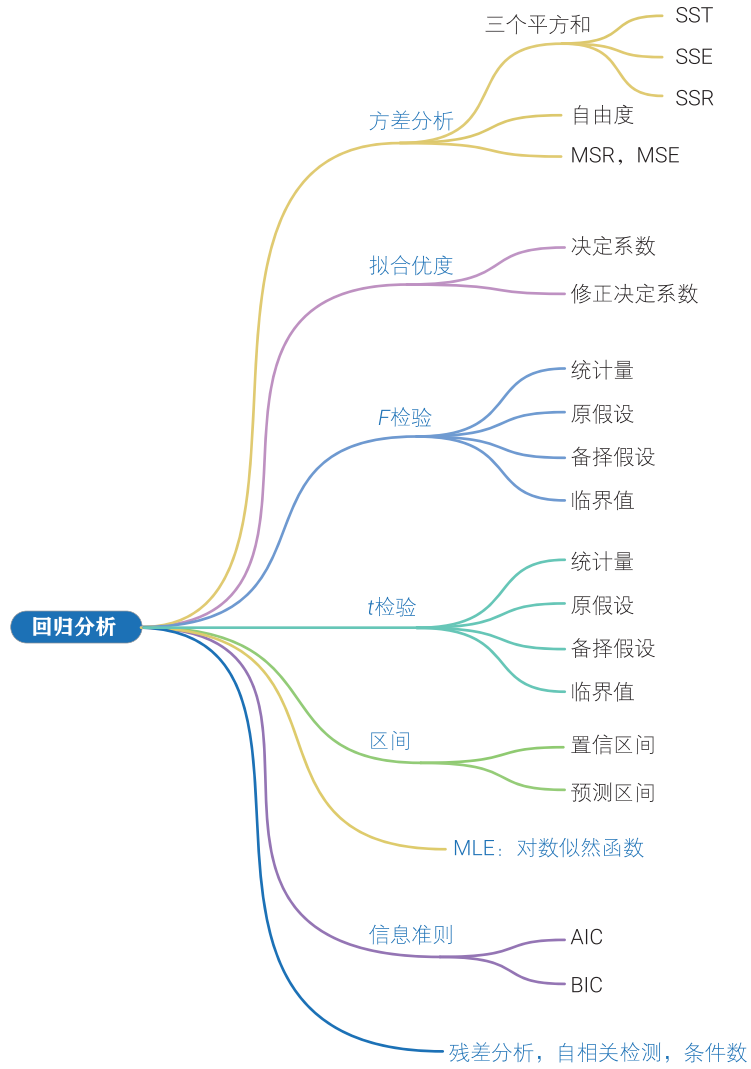
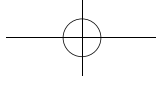
真理太复杂了，除了近似，我们别无他法。

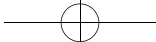
Truth is much too complicated to allow anything but approximations.

—— 约翰·冯·诺依曼 (John von Neumann) | 美国籍数学家 | 1903 — 1957年



- ◀ `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





2.1 线性回归：一个表格、一条直线

一个表格

表2.1汇总了某个线性回归分析的结果。本章的主要目的就是让大家理解这个表格各项数值的含义。

大家是否还记得我们在《统计至简》第24章结尾给出过表2.1。

表2.1 一元线性回归结果 | [Bk7_Ch02_01.ipynb](#)

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

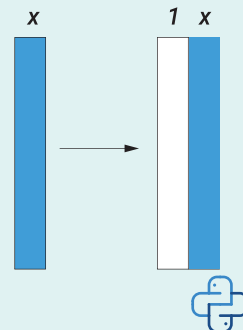
Bk7_Ch02_01.ipynb绘制了本节图像。下面，让我们一起简单聊聊其中关键语句。

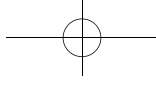
我们可以通过代码2.1下载、处理数据，并完成回归运算。

代码2.1 下载、处理数据并完成回归运算 | [Bk7_Ch02_01.ipynb](#)

```
a y_x_df = yf.download(['AAPL', '^GSPC'],
                        start = '2020-01-01',
                        end = '2020-12-31', )
b y_x_df = y_x_df['Adj Close'].pct_change()
c y_x_df.dropna(inplace = True)
d y_x_df.rename(columns = { '^GSPC': 'SP500' }, inplace = True)
e x_df = y_x_df[['SP500']]
f y_df = y_x_df[['AAPL']]

# 增加一列全1
g X_df = sm.add_constant(x_df)
h model = sm.OLS(y_df, X_df)
i results = model.fit()
j print(results.summary())
```





a 用`yfinance.download`，简写作`yf.download`，下载金融数据。在此之前，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` 中。

此外，如果大家下载数据遇到困难，还要用`to_csv()` 和`to_pickle()` 将数据保存下来。而且，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，也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。设置`inplace=True`后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，设置`inplace=True`后会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自变量和因变量散点数据。

g 用`statsmodels.api`，简写作`sm`，其中的`add_constant()` 给数据帧`x_df`增加一列全1列。目的是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OLS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2.1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 (以下简称股指) 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。表2.1是利用`statsmodels.api.OLS()` 函数构造的线性模型结果。图2.2所示为用`seaborn.jointplot()` 绘制的回归图，并且其中绘制了边际分布。

! 再次强调，线性回归不代表“因果关系”。

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。

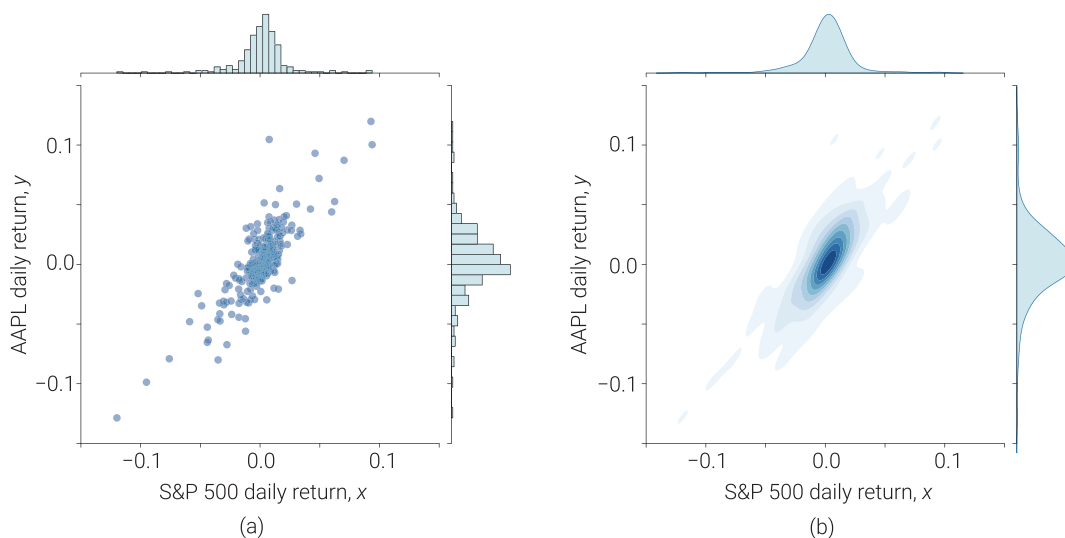


图2.1 日收益率数据关系 | [Bk7_Ch02_01.ipynb](#)

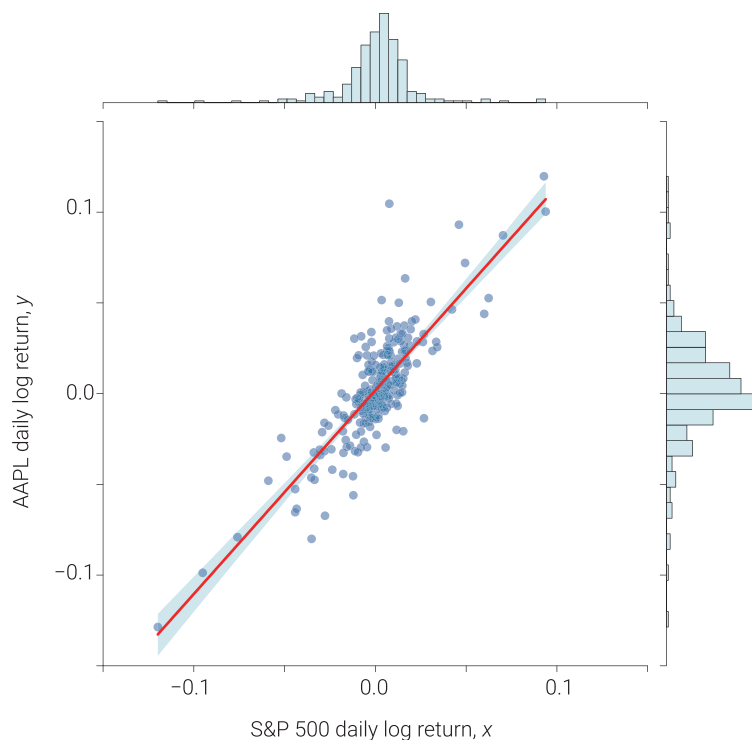
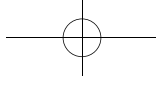


图2.2 用`seaborn.jointplot()` 绘制回归直线 | [Bk7_Ch02_01.ipynb](#)

统计特征

图2.3(a) 所示为数据的协方差矩阵。

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2.3(b) 所示为相关性系数矩阵热图。

《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。

《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2.3(c) 所示。

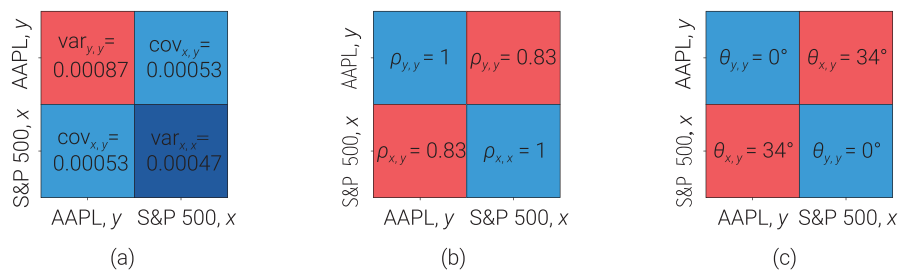


图2.3 $[y, x]$ 数据的协方差矩阵、相关性和夹角热图 | [Bk7_Ch02_01.ipynb](#)

图2.4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 能更充分地解释因变量向量 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，建议大家主要理解方差分析(ANOVA)，并且在有实际需要时再回头查阅本章其余内容。

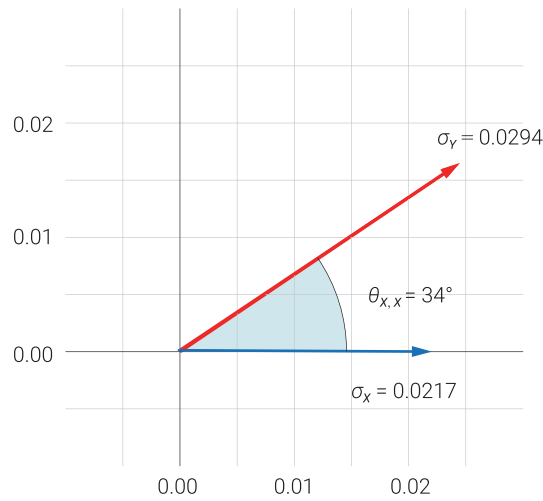
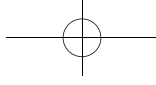


图2.4 标准差向量空间角度解释夹角 | [Bk7_Ch02_01.ipynb](#)

2.2 方差分析 (ANOVA)

本节开始先介绍如何理解表2.2所示的ANOVA表格结果。ANOVA的含义是**方差分析** (Analysis of Variance)。ANOVA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。ANOVA通过比较模型中不同变量的平均方差，来确定哪些变量对目标变量的解释程度更高。

表2.2 一元线性回归ANOVA表格

	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

ANOVA是表2.1的重要组成部分之一。

我们可以通过代码2.2完成ANOVA，下面讲解其中关键语句。

代码2.2 ANOVA | [Bk7_Ch02_02.ipynb](#)

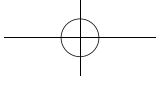
```
from statsmodels.formula.api import ols
from statsmodels.stats.anova import anova_lm

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)
```

a 用pandas.DataFrame() 构造一个数据帧，“x”列中导入SP500数据，“y”列中导入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。“ $y \sim x$ ”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ANOVA。`model_V2` 是线性回归模型的对象，而 `typ=1` 表示使用 “Type I” ANOVA。Type I ANOVA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ANOVA的结果，具体如表2.2所示。

表2.3所示为标准ANOVA表格对应的统计量。标准ANOVA表格比表2.2多一行。此外，表2.3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。

表2.3 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			

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2.3中的每一个值的含义，以及它们和线性回归的关系。

三个平方和

为了解ANOVA表格，我们首先要了解以下三个平方和。

- ◀ **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2.5给出了计算三个平方和所需的数值。表2.4总结了三个平方和的定义。

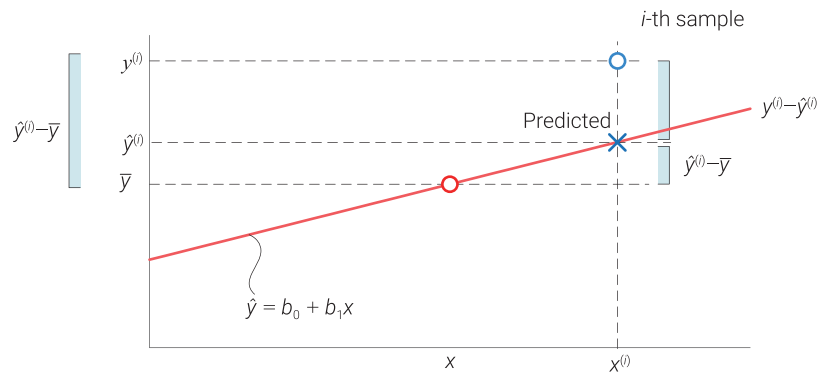
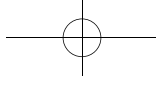


图2.5 通过一元线性回归模型分解因变量的变化

表2.4 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)	$SST = \sum_{i=1}^n (y^{(i)} - \bar{y})^2$	
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)	$SSE = \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2$	

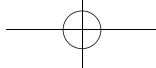
等式关系

对于线性回归来说，ANOVA实际上就是把SST分解成SSE和SSR：

$$SST = SSR + SSE \quad (2.1)$$

即：

$$\underbrace{\sum_{i=1}^n (y^{(i)} - \bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^n (\hat{y}^{(i)} - \bar{y})^2}_{SSR} + \underbrace{\sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2}_{SSE} \quad (2.2)$$



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系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：

$$SST = \sum_{i=1}^n (y^{(i)} - \bar{y})^2 \quad (2.3)$$

其中，期望值 $\bar{y}$ 为：

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y^{(i)} \quad (2.4)$$

如图2.6所示，我们可以把SST看作一系列正方形面积之和。这些正方形的边长为 $|y^{(i)} - \bar{y}|$ 。图2.6中这些正方形的一条边都在期望值 $\bar{y}$ 这个高度上。

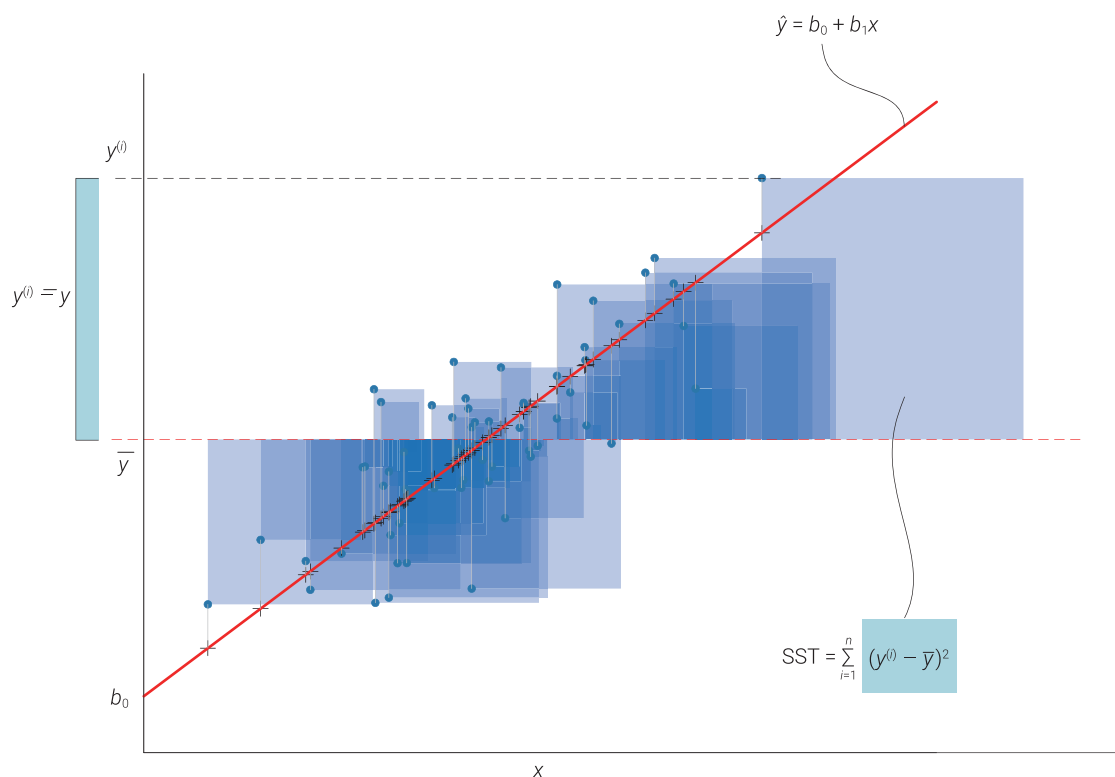
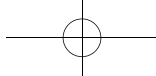


图2.6 总离差平方和(SST)



总离差自由度(DFT)

总离差自由度 (Degree of Freedom Total, DFT) 的定义为:

$$DFT = n - 1 \quad (2.5)$$

n 是样本数据的数量 (NaN除外)。

三个自由度之间的关系

总离差自由度 (DFT)、**回归自由度** (Degree of Freedom for Regression, DFR)、**残差自由度** (Degree of Freedom for Error, DFE) 三者之间的关系为:

$$DFT = n - 1 = DFR + DFE = \underbrace{(k - 1)}_{DFR} + \underbrace{(n - k)}_{DFE} = \underbrace{(D)}_{DFR} + \underbrace{(n - D - 1)}_{DFE} \quad (2.6)$$

k 是回归模型的参数, 其中包括截距项。因此,

$$k = D + 1 \quad (2.7)$$

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:

$$\begin{cases} DFT = 252 - 1 = 251 \\ k = D + 1 = 2 \\ DFR = k - 1 = D = 1 \\ DFE = n - k = n - D - 1 = 252 - 2 = 250 \end{cases} \quad (2.8)$$

平均总离差(MST)

平均总离差 (Mean Square Total, MST) 的定义为:

$$MST = \text{var}(Y) = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1} = \frac{SST}{DFT} \quad (2.9)$$

实际上, 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, 下面讲解代码2.3。

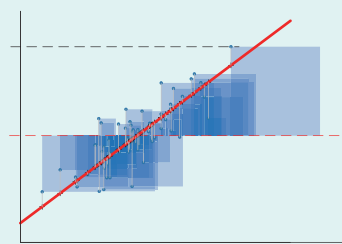
代码2.3 计算SST和MST | Bk7_Ch02_02.ipynb

```
a y_mean = y_df.mean()

# Sum of Squares for Total, SST
b SST = ((y_df - y_mean)**2).sum()
n = len(y_df)

# degree of freedom total, DFT
c DFT = n-1

# mean square total, MST
d MST = SST/DFT
```



- a 计算期望值 $\bar{y}$ 。
- b 计算 $SST = \sum_{i=1}^n (y^{(i)} - \bar{y})^2$ 。
- c 计算 $DFT = n - 1$ ，其中 n 为参与拟合的样本数。
- d 计算 $MST = SST/DFT$ 。

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：

$$SSR = \sum_{i=1}^n (\hat{y}^{(i)} - \bar{y})^2 \quad (2.10)$$

图2.7所示为SSR的几何意义。图2.7中的每个正方形边长为 $|\hat{y}^{(i)} - \bar{y}|$ 。

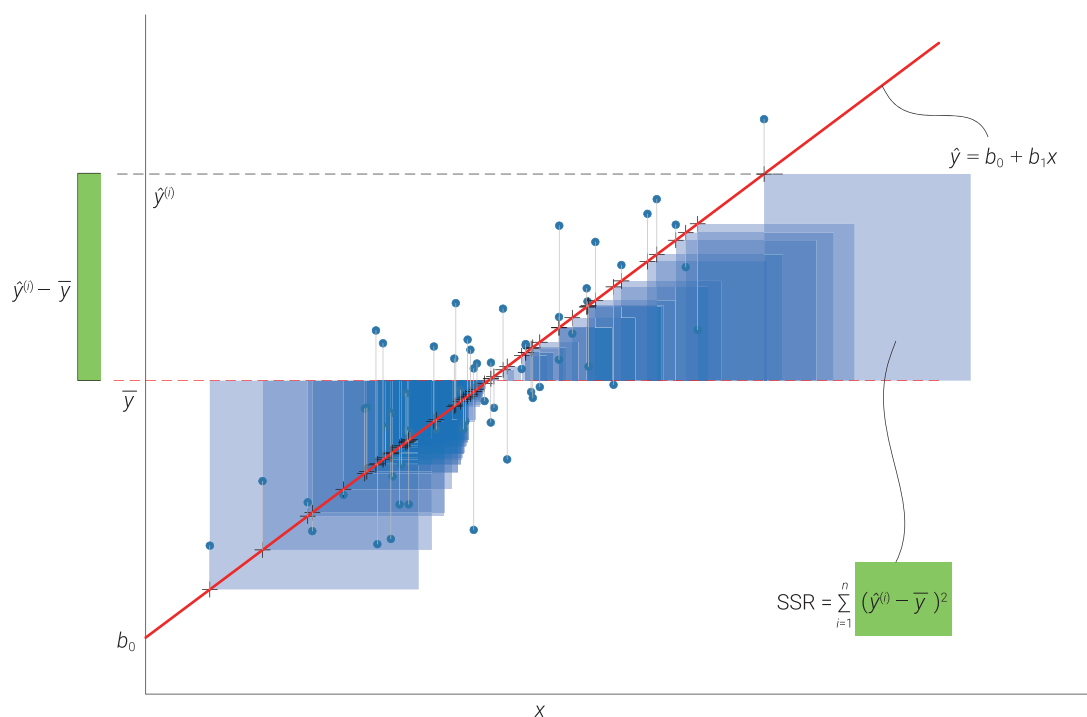


图2.7 回归平方和(SSR)

图中所有正方形的一个顶点都在回归直线上。

回归自由度(DFR)

回归自由度 (Degrees of Freedom for Regression, DFR) 的定义为:

$$DFR = k - 1 = D \quad (2.11)$$

本例中, $D = 1$ 。

平均回归平方(MSR)

平均回归平方 (Mean Square Regression, MSR) 为:

$$MSR = \frac{SSR}{DFR} = \frac{SSR}{k-1} = \frac{SSR}{D} \quad (2.12)$$

我们可以通过代码2.4计算SSR和MSR。

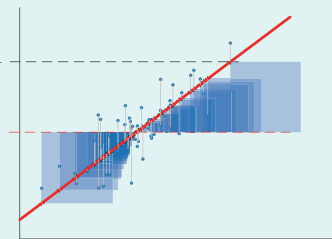
代码2.4 计算SSR和MSR | Bk7_Ch02_02.ipynb

```
# predicted
a y_hat = results.fittedvalues
  y_hat = y_hat.to_frame()
  y_hat = y_hat.rename(columns = {1: 'AAPL'})

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

# Degrees of Freedom for Regression, DFR
c DFR = 1

# Mean Square Regression, MSR
d MSR = SSR/DFR
```



a 从results中获取预测值 $\hat{y}^{(i)}$ 。预测值 $\hat{y}^{(i)}$ 在图中红线上。

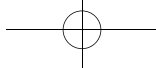
b 计算 $SSR = \sum_{i=1}^n (\hat{y}^{(i)} - \bar{y})^2$ 。

c 设定DFR (因变量数量), 本例中DFR = 1。

d 计算MSR = SSR/DFR。

2.5 残差平方和 (SSE)

残差平方和 (Sum of Squares for Error, SSE) 的定义为:



$$SSE = \sum_{i=1}^n (\varepsilon^{(i)})^2 = \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2 \quad (2.13)$$

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

图2.8所示为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。

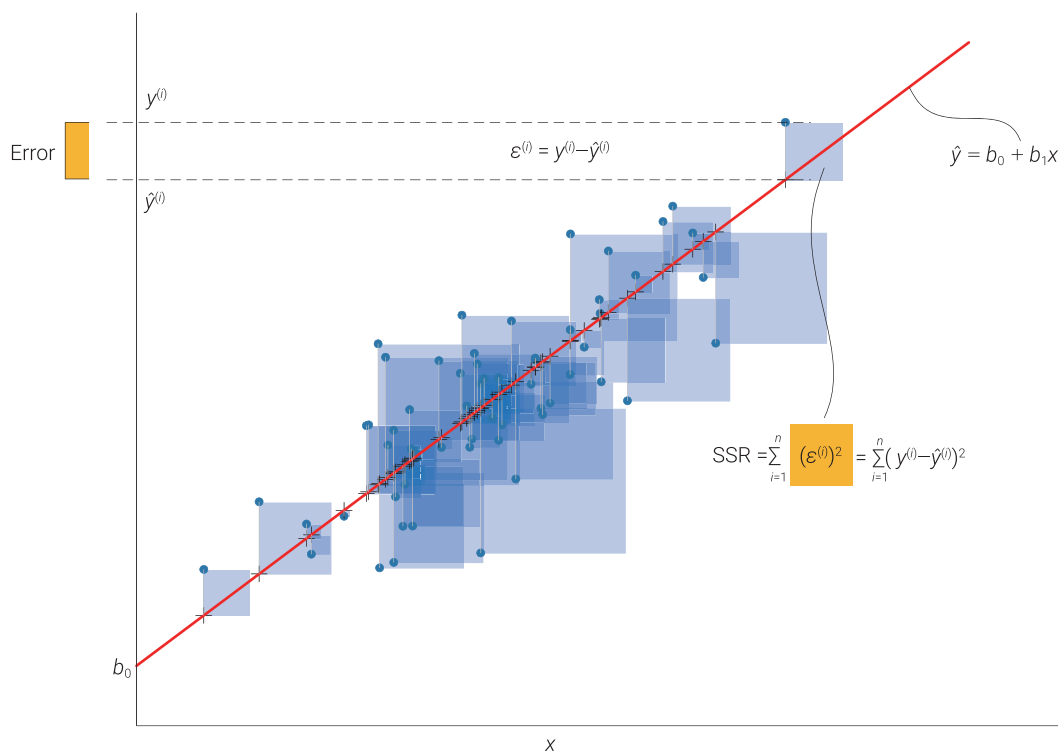


图2.8 残差平方和(SSE)

残差自由度(DFE)

残差自由度 (Degrees of Freedom for Error, DFE) 的定义为：

$$DFE = n - k = n - D - 1 \quad (2.14)$$

残差平均值(MSE)

残差平均值 (Mean Squared Error, MSE) 的定义为：

$$MSE = \frac{SSE}{DFE} = \frac{SSE}{n - k} = \frac{SSE}{n - D - 1} \quad (2.15)$$

均方根残差(RMSE)

均方根残差 (Root Mean Square Error, RMSE) 为MSE的平方根:

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{\text{SSE}}{\text{DFE}}} = \sqrt{\frac{\text{SSE}}{n-k}} = \sqrt{\frac{\text{SSE}}{n-D-1}} \quad (2.16)$$

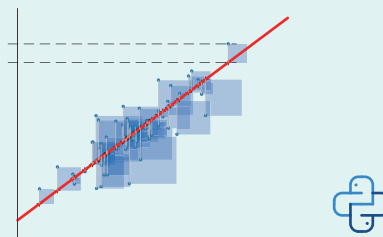
我们可以通过代码2.5计算SSE和MSE, 请大家自行分析这几句。

代码2.5 计算SSE和MSE | Bk7_Ch02_02.ipynb

```
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
```



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

下面我们用向量范数算式完成式(2.2)中的三个求和运算:

$$\begin{aligned} \text{SST} &= \sum_{i=1}^n (y^{(i)} - \bar{y})^2 = \|\mathbf{y} - \bar{y}\mathbf{1}\|_2^2 \\ \text{SSR} &= \sum_{i=1}^n (\hat{y}^{(i)} - \bar{y})^2 = \|\hat{\mathbf{y}} - \bar{y}\mathbf{1}\|_2^2 \\ \text{SSE} &= \sum_{i=1}^n (y^{(i)} - \hat{y}^{(i)})^2 = \|\mathbf{y} - \hat{\mathbf{y}}\|_2^2 \end{aligned} \quad (2.17)$$

根据式2.2, 我们可以得到如下等式:

$$\underbrace{\|\mathbf{y} - \bar{y}\mathbf{1}\|_2^2}_{\text{SST}} = \underbrace{\|\hat{\mathbf{y}} - \bar{y}\mathbf{1}\|_2^2}_{\text{SSR}} + \underbrace{\|\mathbf{y} - \hat{\mathbf{y}}\|_2^2}_{\text{SSE}} \quad (2.18)$$

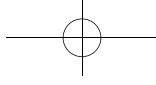


注意: 图2.9中 $\mathbf{y} - \bar{y}\mathbf{1}$ 和 $\hat{\mathbf{y}} - \bar{y}\mathbf{1}$ 的起点为 $\bar{y}\mathbf{1}$ 的终点, 这相当于去均值。

相信大家一眼就会看出来, 式(2.18)代表着直角三角形的勾股定理!

如图2.9(a)所示, $\mathbf{y} - \bar{y}\mathbf{1}$ 就是斜边对应的向量, 斜边长度为 $\|\mathbf{y} - \bar{y}\mathbf{1}\|$ 。 $\hat{\mathbf{y}} - \bar{y}\mathbf{1}$ 为第一条直角边, $\hat{\mathbf{y}} - \bar{y}\mathbf{1}$ 代表回归模型解释的部分。 $\mathbf{y} - \hat{\mathbf{y}}$ 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。

如图2.9(b)所示, 勾股定理还可以写成:



$$(\sqrt{\text{SST}})^2 = (\sqrt{\text{SSR}})^2 + (\sqrt{\text{SSE}})^2 \quad (2.19)$$

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

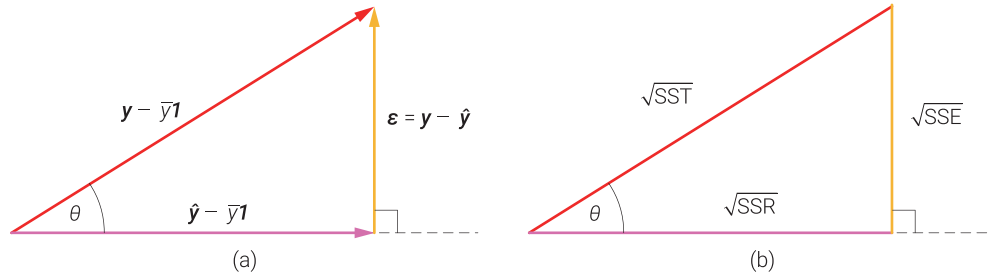


图2.9 几何视角看三个平方和

四个直角三角形

图2.9的直角三角形是图2.10中四面体的一个面 (灰色底色)。而图2.10中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。

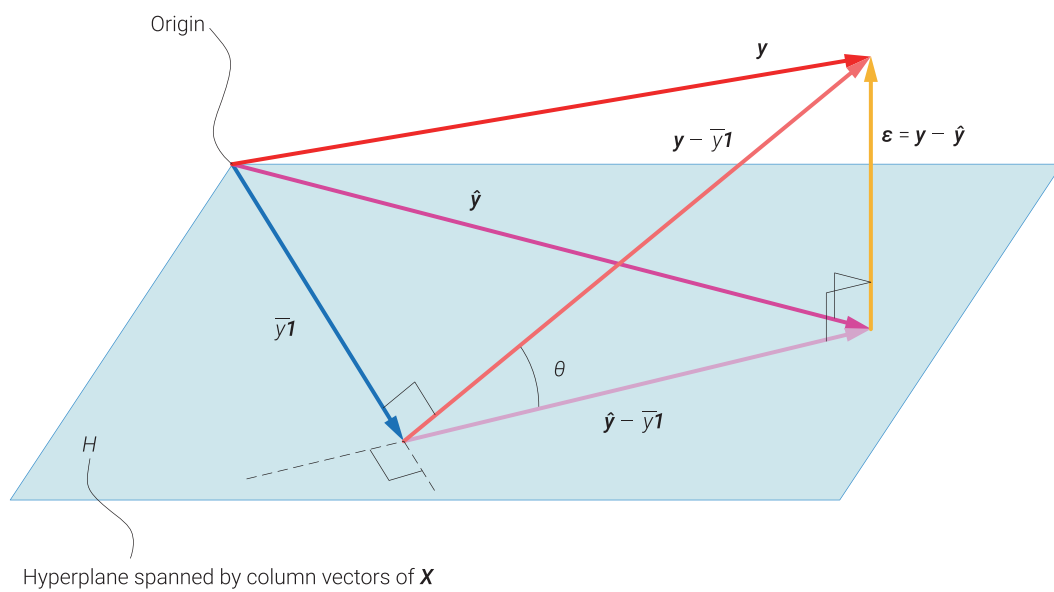


图2.10 四面体的四个面都是直角三角形

2.7 拟合优度：评价拟合程度

如图2.11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，即代表拟合效果越好。

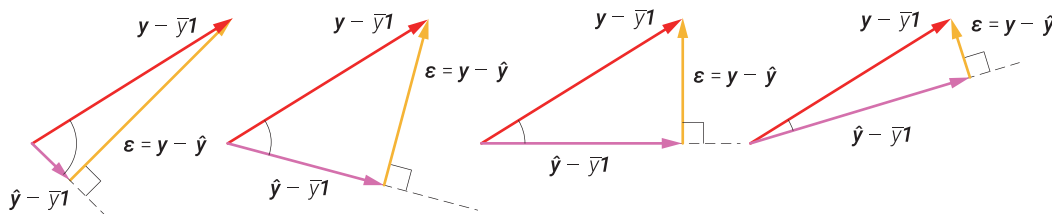


图2.11 因变量向量和预测值向量夹角从大到小

在回归模型创建之后，很自然地就要考虑这个模型是否能够很好地解释数据，即考查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，拟合优度考察回归分析中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2.10中 θ 余弦值 $\cos\theta$ 的平方：

$$R^2 = \cos^2 \theta \quad (2.20)$$

利用图2.9(b)中直角三角形三边之间的关系， R^2 可以整理为：

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (2.21)$$

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。

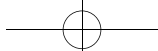
一元线性回归

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。

$$R^2 = \rho_{X,Y}^2 = \left(b_1 \frac{\sigma_X}{\sigma_Y} \right)^2 \quad (2.22)$$

其中，

$$b_1 = \rho_{X,Y} \frac{\sigma_Y}{\sigma_X} \quad (2.23)$$



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：

$$\begin{aligned} R_{\text{adj}}^2 &= 1 - \frac{\text{MSE}}{\text{MST}} \\ &= 1 - \frac{\text{SSE}/(n-k)}{\text{SST}/(n-1)} \\ &= 1 - \left(\frac{n-1}{n-k} \right) \frac{\text{SSE}}{\text{SST}} \\ &= 1 - \left(\frac{n-1}{n-k} \right) (1 - R^2) \\ &= 1 - \left(\frac{n-1}{n-D-1} \right) \frac{\text{SSE}}{\text{SST}} \end{aligned} \quad (2.24)$$

修正决定系数的作用在于，当模型中自变量的数量增加时，能够惩罚**过拟合** (overfitting)，并避免决定系数因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。

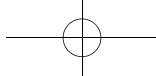
过拟合是指一个模型在训练数据上表现良好，但在测试数据上表现较差的现象。在过拟合的情况下，模型过度地学习了训练数据的特征和噪声，导致其在测试数据上的预测能力下降。

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章讲解正则化回归。

接着前文代码，我们可以通过代码2.6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) 来理解这些算法。



代码2.6 计算决定系数和修正决定系数 | Bk7_Ch02_02.ipynb

```
# 计算决定系数
a R2 = SSR/SST

# 计算修正决定系数
b R2_adj = 1 - MSE/MST
```



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：

$$\begin{aligned} F &= \frac{MSR}{MSE} = \frac{\frac{SSR}{k-1}}{\frac{SSE}{n-k}} = \frac{SSR(n-k)}{SSE(k-1)} \\ &= \frac{\frac{SSR}{D}}{\frac{SSE}{n-D-1}} = \frac{SSR \cdot (n-D-1)}{SSE \cdot (D)} \sim F(k-1, n-k) \end{aligned} \quad (2.25)$$

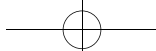
代码2.7展示了如何计算 F 检验统计量，并验证Statsmodels回归分析结果。

代码2.7 F 检验的统计量 | Bk7_Ch02_02.ipynb

```
# 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)
```





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：

$$\begin{aligned} H_0 : b_1 = b_2 = \dots = b_p = 0 \\ H_1 : b_j \neq 0 \text{ 至少有一个 } j \end{aligned} \quad (2.26)$$

具体来说， F 检验的零假设是指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

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，

$$F > F_{1-\alpha}(k-1, n-k) \quad (2.27)$$

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 $\alpha = 0.01$ ， $F_{1-\alpha}(1, 250) = 6.7373$ 。表2.2结果告诉我们， $F = 549.7 > 6.7373$ ，表明可以显著地拒绝 H_0 。

也可以用表2.1中 p 值，

$$p\text{-value} = P(F < F_\alpha(k-1, n-k)) \quad (2.28)$$

如果 p 值小于 α ，则可以拒绝零假设 H_0 。



Bk7_Ch02_02.ipynb中计算了表2.2所示ANOVA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显著不为零，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：

$$\begin{cases} H_0: b_1 = b_{1,0} \\ H_1: b_1 \neq b_{1,0} \end{cases} \quad (2.29)$$

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：

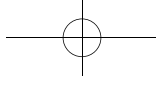
$$t_{b1} = \frac{\hat{b}_1 - b_{1,0}}{\text{SE}(\hat{b}_1)} \quad (2.30)$$

$\hat{b}_1$ 为OLS线性回归估算得到的系数， $\text{SE}(\hat{b}_1)$ 为其标准误：

$$\text{SE}(\hat{b}_1) = \sqrt{\frac{\text{MSE}}{\sum_{i=1}^n (x^{(i)} - \bar{x})^2}} = \sqrt{\frac{\frac{\sum_{i=1}^n (\varepsilon^{(i)})^2}{n-2}}{\sum_{i=1}^n (x^{(i)} - \bar{x})^2}} \quad (2.31)$$

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。

我们可以通过代码2.8计算 b_1 的 t 检验统计量。请大家对照式 (2.30) 和式 (2.31)，逐句分析代码。



代码2.8 b_1 的 t 检验统计量 | Bk7_Ch02_03.ipynb

```
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
```

$$SE(\hat{b}_1) = \frac{\sqrt{MSE}}{\sqrt{\sum_{i=1}^n (X^{(i)} - \bar{X})^2}} = \sqrt{\frac{\sum_{i=1}^n (\mathcal{E}^{(i)})^2}{n-2}} \quad t_{b1} = \frac{\hat{b}_1 - b_{1,0}}{SE(\hat{b}_1)}$$

临界值

如果下式成立，接受零假设 H_0 ：

$$-t_{1-\alpha/2, n-2} < T < t_{1-\alpha/2, n-2} \quad (2.32)$$

否则，则拒绝零假设 H_0 。

特别地，如果原假设和备择假设为：

$$\begin{cases} H_0 : b_1 = 0 \\ H_1 : b_1 \neq 0 \end{cases} \quad (2.33)$$

如果式 (2.32) 成立，接受零假设 H_0 ，即回归系数不具有显著统计性；通俗地说，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：

$$t_{b0} = \frac{\hat{b}_0 - b_{0,0}}{SE(\hat{b}_0)} \quad (2.34)$$

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

$$SE(\hat{b}_0) = \sqrt{MSE \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x^{(i)} - \bar{x})^2} \right]} = \sqrt{\frac{\sum_{i=1}^n (\mathcal{E}^{(i)})^2}{n-2} \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x^{(i)} - \bar{x})^2} \right]} \quad (2.35)$$

请大家对照式 (2.34) 和式 (2.35)，逐句学习代码 2.9。