第10章　产品销售预测模型实例

Jupyter金融应用从入门到实践

在金融领域，经常需要对某一指标进行预测，例如对房价的预测、贷款额度的预测。预测的指标可以被称为被解释变量，影响预测指标的变量被称为解释变量。对未来的把握，可以帮助从业者提前预警甚至规避风险，也可以通过掌握事物发展的内在逻辑，有条件地控制强化或减弱解释变量推动被解释变量的发展。

在本章中，我们将以产品销售额度的预测为案例来进行说明，使用Kaggle上的广告数据集（advertising-dataset）给出在电视（TV）、广播（Radio）、报纸（Newspaper）投放不等的广告带来的销售额Sales的变化。

1．加载数据

导入必需的库并加载数据集，代码如下所示。

[In]

```
# 导入库
import pandas as pd
import numpy as np

# 读取数据
df = pd.read_csv("./dataset/advertising.csv")
df.head(3)    # 展示前三行
```

[Out]

	TV	Radio	Newspaper	Sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	12.0

查看数据维度，代码如下所示。

[In]

```
# 查看数据维度
df.shape
```

[Out]

```
(200, 4)
```

2．描述统计

观察数据各个变量的范围、大小、波动趋势等，有利于判断后续对数据采取哪类模型更合适。

[In]

```
1.  # 统计数据
2.  df.describe()
```

[Out]

	TV	Radio	Newspaper	Sales
count	200.000000	200.000000	200.000000	200.000000
mean	147.042500	23.264000	30.554000	15.130500
std	85.854236	14.846809	21.778621	5.283892
min	0.700000	0.000000	0.300000	1.600000
25%	74.375000	9.975000	12.750000	11.000000
50%	149.750000	22.900000	25.750000	16.000000
75%	218.825000	36.525000	45.100000	19.050000
max	296.400000	49.600000	114.000000	27.000000

统计变量（列）缺失值，可以看出该数据不存在缺失值，代码如下所示。

[In]

```
1.  # 统计列缺失值
2.  df.isnull().sum(axis=0)
```

[Out]

```
TV           0
Radio        0
Newspaper    0
Sales        0
dtype: int64
```

3．特征相关性

相关性是指解释变量和被解释变量之间存在某种相关关系。相关系数是用以反映变量之间关系密切程度的统计指标，刻画了变量之间的线性相关程度。

相关性可以通过查看相关矩阵观察，代码如下所示。

[In]

```
1.  # 线性相关性
2.  df.corr() # 只看解释变量
```

[Out]

	TV	Radio	Newspaper	Sales
TV	1.000000	0.054809	0.056648	0.901208
Radio	0.054809	1.000000	0.354104	0.349631
Newspaper	0.056648	0.354104	1.000000	0.157960
Sales	0.901208	0.349631	0.157960	1.000000

1．一元线性回归模型

根据相关矩阵，我们选取与 Sales 相关程度最高的解释变量 TV 来训练一元线性回归模型，代码如下所示。

[In]

```
# 训练一元线性回归模型
# 提取第一列 TV 作为解释变量
 X_train_one, X_test_one=X_train[:,0].reshape(-1,1), X_test[:,0].reshape(-1,1)
from sklearn.linear_model import LinearRegression
# 创建模型
model_one = LinearRegression()
# 拟合
model_one.fit(X_train_one,y_train)
# 预测
y_train_hat_one=model_one.predict(X_train_one)
```

2. 多元线性回归模型

当线性回归中有多个解释变量时，被定义为多元回归，多的特征不一定能更优。

[In]

```
# 建模
model=LinearRegression()
# 拟合
model.fit(X_train, y_train)
# 预测
y_train_hat=model.predict(X_train)

# 计算残差平方和
print('均方误差:{:.2f}'.format(np.mean((y_train_hat - y_train) ** 2)))
# 计算 R 方
print('可决系数:{:.2f}'.format(model.score(X_train, y_train)))
```

[Out]

```
均方误差:2.40
可决系数:0.91
```

验证评估

通过训练和测试数据的评估指标之间的差异来判定是否存在过拟合。

[In]

```
# 计算残差平方和
print('均方误差:{:.2f}'.format(np.mean((y_test_hat -y_test) ** 2)))
# 计算 R 方
print('可决系数:{:.2f}'.format(model.score(X_test, y_test)))
```

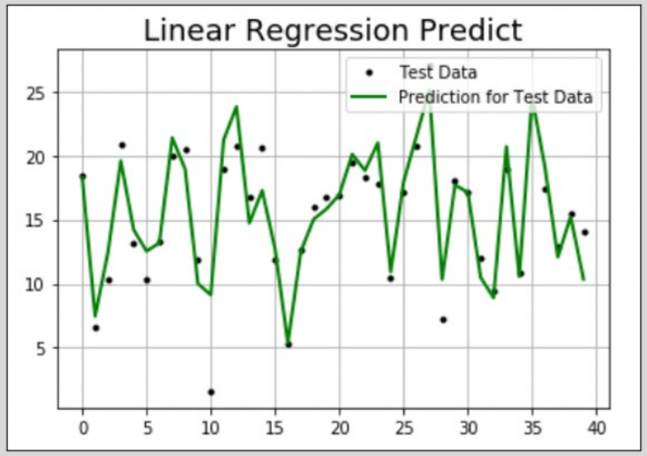
[Out]

```
均方误差:3.98
可决系数:0.85
```

[In]

```
# 模型对测试集的预测能力的可视化展示
plt.figure()
t = np.arange(len(X_test[:,0]))
plt.plot(t, y_test, 'k.', linewidth=2, label=u'Test Data')
plt.plot(t, y_test_hat, 'g-', linewidth=2, label=u'Prediction for Test Data')
plt.legend(loc='upper right')
plt.title(u'Linear Regression Predict', fontsize=18)
plt.grid()
plt.show()
```

[Out]

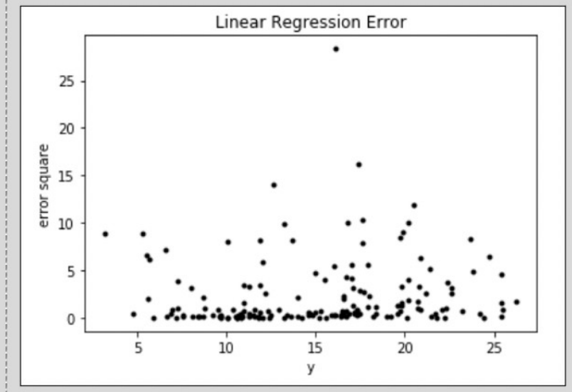


我们将 RSS 称为被解释变量的变异性中能被估计的回归方程解释的部分，与之相对的 ESS 就是未能解释的部分。如何知道未被解释的残差当中是否还隐藏着未被挖掘的解释变量呢？可以通过绘制残差平方和被解释变量的散点图来观察。在运行结果中，e_i^2 并不随 y_i 的变化而变化，说明误差项已经独立于被解释变量，被解释变量能够被已有的解释变量很好地解释。

[In]

```
# 线性回归残差图
plt.plot(y_train,(y_train_hat - y_train) ** 2, 'k.')
plt.title('Linear Regression Error')
plt.xlabel('y')
plt.ylabel('error square')
plt.show()
```

[Out]



回归模型的各项参数如下。

```
[In]
1.  print(model.intercept_)  #  打印截距
2.  print(model.coef_)     # 打印模型系数
```

```
[Out]
4.6956704711911605
[ 0.0545854   0.11158486 -0.00390938]
```

根据以上参数，多元回归模型方程式表达如下。

→ $Sales_i = 4.695 + 0.054TV_i + 0.111Radio_i - 0.003Newpaper_i + e_i, i = 1,2,3...$

预测时，可以忽略残差 e_i 带来的波动。

→ $Sales_i = 4.695 + 0.054TV_i + 0.111Radio_i - 0.003Newpaper_i$

模型的对 *TV*、*Radio*、*Newspaper* 的回归系数分别为 0.054、0.111、0.003，说明 Radio 的广告支出对销售额的影响最大，TV 次之。这意味着每增加一单位的 Radio 广告支出，可带来大约 10%的销售额提升。显然，在广告投放上，应该大力增加 Radio 渠道的支出，削减甚至放弃在 Newspaper 上的投放，以获得更高的销售额。

谢谢！