



线性回归的评价



覃雄派

提纲

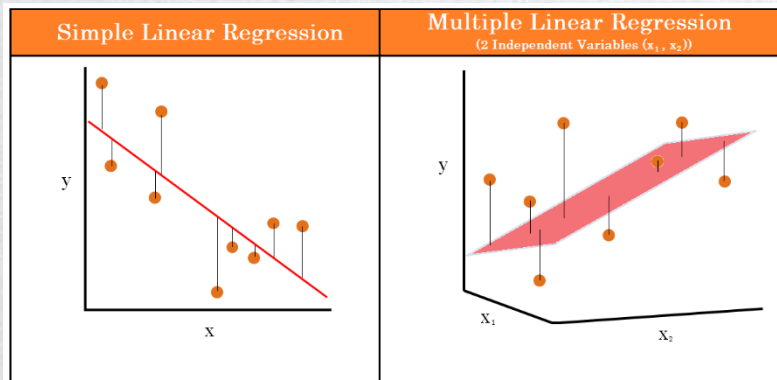


线性回归的评价

- 线性回归的评价
 - 拟合优度
 - 回归方程显著性检验
 - 回归系数的显著性检验
 - 线性回归模型评价的实例
 - 自变量筛选法
- 线性回归评价的实践

线性回归的评价

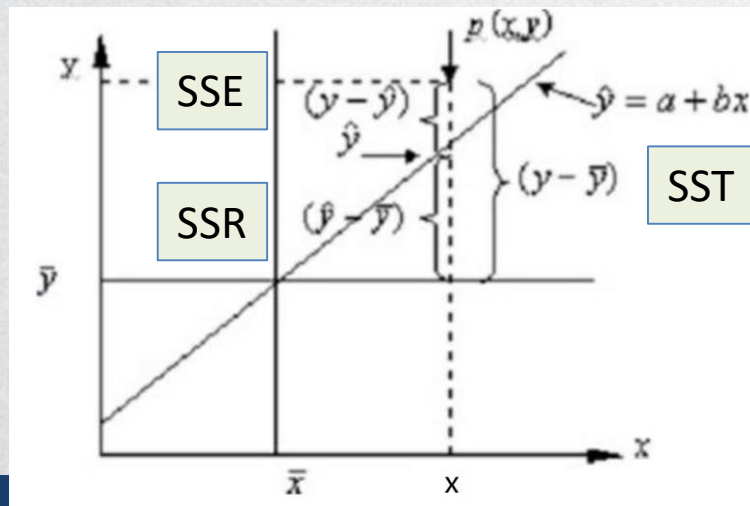
- **回归分析**是应用广泛的统计分析方法，用于分析事物之间的相关关系
 - 其中**一元线性回归**(Linear Regression)模型，指的是只有一个解释变量的线性回归模型
 - 而**多元线性回归模型**，则是包含多个解释变量的线性回归模型
 - 所谓**解释变量**就是自变量，而**被解释变量**则是因变量
 - **回归模型**就是描述因变量和自变量之间依存的数量关系的模型



如何评价回归模型的好坏呢？

线性回归的评价

- 我们把因变量的总变差(Sum of Squares for Total, SST)
 - 分解成自变量变动引起的变差(Sum of Squares for Regression, SSR)
 - 和其它因素造成的变差(Sum of Squares for Error, SSE)
 - 用数学语言来表达为
 - $SST = \sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y - \hat{y})^2 = SSR + SSE$
 - 式中, y 表示因变量的实际值, \bar{y} 表示样本均值, \hat{y} 表示模型预测值,



线性回归的评价

- (1) **拟合优度检验**：回归方程的拟合优度，指的是回归方程对样本的各个数据点的拟合程度
- 拟合优度的度量一般使用判定系数 R^2
 - 是在因变量的总变差中，由回归方程解释的变动(回归平方和)所占的比重
 - R^2 越大，方程的拟合程度越高
 - R^2 的计算公式为 $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
 - 当一个多元线性回归模型的判定系数接近1.0时，说明其拟合优度较高

线性回归的评价

- (2) 回归方程显著性检验：回归方程的显著性检验
- 目的是评价所有自变量和因变量的线性关系是否密切
 - 常用F检验统计量进行检验，F检验是对模型整体回归显著性的检验
 - F统计量的计算公式为 $F = \frac{SSR/k}{SSE/(n-k-1)}$ 式中，n为样本容量，k为自变量个数
 - F检验的原假设(H_0)为，自变量和因变量的线性关系不显著；备择假设(H_1)为，自变量和因变量的线性关系显著
 - 在给定的显著性水平(一般选0.05)下，查找自由度为(k, n-k-1)的F分布表，得到相应的临界值 F_α
 - 如果上述公式计算得的 $F > F_\alpha$ ，那么拒绝原假设，回归方程具有显著意义，回归效果显著
 - 否则， $F < F_\alpha$ ，那么接受原假设，回归方程不具有统计上的显著意义，回归效果不显著

$$\begin{cases} H_0 : \beta_0 = \beta_1 = \cdots = \beta_p = 0 \\ H_1 : \text{系数} \beta_0, \beta_1, \cdots, \beta_p \text{ 不全为} 0 \end{cases}$$

也可以用原假设的概率p值来进行判断, $p < 0.05$ ，拒绝原假设（即原假设不可能发生）， $p \geq 0.05$ 接受原假设

线性回归的评价

- (3) **回归系数的显著性检验**：使用t检验，分别检验回归模型中的各个回归系数是否具有显著性，以便使模型中只保留那些对因变量有显著影响的因素
 - t检验是对单个解释变量回归系数的显著性检验
 - 回归系数 i 的t检验统计量为 $t_i = \frac{\beta_i}{S_{\beta_i}}$ ，其中 S_{β_i} 表示回归系数 β_i 的标准误差
 - t检验的原假设(H_0)为， a_i 的值为0，即对应变量 x_i 的系数为0，该变量无需进入方程；备择假设(H_1)为， a_i 的值不为0，即对应变量 x_i 的系数不为0，该变量需要进入方程
 - 给定显著性水平 α (一般选0.05)，查找自由度为 $n-k-1$ 的t分布表，得到临界值 t_α
 - 如果 $t_i > t_\alpha$ ，拒绝原假设，回归系数 a_i 与0有显著差异，对应的自变量 x_i 对因变量 y 有解释作用
 - 否则 $t_i < t_\alpha$ ，接受原假设，回归系数 a_i 与0没有显著差异，对应的自变量 x_i 对因变量 y 没有解释作用

也可以用原假设的概率 p 值来进行判断, $p < 0.05$ ，拒绝原假设（即原假设不可能发生）， $p \geq 0.05$ 接受原假设

线性回归的评价

- (3) **回归系数的显著性检验**：使用t检验，分别检验回归模型中的各个回归系数是否具有显著性，以便使模型中只保留那些对因变量有显著影响的因素
 - t检验的原假设、备择假设以及计算方法（假设有k个变量）

$$\begin{cases} H_0 : \beta_j = 0, j = 1, 2, \dots, k \\ H_1 : \beta_j \neq 0 \end{cases}$$

$$t = \frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t(n - k - 1)$$

- $\hat{\beta}_j$ 为线性回归模型第j个系数估计值，
- β_j 为原假设的值，也就是0，
- $se(\hat{\beta}_j)$ 为回归系数的标准差

线性回归的评价



线性回归的评价

- 线性回归模型评价的实例

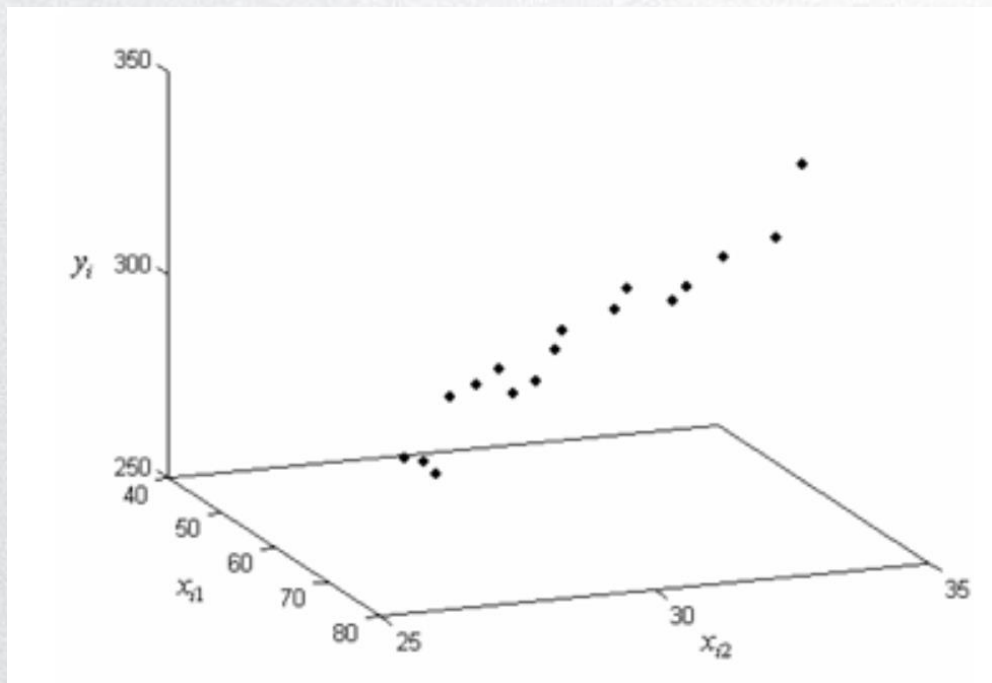
- 数据集
- a chemical process expects the yield to be affected by the levels of two factors
- That are x_1 and x_2

Observation Number	Factor 1 (x_{i1})	Factor 2 (x_{i2})	Yield (y_i)
1	41.9	29.1	251.3
2	43.4	29.3	251.3
3	43.9	29.5	248.3
4	44.5	29.7	267.5
5	47.3	29.9	273.0
6	47.5	30.3	276.5
7	47.9	30.5	270.3
8	50.2	30.7	274.9
9	52.8	30.8	285.0
10	53.2	30.9	290.0
11	56.7	31.5	297.0
12	57.0	31.7	302.5
13	63.5	31.9	304.5
14	65.3	32.0	309.3
15	71.1	32.1	321.7
16	77.0	32.5	330.7
17	77.8	32.9	349.0

http://reliawiki.org/index.php/Multiple_Linear_Regression_Analysis

线性回归的评价

- 线性回归模型评价的实例
 - 数据集
 - A scatter plot



线性回归的评价

- 线性回归模型评价的实例
 - regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$



error

线性回归的评价

- 线性回归模型评价的实例
 - 解析解

$$X = \begin{bmatrix} 1 & 41.9 & 29.1 \\ 1 & 43.4 & 29.3 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & 77.8 & 32.9 \end{bmatrix} \quad y = \begin{bmatrix} 251.3 \\ 251.3 \\ \cdot \\ \cdot \\ \cdot \\ 349.0 \end{bmatrix}$$

$$\begin{aligned} \hat{\beta} &= \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (X'X)^{-1}X'y \\ &= \begin{bmatrix} 17 & 941 & 525.3 \\ 941 & 54270 & 29286 \\ 525.3 & 29286 & 16254 \end{bmatrix}^{-1} \begin{bmatrix} 4902.8 \\ 276610 \\ 152020 \end{bmatrix} \\ &= \begin{bmatrix} -153.51 \\ 1.24 \\ 12.08 \end{bmatrix} \end{aligned}$$

线性回归的评价

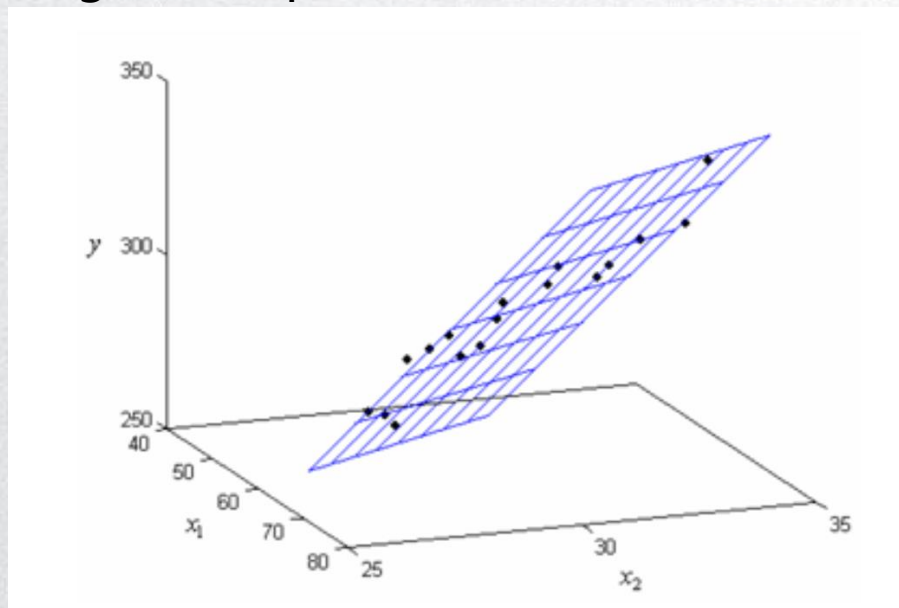
- 线性回归模型评价的实例
 - 线性回归模型

$$\hat{\beta}_0 = -153.51, \hat{\beta}_1 = 1.24 \text{ and } \hat{\beta}_2 = 12.08$$

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \\ &= -153.5 + 1.24x_1 + 12.08x_2\end{aligned}$$

线性回归的评价

- 线性回归模型评价的实例
 - 线性回归模型
 - A plot of the fitted regression plane



线性回归的评价

- 线性回归模型评价的实例
 - 线性回归模型：预测值与误差

$$\hat{y}_i = -153.5 + 1.24x_{i1} + 12.08x_{i2}$$

$$e_i = y_i - \hat{y}_i$$

– 比如

$$\begin{aligned}\hat{y}_5 &= -153.5 + 1.24x_{51} + 12.08x_{52} \\ &= -153.5 + 1.24(47.3) + 12.08(29.9) \\ &= 266.3\end{aligned}$$

$$\begin{aligned}e_5 &= y_5 - \hat{y}_5 \\ &= 273.0 - 266.3 \\ &= 6.7\end{aligned}$$

线性回归的评价

- 线性回归模型评价的实例
 - 线性回归模型：拟合优度计算R平方

$$R^2 = \frac{SS_R}{SS_T}$$

$$= 1 - \frac{SS_E}{SS_T}$$

- R平方=12816.3459/13239.72
- =96.80%

Source of Variation	Degrees of Freedom	Sum of Squares [Sequential]	Mean Squares [Sequential]	F Ratio	P Value
Model	2	12816.3459	6408.173	211.9034	3.42E-11
Factor 1	1	12530.8447	12530.8447	414.366	8.46E-12
Factor 2	1	285.5012	285.5012	9.4409	0.0083
Residual	14	423.3741	30.241		
Lack of Fit	14	423.3741	30.241		
Total	16	13239.72			

S = 5.4992	PRESS = 639.5261
R-sq = 96.80%	R-sq(pred) = 95.17%
R-sq(adj) = 96.35%	

$$SST = \sum (y - \bar{y})^2 = \sum (\hat{y} - \bar{y})^2 + \sum (y - \hat{y})^2 = SSR + SSE$$

线性回归的评价

• 线性回归模型评价的实例

- 线性回归模型：F检验——对模型整体回归显著性的检验(Significance of Regression)

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0$$

- 原假设与备择假设

$$H_1 : \beta_j \neq 0 \text{ for at least one } j$$

- F统计量

$$F_0 = \frac{MS_R}{MS_E}$$

$$\begin{aligned} MS_R &= \frac{SS_R}{dof(SS_R)} \quad \text{即} \frac{SSR}{k} \\ &= \frac{12816.35}{2} \\ &= 6408.17 \end{aligned}$$

$$\begin{aligned} MS_E &= \frac{SS_E}{dof(SS_E)} \\ &= \frac{SS_E}{(n - (k + 1))} \\ &= \frac{423.37}{(17 - (2 + 1))} \\ &= 30.24 \end{aligned}$$

- $F_0 = 6408.17 / 30.24 = 211.9$

- 在给定的显著性水平(significance level of 0.1)下，查找自由度为(k, n-k-1)即(2, 17-2-1)的F分布表，得到相应的临界值F

$$f_{0.1, 2, 14} = 2.726$$

线性回归的评价

- 线性回归模型评价的实例

- 线性回归模型：F检验——对模型整体回归显著性的检验(Significance of Regression)

- 结论

- 由于

$$f_0 > f_{0.1,2,14}$$

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Squares	<i>F</i> Statistic	<i>P</i> Value
Regression	2	12816.35	6408.17	211.9	0.00
Error	14	423.37	30.24		
Total	16	13239.72			

- 所以拒绝原假设，选择备择假设

- at least one coefficient out of β_1 and β_2 is significant
 - In other words, it is concluded that a regression model exists between yield and either one or both of the factors in the table

线性回归的评价

- 线性回归模型评价的实例

- 线性回归模型：t检验——各个系数
- 原假设与备择假设

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

- T统计量

$$T_0 = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

- 结论

- $-t_{\alpha/2, n-2} < T_0 < t_{\alpha/2, n-2}$, 不能拒绝原假设; 否则拒绝原假设

线性回归的评价

- 线性回归模型评价的实例
 - 线性回归模型：t检验——各个系数
 - The variance-covariance matrix of the estimated regression coefficients is obtained as follows
 - (具体推导这里不展开)

$$C = \hat{\sigma}^2 (X'X)^{-1}$$

$$se(\hat{\beta}_j) = \sqrt{C_{jj}}$$

$$\hat{\beta}_0 = -153.51, \hat{\beta}_1 = 1.24 \text{ and } \hat{\beta}_2 = 12.08$$

$$\begin{aligned} \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \\ &= -153.5 + 1.24x_1 + 12.08x_2 \end{aligned}$$

两个变量，两个系数



线性回归的评

- 线性回归模型评价的实例
 - 线性回归模型：t检验——各个系数

$$\begin{aligned}\hat{\sigma}^2 &= MS_E = \frac{SS_E}{dof(SS_E)} \\ &= \frac{SS_E}{(n - (k + 1))} \\ &= \frac{423.37}{(17 - (2 + 1))} \\ &= 30.24\end{aligned}$$

$$C = \hat{\sigma}^2 (X'X)^{-1}$$

$$se(\hat{\beta}_j) = \sqrt{C_{jj}}$$

$$\begin{aligned}C &= \hat{\sigma}^2 (X'X)^{-1} \\ &= 30.24 \begin{bmatrix} 336.5 & 1.2 & -13.1 \\ 1.2 & 0.005 & -0.049 \\ -13.1 & -0.049 & 0.5 \end{bmatrix} \\ &= \begin{bmatrix} 10176.75 & 37.145 & -395.83 \\ 37.145 & 0.1557 & -1.481 \\ -395.83 & -1.481 & 15.463 \end{bmatrix}\end{aligned}$$

$$se(\hat{\beta}_1) = \sqrt{0.1557} = 0.3946$$

$$se(\hat{\beta}_2) = \sqrt{15.463} = 3.93$$

线性回归的评价

- 线性回归模型评价的实例
 - 线性回归模型：t检验——各个系数

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \\ &= -153.5 + 1.24x_1 + 12.08x_2\end{aligned}$$

$$(t_0)_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{1.24}{0.3946} = 3.1393$$

$$(t_0)_{\hat{\beta}_2} = \frac{\hat{\beta}_2}{se(\hat{\beta}_2)} = \frac{12.08}{3.93} = 3.0726$$

The critical values for the present t test at a significance of 0.1 are

$$\begin{aligned}t_{\alpha/2, n-(k+1)} &= t_{0.05, 14} = 1.761 \\ -t_{\alpha/2, n-(k+1)} &= -t_{0.05, 14} = -1.761\end{aligned}$$

两个t统计量都处在阈值之外，应该拒绝原假设，接受备择假设，即两个系数都不为0

线性回归的评价

- 线性回归模型评价的实例
 - 线性回归模型：t检验——各个系数

$$(t_0)_{\hat{\beta}_1} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{1.24}{0.3946} = 3.1393$$

$$(t_0)_{\hat{\beta}_2} = \frac{\hat{\beta}_2}{se(\hat{\beta}_2)} = \frac{12.08}{3.93} = 3.0726$$

Regression Table								
Regression Information								
Term	Effect	Coefficient	Standard Error	Low Confidence	High Confidence	T Value	P Value	Variance Inflation Factor
Intercept		-153.5117	100.8799	-331.1924	24.169	-1.5217	0.1503	
Factor 1	2.4774	1.2387	0.3946	0.5437	1.9337	3.1393	0.0072	11.2389
Factor 2	24.1647	12.0824	3.9323	5.1564	19.0083	3.0726	0.0083	11.2389

线性回归的评价





线性回归的评价

- **自变量筛选法**

- 在多元线性回归中，存在一个自变量选择的问题，因为并不是所有的自变量都对因变量有解释作用
 - 比如，我们通过身高、体重(自变量)和肺活量(因变量)，建立的回归模型中
 - 这时候，我们引入一个血压数据，就可能和肺活量没有什么关系
 - 自变量间可能存在较强的线性关系，即共线性
 - 所以不能把所有的变量全部引入方程
 - 变量选择的方法，包括前向筛选法、后向筛选法、和逐步筛选法三种



线性回归的评价

- **自变量筛选法：前向筛选法(Forward)**
 - 自变量不断进入回归方程的过程
 - 选择与因变量具有最高相关系数的自变量进入方程，并进行各种检验
 - 其次，在剩余的自变量中寻找偏相关系数最高的变量，进入回归方程，并进行检验
 - 反复上述步骤，直到没有可进入方程的自变量为止
 - 回归系数检验的概率 P 值小于 $P_{in}(0.05)$ ，才可以进入方程

线性回归的评价

- **自变量筛选法：后向筛选法(Backward)**
 - 自变量不断剔除出回归方程的过程
 - 首先，将所有自变量全部引入回归方程
 - 其次，在一个或多个t值不显著的自变量中，将t值最小的那个变量剔除出去，并重新建立方程和进行检验
 - 回归系数检验P值大于 $P_{out}(0.10)$ ，则剔除出方程
 - 如果新方程中所有变量的回归系数t值都是显著的，则变量筛选过程结束
 - 否则，重复上述过程，直到没有变量可剔除为止

线性回归的评价

- **自变量筛选法：逐步筛选法(Stepwise)**
 - 是“前向筛选法”和“后向筛选法”的结合
 - 前向筛选法，只对进入方程的变量的回归系数进行显著性检验，而对已经进入方程的其它变量的回归系数不再进行显著性检验，也就是，变量一旦进入方程就不会被剔除
 - 随着变量的逐个引进，由于变量之间存在着一定程度的相关性，使得已经进入方程的变量，其回归系数不再显著，因此会造成最后的回归方程可能包含不显著的变量
 - 逐步筛选法则在变量选择的每一个阶段，都考虑剔除一个变量的可能性

线性回归的评价






线性回归的评价

- 线性回归评价的实践

2021-07-18 《数据科学概论》 new plan › 2022newPPT › 0305-回归：多元回归（解析解与梯度下降算法） ›

名称	类型	大小	修改日期
 05_evaluate_regression_model.ipynb	IPYNB 文件	66 KB	2021/12/1 16:15

线性回归的评价

- 线性回归评价的实践
 - 装载数据集，显示前5行

```
profit=pd.read_csv('50_Startups.csv', sep=",")  
profit.head()
```

	R&D Spend	Administration	Marketing Spend	State	Profit
0	165349.20	136897.80	471784.10	New York	192261.83
1	162597.70	151377.59	443898.53	California	191792.06
2	153441.51	101145.55	407934.54	Florida	191050.39
3	144372.41	118671.85	383199.62	New York	182901.99
4	142107.34	91391.77	366168.42	Florida	166187.94

线性回归的评价

- 线性回归评价的实践
 - 修改数据集的列名称（为了进行回归），显示前5行

```
profit = profit.rename(columns={'R&D Spend': 'RD_Spend', 'Administration': 'Administration',  
                               'Marketing Spend': 'Marketing_Spend', 'State': 'State', 'Profit': 'Profit'})  
profit.head()
```

	RD_Spend	Administration	Marketing_Spend	State	Profit
0	165349.20	136897.80	471784.10	New York	192261.83
1	162597.70	151377.59	443898.53	California	191792.06
2	153441.51	101145.55	407934.54	Florida	191050.39
3	144372.41	118671.85	383199.62	New York	182901.99
4	142107.34	91391.77	366168.42	Florida	166187.94

线性回归的评价

- 线性回归评价的实践
 - 回归，显示截距和系数

```
fit=sm.formula.ols('Profit~RD_Spend+Administration+Marketing_Spend', data=profit).fit()  
print(fit.params)
```

Intercept	50122.192990
RD_Spend	0.805715
Administration	-0.026816
Marketing_Spend	0.027228
dtype:	float64

Profit=50122.192990 + 0.875715RD_Spend - 0.026816Administration
+0.027228Marketing_Spend

线性回归的评价

- 线性回归评价的实践
 - 显示R平方系数、F检验结果、t检验结果

```
print(fit.summary())
```

OLS Regression Results

Dep. Variable:	Profit	R-squared:	0.951
Model:	OLS	Adj. R-squared:	0.948
Method:	Least Squares	F-statistic:	296.0
Date:	Wed, 01 Dec 2021	Prob (F-statistic):	4.53e-30
Time:	16:18:47	Log-Likelihood:	-525.39
No. Observations:	50	AIC:	1059.
Df Residuals:	46	BIC:	1066.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	5.012e+04	6572.353	7.626	0.000	3.69e+04	6.34e+04
RD_Spend	0.8057	0.045	17.846	0.000	0.715	0.897
Administration	-0.0268	0.051	-0.526	0.602	-0.130	0.076
Marketing_Spend	0.0272	0.016	1.655	0.105	-0.006	0.060

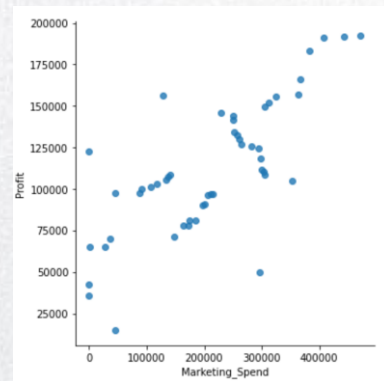
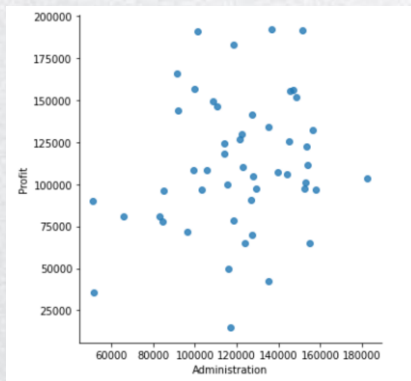
Omnibus:	14.838	Durbin-Watson:	1.282
Prob(Omnibus):	0.001	Jarque-Bera (JB):	21.442
Skew:	-0.949	Prob(JB):	2.21e-05
Kurtosis:	5.586	Cond. No.	1.40e+06

- R平方为0.951，拟合优度较好
- F-statistic : 296.0 , Prob (F-statistic):4.53e-30, F统计量值为296.0, 对应的概率值P远远小于0.05, 说明应该拒绝原假设(原假设不成立), 认为模型是显著的
- 在各自变量的t统计中, Administration 和 Marketing_Spend 变量所对应的概率值p大于0.05, 说明不能拒绝原假设, 这些变量是不显著的, 无法认定其实影响Profit的重要因素

线性回归的评价

- 线性回归评价的实践
 - 由于Administration和Marketing_Spend变量的t检验结果是不显著的
 - 故可以探索这些变量与Profit之间的散点关系，如果确实没有线性关系，可将其从模型中剔除

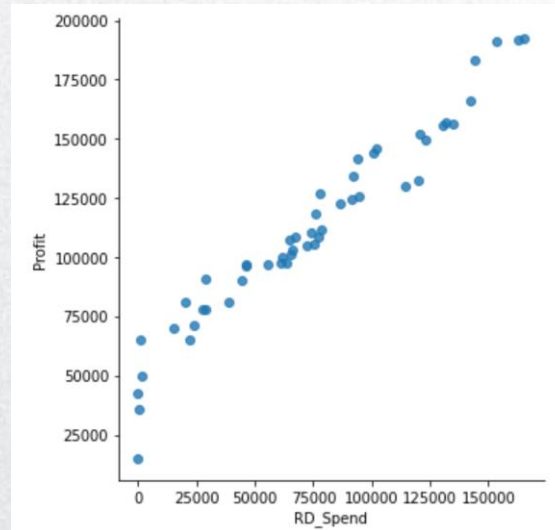
```
sns.lmplot(x='Administration', y='Profit', data=profit,  
           legend_out=False, #将图例呈现在图框内  
           fit_reg=False #不显示拟合曲线  
           )  
sns.lmplot(x='Marketing_Spend', y='Profit', data=profit,  
           legend_out=False, #将图例呈现在图框内  
           fit_reg=False #不显示拟合曲线  
           )  
plt.show()
```



线性回归的评价

- 线性回归评价的实践
 - 看看RD_Spend和目标变量的关系

```
sns.lmplot(x='RD_Spend', y='Profit', data=profit,  
           legend_out=False, #将图例呈现在图框内  
           fit_reg=False #不显示拟合曲线  
           )  
plt.show()
```



线性回归的评价

- 线性回归评价的实践
 - 剔除变量, 重建模型fit2

$$\text{Profit} = 49032.899141 + 0.854291 \text{RD_Spend}$$

```
#将Administration、Marketing_Spend变量从模型中剔除
fit2 = sm.formula.ols('Profit~RD_Spend', data=profit).fit()
print(fit2.params)
print(fit2.summary())
```

```
Intercept    49032.899141
RD_Spend      0.854291
dtype: float64
```

OLS Regression Results

```
=====
Dep. Variable:          Profit    R-squared:          0.947
Model:                  OLS      Adj. R-squared:       0.945
Method:                 Least Squares    F-statistic:       849.8
Date:                  Wed, 01 Dec 2021    Prob (F-statistic): 3.50e-32
Time:                  16:26:33    Log-Likelihood:    -527.44
No. Observations:      50    AIC:                  1059.
Df Residuals:          48    BIC:                  1063.
Df Model:               1
Covariance Type:        nonrobust
=====
```

	coef	std err	t	P> t	[0.025	0.975]
Intercept	4.903e+04	2537.897	19.320	0.000	4.39e+04	5.41e+04
RD_Spend	0.8543	0.029	29.151	0.000	0.795	0.913

```
=====
Omnibus:                 13.727    Durbin-Watson:          1.116
Prob(Omnibus):            0.001    Jarque-Bera (JB):       18.536
Skew:                     -0.911    Prob(JB):               9.44e-05
Kurtosis:                  5.361    Cond. No.                1.65e+05
=====
```

- 新模型fit2通过了显著性检验
- 新模型fit2的每个自变量所对应的系数也是通过显著性检验的

线性回归的评价



线性回归的评价

- 异常值

- 回归模型计算过程会依赖于自变量的均值，均值的最大弊端是其容易受到异常点（或极端值）的影响
- 建模数据中存在异常点，一定程度上会影响到建模的有效性
- 对于现行回归模型来说，通常利用帽子矩阵、DFFITS准则、学生化残差或Cook距离进行异常点检测
 - （这4种异常值检测方法的原理，在此不展开讨论）
 - 使用以上4种方法判别数据集的第 i 个样本是否为异常点
 - 前提是已构建好一个线性回归模型，然后基于由get_influence方法获得4种统计量的值

线性回归的评价

- 异常值
 - 计算每个样本的异常值指标

```
#异常值检验
outliers=fit2.get_influence()

#高杠杆值点（帽子矩阵）
leverage=outliers.hat_matrix_diag
#dffits值
dffits=outliers.dffits[0]
#学生化残差
resid_stu=outliers.resid_studentized_external
#cook距离
cook=outliers.cooks_distance[0]

concat_result = pd.concat([pd.Series(leverage, name='leverage'), pd.Series(dffits, r
    pd.Series(resid_stu, name='resid_stu'), pd.Series(cook, name='cook')], axis=1)

raw_outliers = pd.concat([profit, concat_result], axis=1)
raw_outliers.head()
```

	RD_Spend	Administration	Marketing_Spend	State	Profit	leverage	diffits	resid_stu	cook
0	165349.20	136897.80	471784.10	New York	192261.83	0.101318	0.073456	0.218771	0.002753
1	162597.70	151377.59	443898.53	California	191792.06	0.096508	0.139502	0.426837	0.009899
2	153441.51	101145.55	407934.54	Florida	191050.39	0.081556	0.362855	1.217678	0.065177
3	144372.41	118671.85	383199.62	New York	182901.99	0.068347	0.315040	1.163141	0.049263
4	142107.34	91391.77	366168.42	Florida	166187.94	0.065297	-0.122261	-0.462571	0.007598



线性回归的评价

- 异常值

- 通过学生化残差识别出异常值（一定阈值下），异常值比例为4%
- 由于异常值比例非常小，故可以考虑将其直接从数据集中删除，由此继续建模将会得到更加稳健且合理的模型

```
outliers_ratio=sum(np.where(np.abs(raw_outliers.resid_stu)>2,1,0))/raw_outliers.shape[0]  
print(outliers_ratio)
```

0.04



线性回归的评价

- 异常值
 - 去掉异常值，建立模型fit3

$$\text{Profit} = 51454.448622 + 0.836629 \text{RD_Spend}$$

```
#通过筛选的方法，将异常点排除
none_outliers=raw_outliers.loc[np.abs(raw_outliers.resid_stu)<=2,]

#应用无异常值的数据集重新建模
fit3=sm. formula.ols('Profit~RD_Spend',data=none_outliers).fit() #none_outliers

#返回模型的概览信息
print(fit3.params)
print(fit3.summary())

Intercept    51454.448622
RD_Spend      0.836629
dtype: float64
```

排除异常点之后得到的模型fit3，不管是模型的显著性检验还是系数的显著性检验，各自的概率P值均小于0.05，说明它们均通过显著性检验

OLS Regression Results

Dep. Variable:	Profit	R-squared:	0.962
Model:	OLS	Adj. R-squared:	0.961
Method:	Least Squares	F-statistic:	1171.
Date:	Wed, 01 Dec 2021	Prob (F-statistic):	2.28e-34
Time:	16:32:06	Log-Likelihood:	-495.77
No. Observations:	48	AIC:	995.5
Df Residuals:	46	BIC:	999.3
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	5.145e+04	2119.025	24.282	0.000	4.72e+04	5.57e+04
RD_Spend	0.8366	0.024	34.221	0.000	0.787	0.886

Omnibus:	0.188	Durbin-Watson:	1.550
Prob(Omnibus):	0.910	Jarque-Bera (JB):	0.381
Skew:	0.089	Prob(JB):	0.827
Kurtosis:	2.601	Cond. No.	1.68e+05

线性回归的评价

- 异常值
 - 进行预测验证一下

```
pred=fit3.predict(profit[['RD_Spend']])
```

#对于实际值与预测值的比较

```
df=pd.concat([pd.Series(profit.Profit/100, name='real'), pd.Series(pred/100, name='prediction')], axis=1)  
df['误差绝对值']=np.abs((df['real']-df['prediction'])/100)
```

```
df.head(10)
```

	real	prediction	误差绝对值
0	1922.6183	1897.903980	0.247143
1	1917.9206	1874.884131	0.430365
2	1910.5039	1798.280783	1.122231
3	1829.0199	1722.406055	1.066138
4	1661.8794	1703.455820	0.415764
5	1569.9112	1617.864984	0.479538
6	1561.2251	1640.776573	0.795515
7	1557.5260	1604.656535	0.471305
8	1522.1177	1523.038265	0.009206
9	1497.5996	1546.399960	0.488004

- 针对原始数据profit目标变量，根据fit3模型重新预测各成本下的利润预测值
- 从结果上看有的预测值比较接近实际值，有的预测测偏离实际值稍微远一点
- 但从总体上来说，预测值与实际值之间的差异并不是特别大

线性回归的评价



线性回归的评价

- 变量的相关系数
 - 考察自变量、因变量的相关系数
 - 考察自变量之间的相关系数
 - (把冗余的自变量剔除)

$$\rho_{x,y} = \frac{COV(x,y)}{\sqrt{D(x)} \sqrt{D(y)}}$$

$COV(x,y)$ 为自变量 x 与因变量 y 之间的协方差, $D(x)$ 和 $D(y)$ 分别为自变量 x 和因变量 y 的 方差

$ \rho \geq 0.8$	$0.5 \leq \rho < 0.8$	$0.3 \leq \rho < 0.5$	$ \rho < 0.3$
高度相关	中度相关	弱相关	几乎不相关

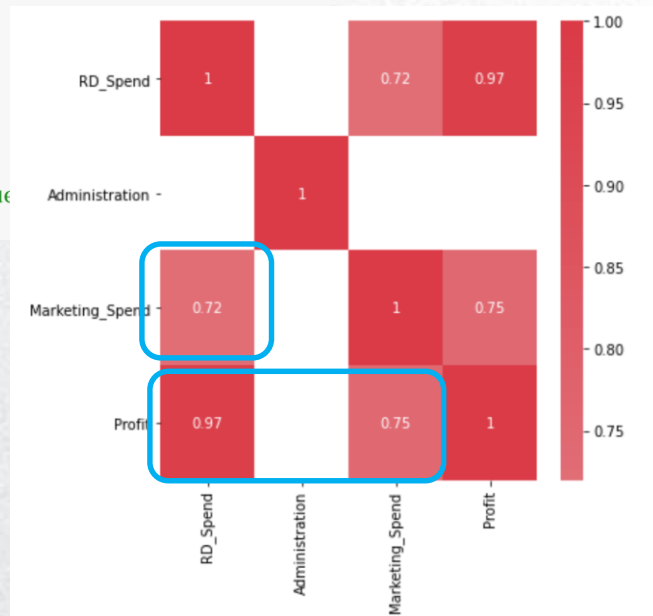
线性回归的评价

• 变量的相关系数

```
ContinuousCols = ['RD_Spend', 'Administration', 'Marketing_Spend', 'Profit']

df2 = profit[ContinuousCols]
cor_matrix = df2.corr().round(2)

# Plotting heatmap
fig = plt.figure(figsize=(6,6));
kot = cor_matrix[abs(cor_matrix)>=.45]
sns.heatmap(kot, annot=True, center=0, cmap = sns.diverging_palette(250, 10, as_cmap=True))
plt.show()
```



- 目标变量profit和RD_spend、Marketing_Spend（自变量与因变量）的相关系数比较高
- 但是Marketing_Spend和RD_spend（自变量之间）的相关系数比较高