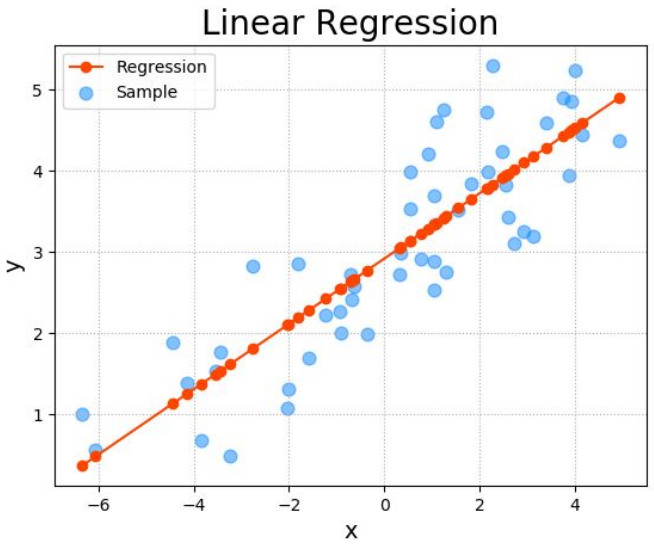
简单线性回归

2 课 程 目 录



01

简单线性回归模型

4 线性回归介绍

- 管理决策经常取决于对两个或多个变量之间关系的分析：
  - 把预测的变量称为因变量
  - 把用来预测因变量的一个或多个变量成为自变量
  - 只包括一个自变量和一个因变量，二者之间的关系可以用一条直线近似表示，这种回归分析被称为简单线性回归

5 回 归 模 型

Armand披萨店案列

Armand披萨店在美国都开在大学附近，其管理人员认为连锁店的销售收入y与学生人数x是正相关的。我们将会利用回归分析，求出一个能说明**因变量y**是如何依赖**自变量x**的方程。

在案例中，总体是由所有的Armand披萨店连锁组成的。对于总体中的每一个连锁店，都有一个学生值x和一个对应的季度销售收入y值。描述y如何依赖于x和误差项的方程称为回归模型

简单线性回归模型

$$y = \beta_0 + \beta_1 x + \epsilon$$

式中β_0，β_1被称为模型的参数；ϵ是一个随机变量，被称为模型的误差项。误差项说明了包含在y里，但不能被x和y之间的线性关系解释的变异性

6 回 归 模 型

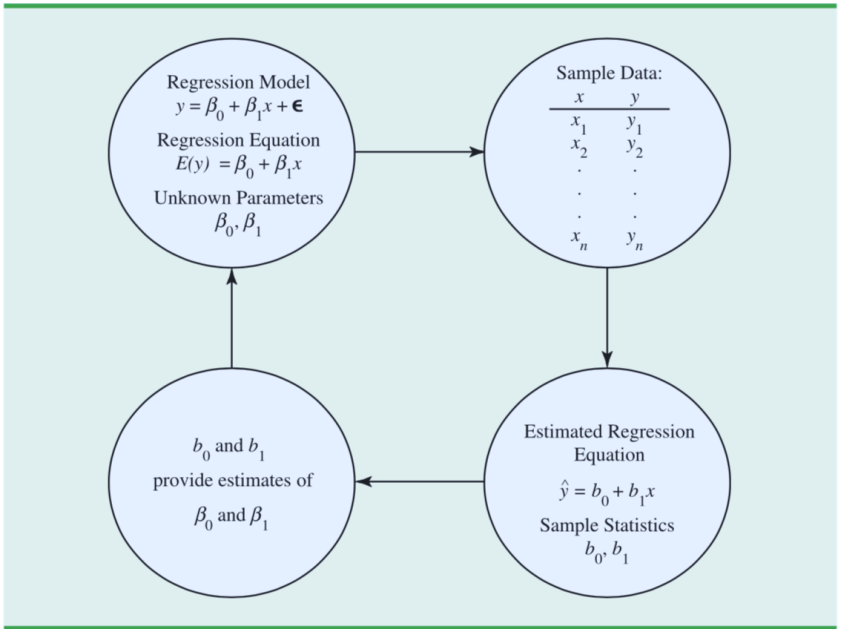
- Armand披萨店的总体还可以视为由若干个子总体组成的集合，每一个子总体都对应一个不同的x值：
  - 一个子总体由8000名学生的校园附近的Armand披萨店组成的
  - 一个子总体由9000名学生的校园附近的Armand披萨店组成的
- 每个子总体都会对应一个y值的分布，y值的每一个分布都有它自己的平均值或期望值。
- 描述y的期望值E(y)如何依赖于x的方程称为回归方程

简单线性回归方程

$$E(y) = \beta_0 + \beta_1 x$$

7 估计的回归方程

简单线性回归的估计步骤



1

- 如果总体参数β_0，β_1已知，那么对于一个给定的x值，我们就能利用公式：简单线性回归方程，计算y的平均值
- 在实际中，β_0，β_1都是未知的，我们必须使用样本统计量b_0，b_1作为总体参数β_0，β_1的估计量。
- **回归估计方程：**

$$\hat{y} = b_0 + b_1 x$$

2

- b_0是y轴截距，b_1是斜率。下一节，我们使用最小二乘法计算估计回归方程中的值
- 对于x的一个给定值，$\hat{y}$是y的平均值E(y)的一个点估计

3

我们不能把回归分析看作在变量之间建立一个因果关系的过程。回归分析只能表明变量时如何彼此联系在一起的

02

最小二乘法则

9 最小二乘法则

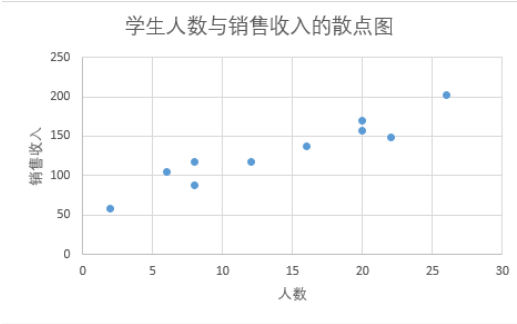
最小二乘法（least squares method），是利用样本数据建立估计的回归方程的一种方法.

x_i: 第i个观测值或者第i家连锁店的学生人数
y_i: 第i个观测值或者第i家连锁店的销售收入
表中的数据显示了x_i与y_i的数据对应关系

散点图可以帮助我们观察到学生人数与销售收入的关系；似乎人数越多，销售收入越高，它们之间可以用一条直线近似的表示！我们使用简单线性回归模型来表示销售收入与学生人数之间的关系！

1	Restaurant	Population	Sales
2	1	2	58
3	2	6	105
4	3	8	88
5	4	8	118
6	5	12	117
7	6	16	137
8	7	20	157
9	8	20	169
10	9	22	149
11	10	26	202

Armand披萨店学生人数与销售收入数据



关系散点图

10 最小二乘法则

- 为了使估计的回归直线能对样本数据有一个好的拟合，我们希望销售收入的观测值与销售收入的预测值之间的差要小
- 最小二乘法利用样本数据，通过使因变量的观测值y_i 与因变量的预测值$\hat{y}_i$之间的差的平方和达到最小的方法求得b_0，b_1的值。

最小二乘法准则

注：德国数学家高斯提出的

$$min \sum (y_i - \hat{y}_i)^2$$

式中，y_i为对于第i次观测因变量的观测值；$\hat{y}_i$为对于第i次观测因变量的预测值

11 微 分 学 求 解

SLOPE AND y-INTERCEPT FOR THE ESTIMATED REGRESSION EQUATION*

$$b_1 = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\Sigma(x_i - \bar{x})^2} \quad (14.6)$$

$$b_0 = \bar{y} - b_1\bar{x} \quad (14.7)$$

where

x_i = value of the independent variable for the ith observation
y_i = value of the dependent variable for the ith observation
$\bar{x}$ = mean value for the independent variable
$\bar{y}$ = mean value for the dependent variable
n = total number of observations

注意：通常计算机软件可以帮助我们直接求解

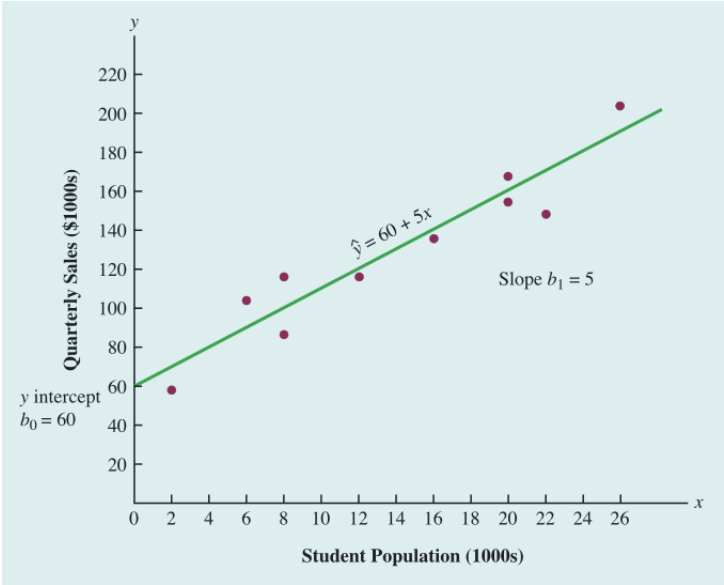
12 Armand案例计算过程

Restaurant i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	2	58	−12	−72	864	144
2	6	105	−8	−25	200	64
3	8	88	−6	−42	252	36
4	8	118	−6	−12	72	36
5	12	117	−2	−13	26	4
6	16	137	2	7	14	4
7	20	157	6	27	162	36
8	20	169	6	39	234	36
9	22	149	8	19	152	64
10	26	202	12	72	864	144
Totals	140	1300			2840	568
	Σx_i	Σy_i			$\Sigma(x_i - \bar{x})(y_i - \bar{y})$	$\Sigma(x_i - \bar{x})^2$

- 计算平均数 $\bar{x}$: 140/10 = 14 $\bar{y}$: 1300/10 = 130
- 利用公式便可求解得出
  b_0=60
  b_1=5
- 估计的回归方程：$\hat{y} = 60 + 5x$

13 方 程 结 论

回归方程的图表示



- 回归方程的斜率是正的，表明随着学生人数的增加，Armand披萨店的季度销售收入也增加。
- 更进一步的结论为学生人数每增加一人，销售收入增加5美元
- 存在的问题：我们如何判定我们的模型是合理的呢？

03

判　定　系　数

15 判 定 系 数

存在的疑问：估计的回归方程 $\hat{y} = 60 + 5x$ 能否很好地拟合了样本数据？

判定系数为估计的回归方程提供了一个拟合优度的度量。

误差平方和	SSE = $\sum(y_i - \hat{y}_i)^2$
总的平方和	SST= $\sum(y_i - \bar{y})^2$
回归平方和	SSR= $\sum(\hat{y}_i - \bar{y})^2$

SST = SSR + SSE

16 判 定 系 数

如何利用这三个平方和为估计的回归方程给出一个拟合优度的度量呢？

- 如果因变量的每一个值y_i都刚好落在估计的回归线上，那么估计的回归方程将给出一个完全的拟合。这时候，对于每一个观测值, $y_i - \hat{y}_i$都等于0，从而导致SSE=0.
- SST = SSR + SSE，对于一个完全拟合，SSR必须等于SST，也就是SSR/SST =1.
- 值SSR/SST将在0到1之间取值，这个比值被称为判定系数

判定系数

$$r^2 = \frac{SSR}{SST}$$

- 回到披萨店的案例中，$r^2 = \frac{SSR}{SST}$=14200/15730 = 0.9027
- 我们把r^2理解为总平方和中能被估计的回归方程解释的百分比。
- 在用估计的回归方程去预测销售收入时，我们能断定总平方和的90.27%能被估计的回归方程所解释。
- 换句话说：销售收入变异性的90.27%能被学生人数和销售收入之间的线性关系所解释

17 相 关 系 数

- **样本相关系数**
  - $r_{xy} = (b_1$的符号$)\sqrt{r^2}$
- 关于Armand披萨店中的样本相关系数： +0.9501 表示的是x和y之间存在着一个强的正向的线性关系
- 在两变量之间的存在一个线性关系的情况下，判定系数和样本相关系数都给出了他们之间线性关系强度的度量。
- 虽然样本相关系数的适用范围被限制在两变量之间存在线性关系的情况，但是判定系数对非线性关系以及有两个以上的自变量的相关关系都适用。因此判定系数有着更广泛的适应范围

04

模 型 的 假 定

19 模型的假定

判定系数r^2即使较大，我们也需要对假定模型的合理性做出进一步的分析。确定是否合理的一个步骤：是要对变量之间关系的显著性进行检验。回归分析中的显著性检验是以对误差项ε的下列假定为依据进行的：

1. 误差项ε是一个平均值或期望值为0的随机变量，即$E(\varepsilon)=0$
2. 对所有的x值，ε的方差都是相同的。我们用σ^2表示ε的方差
3. ε的值是相互独立的
4. 对所有特定的x值，误差项ε是一个正态分布的随机变量

05

显 著 性 检 验

21 显著性检验

- 对于E(y) = $\beta_0 + \beta_1 x$ ，如果 β_1 的值是0，那么y的平均值或者期望值不依赖于x的值，因此我们得出的结论是两变量x和y之间不存在线性关系。
- 我们必须使用一个假设检验来判定β_1的值是否等于0，从而检验出两个变量之间是否存在线性关系

22 σ^2 的估计

- ϵ的方差σ^2也是因变量y的值关于回归直线的方差。
- 残差平方和SSE是实际观测值关于估计的回归线变异性的度量
- 利用SSE除以它自己的自由度，得到均方误差MSE，均方误差给出了σ^2的一个估计量
- 统计学家已经证明，为了计算SSE，必须估计两个参数β_0，β_1，所以SSE的自由度是n-2。
- **均方误差（σ^2的估计量）：**
  - s^2=MSE=$\frac{SSE}{n-2}$
- **估计的标准误差**
  - $s=\sqrt{MSE}$
- 披萨店的案列中，s^2=1530/8=191.25, s=13.829

23 t 检 验

- 构建关于β_1的双侧检验
  - H_0: $\beta_1 = 0$ H_1: $\beta_1 \neq 0$
  - 如果H_0被拒绝，我们将会得到$\beta_1 \neq 0$的结论，于是证明了x与y之间存在一个统计上显著的关系；如果H_0没有被拒绝，我们将没有充分的理由来断定x与y之间存在一个统计上显著的关系
- 我们使用了不同的样本就会得出不同的估计的回归方程。比如披萨店的案列中，我们假设使用了另一批样本，我们得到的估计回归方程只能是和$\hat{y} = 60 + 5x$ 相似，但是不太可能完全一致
- 实际上，最小二乘估计量b_0，b_1是样本统计量，他们有着自己的抽样分布

24 b_1的抽样分布

b_1的抽样分布

SAMPLING DISTRIBUTION OF b_1

Expected Value

$$E(b_1) = \beta_1$$

Standard Deviation

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{\Sigma(x_i - \bar{x})^2}} \qquad (14.17)$$

Distribution Form

Normal

b_1的期望值等于β_1，所以b_1是β_1的无偏估计量

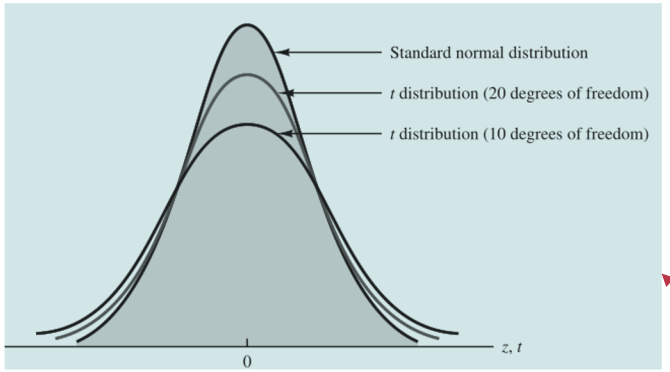
b_1的估计的标准差

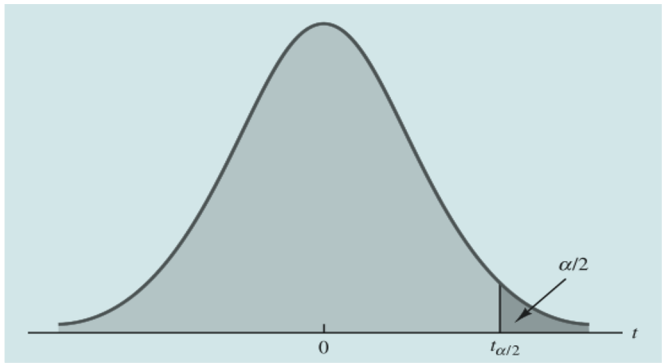
ESTIMATED STANDARD DEVIATION OF b_1

$$s_{b_1} = \frac{s}{\sqrt{\Sigma(x_i - \bar{x})^2}} \qquad (14.18)$$

由于σ是未知的，使用样本标准差

25 t 分 布 回 顾





给t加上下标表明其在t分布上侧的面积

σ未　　知

在建立总体均值的区间估计时，我们通常并没有关于总体标准差一个号的估计。在这种情形下，我们必须利用同一样本估计μ和σ两个参数。

t　分　布

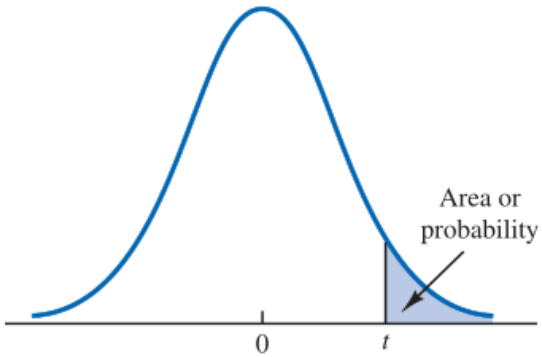
定义：在概率论和统计学中，t-分布（t-distribution）用于根据小样本来估计呈正态分布且方差未知的总体的均值。如果总体方差已知（例如在样本数量足够多时），则应该用正态分布来估计总体均值。

当利用s估计σ时，边际误差和总体均值的区间估计都以t分布的概率分布为依据进行的。虽然t分布的数学推导是以假设总体服从正态分布为依据的，但是许多研究表明在总体分布偏态的情形下，t分布效果也相当不错。

t 分布依赖于自由度的参数。当自由度为1,2,3,...时，有且仅有唯一的t分布与之相对应。随着自由度的增大，t分布与标准正态分布之间的差别变得越来越来越小

t分布的均值为0

26 t分布表回顾



Degrees of Freedom	Area in Upper Tail					
	.20	.10	.05	.025	.01	.005
1	1.376	3.078	6.314	12.706	31.821	63.656
2	1.061	1.886	2.920	4.303	6.965	9.925
3	.978	1.638	2.353	3.182	4.541	5.841
4	.941	1.533	2.132	2.776	3.747	4.604
5	.920	1.476	2.015	2.571	3.365	4.032
6	.906	1.440	1.943	2.447	3.143	3.707
7	.896	1.415	1.895	2.365	2.998	3.499
8	.889	1.397	1.860	2.306	2.896	3.355
9	.883	1.383	1.833	2.262	2.821	3.250
⋮	⋮	⋮	⋮	⋮	⋮	⋮
60	.848	1.296	1.671	2.000	2.390	2.660
61	.848	1.296	1.670	2.000	2.389	2.659
62	.847	1.295	1.670	1.999	2.388	2.657
63	.847	1.295	1.669	1.998	2.387	2.656
64	.847	1.295	1.669	1.998	2.386	2.655
65	.847	1.295	1.669	1.997	2.385	2.654
66	.847	1.295	1.668	1.997	2.384	2.652
67	.847	1.294	1.668	1.996	2.383	2.651
68	.847	1.294	1.668	1.995	2.382	2.650
69	.847	1.294	1.667	1.995	2.382	2.649
⋮	⋮	⋮	⋮	⋮	⋮	⋮
90	.846	1.291	1.662	1.987	2.368	2.632
91	.846	1.291	1.662	1.986	2.368	2.631
92	.846	1.291	1.662	1.986	2.368	2.630
93	.846	1.291	1.661	1.986	2.367	2.630
94	.845	1.291	1.661	1.986	2.367	2.629
95	.845	1.291	1.661	1.985	2.366	2.629
96	.845	1.290	1.661	1.985	2.366	2.628
97	.845	1.290	1.661	1.985	2.365	2.627
98	.845	1.290	1.661	1.984	2.365	2.627
99	.845	1.290	1.660	1.984	2.364	2.626
100	.845	1.290	1.660	1.984	2.364	2.626
∞	.842	1.282	1.645	1.960	2.326	2.576

27 计 算 过 程

有上述公式，可以得出：

- b_1的估计的标准差为：$s_{b_1} = 13.829/\sqrt{568}$=0.5803
- t检验的检验统计量：$\frac{b_1-\beta_1}{s_{b_1}}$,这是一个服从自由度为n=2的t分布
- 如果原假设成立，则β_1=0，并且t=$\frac{b_1}{s_{b_1}}$
- 我们在显著性水平0.01的情况下，对披萨店的例子进行显著性检验，检验统计量为
  - t = 5/0.5803=8.62
  - 可以计算出对应的p-值为0.00
  - 因为p-值小于0.01，所以我们拒绝原假设，得出β_1不为0的的结论

28 t检验过程

对于简单线性回归情形，总结显著性t检验的步骤

t TEST FOR SIGNIFICANCE IN SIMPLE LINEAR REGRESSION

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

TEST STATISTIC

$$t = \frac{b_1}{s_{b_1}} \qquad (14.19)$$

REJECTION RULE

p-value approach:	Reject H_0 if *p*-value $\leq \alpha$
Critical value approach:	Reject H_0 if $t \leq -t_{\alpha/2}$ or if $t \geq t_{\alpha/2}$

where $t_{\alpha/2}$ is based on a *t* distribution with $n - 2$ degrees of freedom.

29 置信区间

- **β_1的置信区间公式是：**
  - **$b_1 \pm t_{\alpha/2} s_{b_1}$**
  - **b_1是β_1的点估计；$t_{\alpha/2} s_{b_1}$为边际误差。置信系数是1-α; $t_{\alpha/2}$为自由度为n-2时，使t分布的上侧面积为$\alpha/2$的t值**
- 回到案列中，对于置信系数0.01和自由度8，我们查表可知$t_{0.005}$=3.355。 于是β_1的99%置信区间是：
  - $b_1 \pm t_{\alpha/2} s_{b_1}$=5 ±1.95 即 3.05 – 6.95
- 在之前t检验中，β_1的假设值为0，**这不在我们的置信区间中，所以我们拒绝了原假设！**

30 F 检 验

- 在仅有一个自变量的情况下，F检验将得出与t 检验同样的结论
- 如果回归方程有两个或两个以上的自变量，F 检验仅仅能被用来检验回归方程总体的显著关系
- F检验的基本原理是基于建立σ^2的两个独立的估计量：
  - **MSE=**$\frac{SSE}{n-2}$
  - **MSR =**$\frac{SSR}{\text{回归自由度}}$**=**$\frac{SSR}{\text{自变量个数}}$ **均方回归**
- **结论解释**
  - 如果原假设H_0: β_1 = 0 不成立，MSE仍然是σ^2的一个无偏估计量，而MSR高估σ^2，在这种情形下，MSR/MSE的值将变得无穷大
  - 如果H_0成立，MSE和MSR都是σ^2的无偏估计量，在这种情况下， MSR/MSE的值接近于1

31 F 检 验

案例计算

- MSR = SSR/1 = 14200
- MSE = 191.25
- $F = \frac{MSR}{MSE} = 74.25$
- 查表或者计算机得出对应的p-value = 0.00
- 因为p-value小于0.01，所以我们拒绝原假设H_0

F检验计算步骤

F TEST FOR SIGNIFICANCE IN SIMPLE LINEAR REGRESSION

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

TEST STATISTIC

$$F = \frac{\text{MSR}}{\text{MSE}} \qquad (14.21)$$

REJECTION RULE

p-value approach:	Reject H_0 if *p*-value $\leq \alpha$
Critical value approach:	Reject H_0 if $F \geq F_\alpha$

where F_α is based on an *F* distribution with 1 degree of freedom in the numerator and $n - 2$ degrees of freedom in the denominator.

PYTHON解析

```
from statsmodels.formula.api import ols
import pandas as pd

df = pd.read_csv("Armand's.CSV")
df_model = ols("Sales ~ Population", data=df).fit()
print(df_model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Sales   R-squared:                       0.903
Model:                            OLS   Adj. R-squared:                  0.891
Method:                 Least Squares   F-statistic:                     74.25
Date:                Sun, 19 Aug 2018   Prob (F-statistic):           2.55e-05
Time:                        10:18:12   Log-Likelihood:                -39.342
No. Observations:                  10   AIC:                             82.68
Df Residuals:                       8   BIC:                             83.29
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     60.0000      9.226      6.503      0.000      38.725      81.275
Population     5.0000      0.580      8.617      0.000       3.662       6.338
==============================================================================
Omnibus:                        0.928   Durbin-Watson:                   3.224
Prob(Omnibus):                  0.629   Jarque-Bera (JB):                0.616
Skew:                          -0.060   Prob(JB):                        0.735
Kurtosis:                       1.790   Cond. No.                         33.6
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

判定系数

F检验

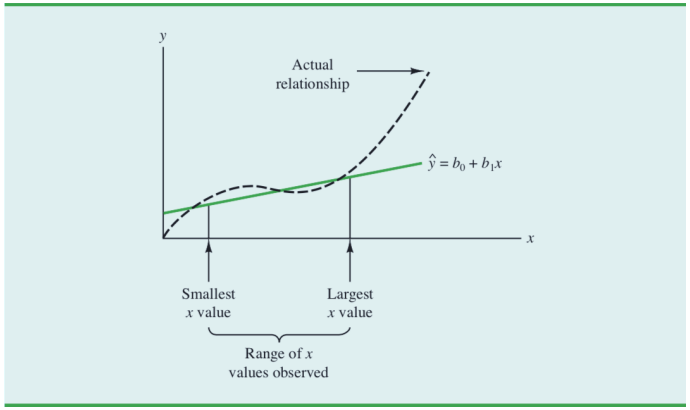
t检验

置信区间

33 显著性解释

- 拒绝原假设H_0: β_1 = 0 并且得出变量x和y之间存在显著性关系的结论，并不意味着我们能得出变量x和y之间存在一个因果关系的结论
- 我们仅仅是证实了变量x和y之间存在统计显著性关系，但这并不能让我们得出变量x和y之间存在线性关系的结论！我们仅仅能说明在x的样本观测值范围内，x和y是相关的。而且这个线性关系只是在x的样本观测值范围内，解释了y的变异性的显著部分

我们利用估计的回归方程对于x的样本观测值范围以内的x值进行预测，应该是完全有把握的。但是超过这一范围就需要十分谨慎！



34 随 堂 练 习

对于一个大都市城区，当地交通管理部门想要确定公共汽车的使用时间和年维修费用之间是否存在某种关系。由10辆公共汽车组成一个样本，收集到的数据如下所示

汽车使用时间(年)	年5维修费用（美元）
1	350
2	370
2	480
2	520
2	590
3	550
4	750
4	800
5	790
5	950

1. 利用最小二乘法，建立估计的回归方程
2. 在α =0.05的显著性水平下，通过检验能否看出两变量之间存在一个显著的关系
3. 最小二乘回归线对观测数据的拟合好么？请做出解释

35 随堂练习-Excel解法

06

残 差 分 析

37 残差分析：证实模型假定

第i次观测的残差

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

y_i代表因变量的观测值；$\hat{y}_i$代表因变量的预测值

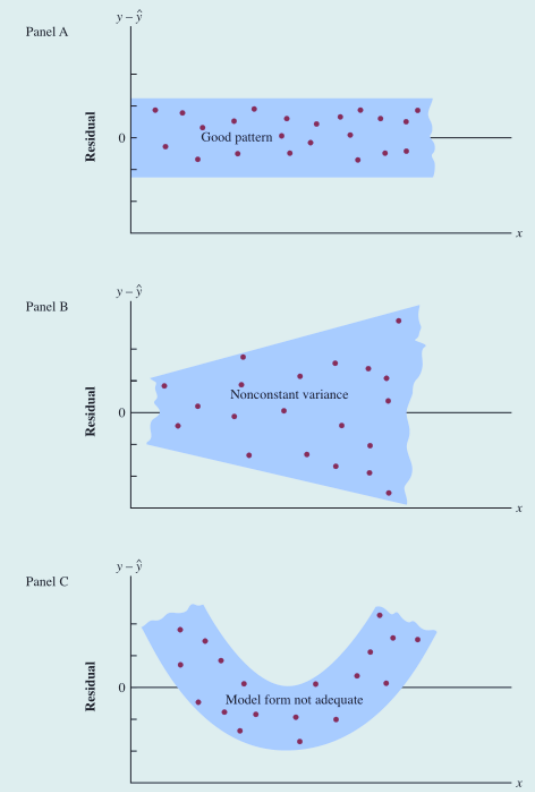
- 这些假定对于显著性检验、置信区间估计都提供了理论上的依据。如果关于误差项ε的假定显得不那么可靠，那么有关回归分析的结果可能会站不住脚
- 残差提供了有关误差项ε的重要信息。因此，残差分析时确定误差项ε的假定是否成立的重要步骤
- 残差分析都是在对残差图形仔细考察的基础上完成的

Armand的残差数据

Student Population x_i	Sales y_i	Estimated Sales $\hat{y}_i = 60 + 5x_i$	Residuals $y_i - \hat{y}_i$
2	58	70	−12
6	105	90	15
8	88	100	−12
8	118	100	18
12	117	120	−3
16	137	140	−3
20	157	160	−3
20	169	160	9
22	149	170	−21
26	202	190	12

38 关于x的残差

三种回归研究的残差图

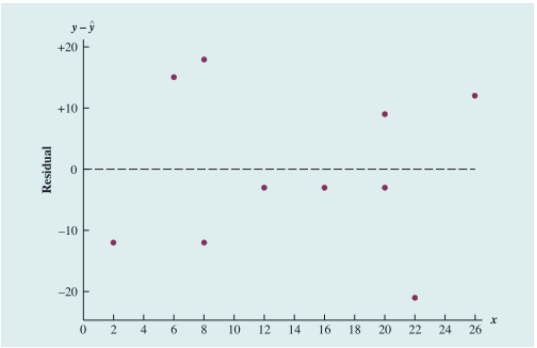


残差图（residual plot）：横轴表示自变量的值，纵轴表示对应的残差值。

- Panel A示意所有的散点都落在一条水平带中间。
- Panel B示意我们违背了残差有一个相同的常数方差的假定。
- Panel C示意回归模型不恰当

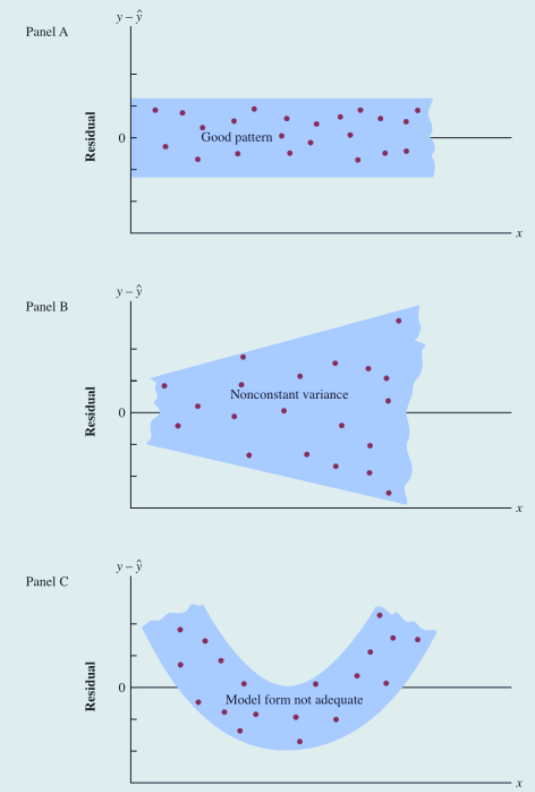
残差的分布接近于Panel A。因此，我们的结论是残差图并没有提供足够的证据，使我们队Armand的回归模型所做的假定表示怀疑

案例的残差图



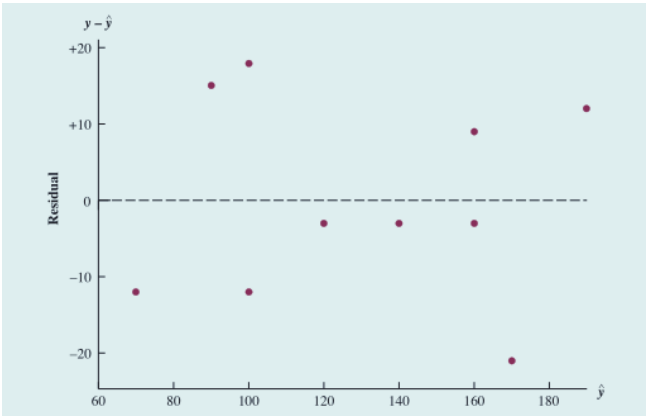
39 关于$\hat{y}$的残差

三种回归研究的残差图



- $\hat{y}$的残差图与x的模式一致，我们无法对模型的假定产生怀疑。
- 对于多元回归分析，因为有一个以上的自变量，所以$\hat{y}$的残差图更适用

案例的残差图



40 标准化残差

第i 个残差的标准差

STANDARD DEVIATION OF THE ith RESIDUAL*

$$s_{y_i-\hat{y}_i} = s\sqrt{1-h_i}$$

where

$s_{y_i-\hat{y}_i}$ = the standard deviation of residual i

s = the standard error of the estimate

$$h_i = \frac{1}{n} + \frac{(x_i-\bar{x})^2}{\Sigma(x_i-\bar{x})^2}$$

标准化残差的计算

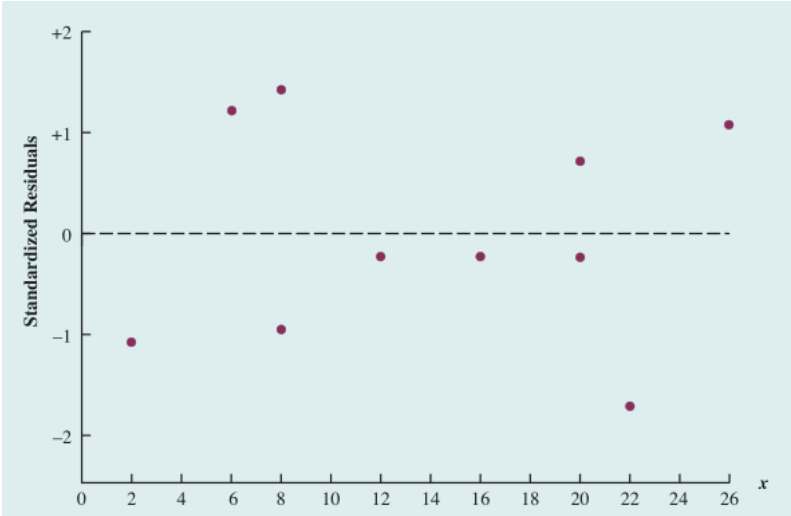
Restaurant i	x_i	$x_i-\bar{x}$	$(x_i-\bar{x})^2$	$\frac{(x_i-\bar{x})^2}{\Sigma(x_i-\bar{x})^2}$	h_i	$s_{y_i-\hat{y}_i}$	$y_i-\hat{y}_i$	Standardized Residual
1	2	−12	144	.2535	.3535	11.1193	−12	−1.0792
2	6	−8	64	.1127	.2127	12.2709	15	1.2224
3	8	−6	36	.0634	.1634	12.6493	−12	−.9487
4	8	−6	36	.0634	.1634	12.6493	18	1.4230
5	12	−2	4	.0070	.1070	13.0682	−3	−.2296
6	16	2	4	.0070	.1070	13.0682	−3	−.2296
7	20	6	36	.0634	.1634	12.6493	−3	−.2372
8	20	6	36	.0634	.1634	12.6493	9	.7115
9	22	8	64	.1127	.2127	12.2709	−21	−1.7114
10	26	12	144	.2535	.3535	11.1193	12	1.0792
		Total	568					

第i次观测的标准化残差

$$\frac{y_i-\widehat{y_i}}{s_{y_i-\widehat{y_i}}}$$

41 标准化残差

关于自变量x的标准化残差图



- 标准化残差图能对随机误差项ε服从正态分布的假定提供一种直观的认识
- 如果这一假定被满足，那么标准化残差分析分布看起来也应该服从一个标准正态分布

42 正态概率图

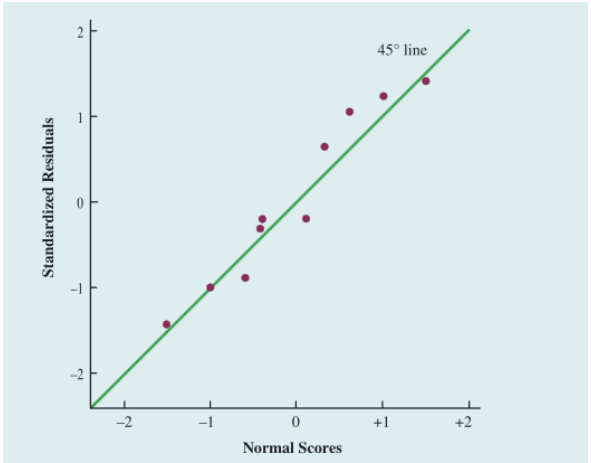
什么是正态分数？

1. 从均值为0，标准差为1的标准正态概率分布的数据中随机抽取10个数值
2. 重复这一抽样多次，然后把每个样本中的数据进行排序
3. 选取每个样本中的最小值，组成一个随机变量，被称为一阶顺序统计量
4. 统计学家已经证明：来自标准正态概率分布的容量为10的样本，一阶顺序统计量的期望值为-1.55。这个期望值被称为正态分数
5. **一般的，如果数据集由n个观测值组成，那么就有n个顺序统计量和n个正态分数**

Armand案例分析

- 对于Armand案例，我们将10个正态分数和10个排好顺序的标准化残差放在一起。如果正态性的假设被满足，那么最小的标准化残差应接近最小的正态分数，以此类推。
- 横轴表示正态分数，纵轴表示标准残差。如果符合正态分布，那么这些点应该围绕在45度直线附近
- 从图中可以看出，随机误差项服从标准正态概率分布的假定是合理的

正态概率图



43 残差分析总结

- 我们利用残差和正态概率图来证实一个回归模型的假定。如果我们的检查表明一个或几个假定是不可靠的，那么我们就应该考虑一个不同的回归模型或者一个数据变换。
- 证实回归模型的假定成立的主要方法是残差分析。即使没有发现假定被违背，但是这并不一定意味着模型将能给出一个好的预测。然而，如果有补充的统计检验支持显著性结论，并且有较大的判定系数，那么我们的模型就有一个比较好的预测。