

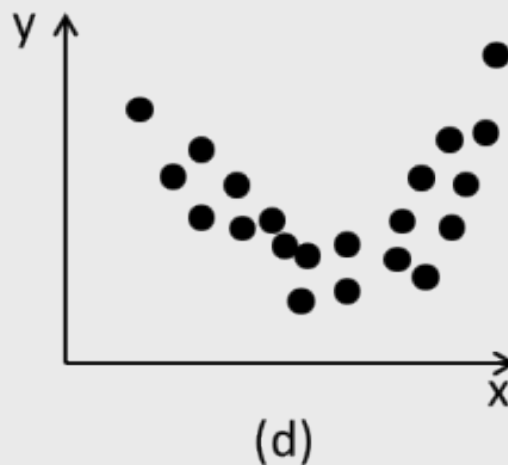
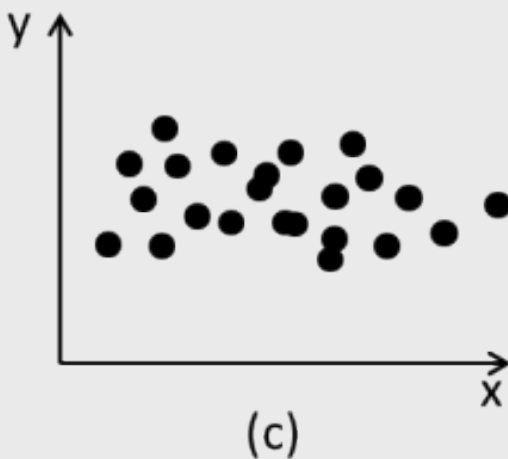
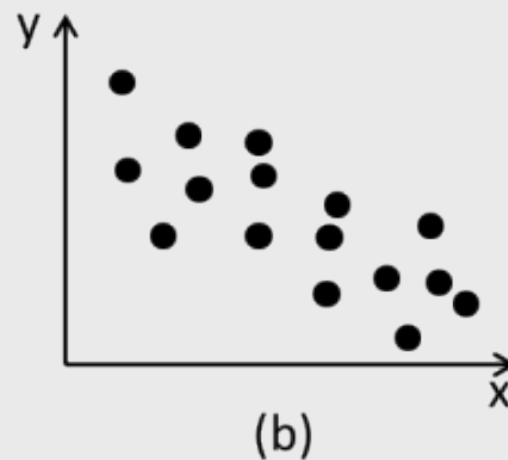
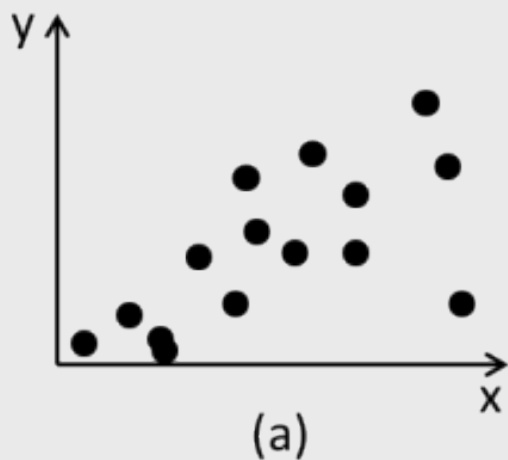


# 多元线性回归模型

讲师：刘顺祥

1. 一元线性回归模型的介绍与应用
2. 多元线性回归模型的系数推导
3. 线性回归模型的假设检验

## 相关分析



## 相关分析

$$\rho = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}$$

~~0.8~~  
||

~~0.5~~  
||

< ||

0.8

~~0.5~~  
||

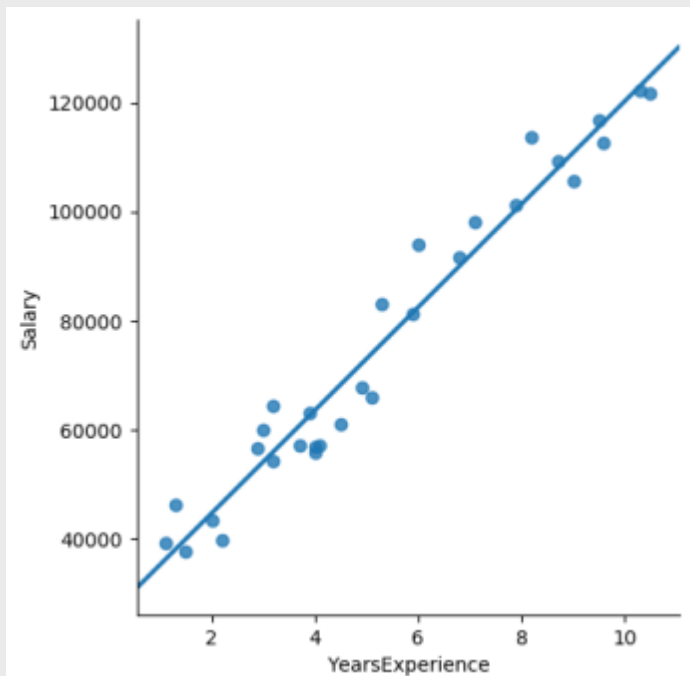
< ||

0.5

~~0.5~~  
||

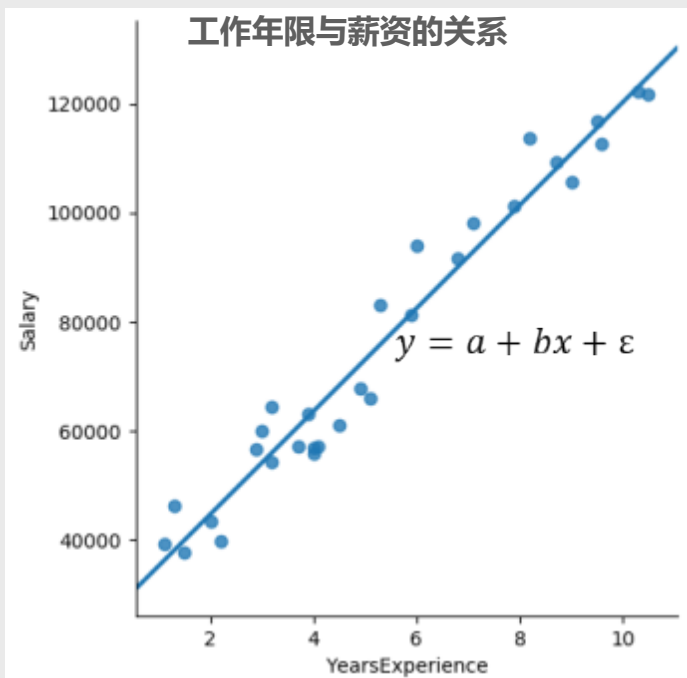
0.5	0.5	0.5	0.5
-----	-----	-----	-----

## 回归分析



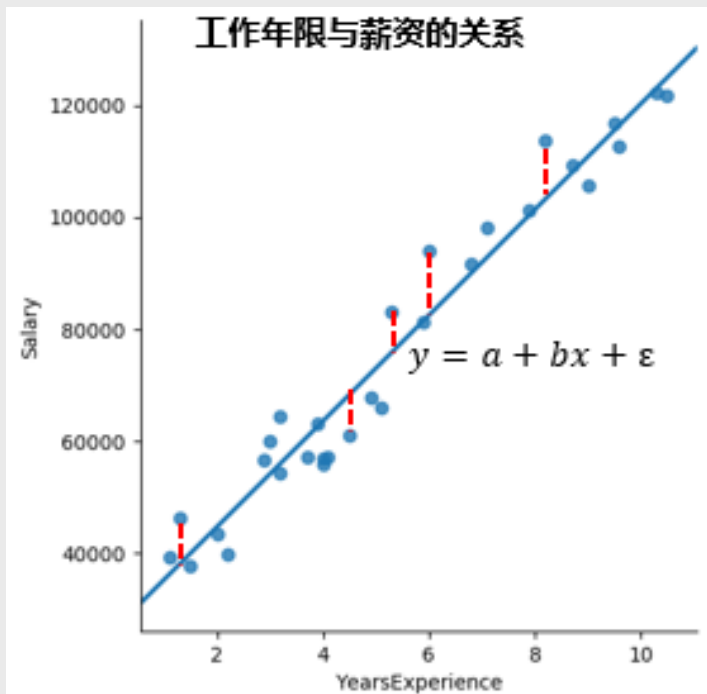
- 1、两边变量之间存在明显的线性关系；
- 2、根据常识，工作年限是因，薪资水平是果；
- 3、是否存在某个模型（即图中的一次函数）可以刻画两个变量之间的关系呢？

## 一元线性回归模型



- 1、模型中的 $x$ 称为自变量， $y$ 称为因变量；
- 2、 $a$ 为模型的截距项， $b$ 为模型的斜率项， $\varepsilon$ 为模型的误差项；
- 3、误差项 $\varepsilon$ 的存在主要是为了平衡等号两边的值，通常被称为模型无法解释的部分；

## 参数 $a$ 和 $b$ 的求解



求解思路：

- 1、如果拟合线能够精确地捕捉到每一个点（即所有散点全部落在拟合线上），那么对应的误差项 $\varepsilon$ 应该为0；
- 2、所以，模型拟合的越好，则误差项 $\varepsilon$ 应该越小。进而可以理解为：求解参数的问题便是求解误差平方和最小的问题；

## 参数 $a$ 和 $b$ 的求解

$$J(a, b) = \sum_{i=1}^n \varepsilon^2 = \sum_{i=1}^n (y_i - [a + bx_i])^2$$

- 1、 $J(a, b)$ 为目标函数，需求解该函数的最小值。
- 2、求解方法便是计算目标函数关于参数 $a$ 和 $b$ 的两个偏导数，最终令偏导数为0即可。



## 参数 $a$ 和 $b$ 的求解

✦ 展开目标函数中的平方项

$$J(a, b) = \sum_{i=1}^n (y_i^2 + a^2 + b^2 x_i^2 + 2abx_i - 2ay_i - 2bx_i y_i)$$

✦ 计算参数 $a$ 和 $b$ 的偏导数，并令导函数为0

$$\begin{cases} \frac{\partial J}{\partial a} = \sum_{i=1}^n (0 + 2a + 0 + 2bx_i - 2y_i + 0) = 0 \\ \frac{\partial J}{\partial b} = \sum_{i=1}^n (0 + 0 + 2bx_i^2 + 2ax_i + 0 - 2x_i y_i) = 0 \end{cases}$$

## 参数 $a$ 和 $b$ 的求解

✚ 公式转换

$$\begin{cases} \frac{\partial J}{\partial a} = 2na + 2b \sum_{i=1}^n x_i - 2 \sum_{i=1}^n y_i = 0 \\ \frac{\partial J}{\partial b} = 2b \sum_{i=1}^n x_i^2 + 2a \sum_{i=1}^n x_i - 2 \sum_{i=1}^n x_i y_i = 0 \end{cases}$$

✚ 化解导函数为0的等式

$$\begin{cases} a = \frac{\sum_{i=1}^n y_i}{n} - \frac{b \sum_{i=1}^n x_i}{n} \\ b \sum_{i=1}^n x_i^2 + \left( \frac{\sum_{i=1}^n y_i}{n} - \frac{b \sum_{i=1}^n x_i}{n} \right) \sum_{i=1}^n x_i - \sum_{i=1}^n x_i y_i = 0 \end{cases}$$

## 参数 $a$ 和 $b$ 的求解

✦ 系数值

$$\begin{cases} a = \bar{y} - b\bar{x} \\ b = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2} \end{cases}$$

## 模型的应用

```
# 导入第三方模块
import statsmodels.api as sm
sm.ols(formula, data, subset=None, drop_cols=None)
```

formula：以字符串的形式指定线性回归模型的公式，如'y~x'就表示简单线性回归模型

data：指定建模的数据集

subset：通过bool类型的数组对象，获取data的子集用于建模

drop\_cols：指定需要从data中删除的变量

## 模型的应用

```
# 导入第三方模块
import pandas as pd
import statsmodels.api as sm

income = pd.read_csv('Salary_Data.csv')
# 利用收入数据集，构建回归模型
fit = sm.formula.ols('Salary ~ YearsExperience', data = income).fit()
# 返回模型的参数值
fit.params
```

```
out:
Intercept      25792.200199
YearsExperience 9449.962321
dtype: float64
```

模型结果:

$\text{Salary} = 25792.20 + 9449.96\text{YearsExperience}$

## 多元线性回归模型的定义

对于一元线性回归模型来说，其反映的是单个自变量对因变量的影响，然而实际情况中，影响因变量的自变量往往不止一个，从而需要将一元线性回归模型扩展到多元线性回归模型。

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

其中， $x_{ij}$ 代表第*i*行的第*j*变量值。如果按照一元线性回归模型的逻辑，那么多元线性回归模型应该就是因变量y与自变量X的线性组合。

## 多元线性回归模型的定义

所以，基于一元线性回归模型的扩展，可以将多元线性回归模型表示为：

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

进一步，根据线性代数的知识，可以将上式表示为矩阵的形式：

$$y = X\beta + \varepsilon$$

## 多元线性回归模型的参数求解

✦ 构造目标函数

$$J(\beta) = \sum \varepsilon^2 = \sum (y - X\beta)^2$$

✦ 展开平方项

$$\begin{aligned} J(\beta) &= (y - X\beta)'(y - X\beta) \\ &= (y' - \beta'X')(y - X\beta) \\ &= (y'y - y'X\beta - \beta'X'y + \beta'X'X\beta) \end{aligned}$$



## 多元线性回归模型的参数求解

✦ 求偏导

$$\frac{\partial J(\beta)}{\partial \beta} = (0 - X'y - X'y + 2X'X\beta) = 0$$

✦ 计算偏回归系数

$$\begin{aligned} X'X\beta &= X'y \\ \beta &= (X'X)^{-1}X'y \end{aligned}$$

## 多元线性回归模型的预测

RD_Spend	Administration	Marketing_Spend	State	Profit
165349.2	136897.8	471784.1	New York	192261.83
162597.7	151377.59	443898.53	California	191792.06
153441.51	101145.55	407934.54	Florida	191050.39
144372.41	118671.85	383199.62	New York	182901.99
142107.34	91391.77	366168.42	Florida	166187.94
131876.9	99814.71	362861.36	New York	156991.12
134615.46	147198.87	127716.82	California	156122.51
130298.13	145530.06	323876.68	Florida	155752.6
120542.52	148718.95	311613.29	New York	152211.77
123334.88	108679.17	304981.62	California	149759.96
101913.08	110594.11	229160.95	Florida	146121.95
100671.96	91790.61	249744.55	California	144259.4
93863.75	127320.38	249839.44	Florida	141585.52

数据集包含5个变量，分别是产品的研发成本、管理成本、市场营销成本、销售市场和销售利润。

## 多元线性回归模型的应用

```
# 导入模块
from sklearn import model_selection

# 导入数据
Profit = pd.read_excel(r'C:\Users\Administrator\Desktop\Predict to Profit.xlsx')
# 将数据集拆分为训练集和测试集
train, test = model_selection.train_test_split(Profit, test_size = 0.2, random_state=1234)
# 根据train数据集建模
model = sm.formula.ols('Profit ~ RD_Spend+Administration+Marketing_Spend+C(State)', data
= train).fit()

print('模型的偏回归系数分别为：\n', model.params)
# 删除test数据集中的Profit变量，用剩下的自变量进行预测
test_X = test.drop(labels = 'Profit', axis = 1)
pred = model.predict(exog = test_X)

print('对比预测值和实际值的差异：\n',pd.DataFrame({'Prediction':pred,'Real':test.Profit}))
```

## 多元线性回归模型的应用

模型的偏回归系数分别为：

Intercept	58581.516503
C(State)[T.Florida]	927.394424
C(State)[T.New York]	-513.468310
RD_Spend	0.803487
Administration	-0.057792
Marketing_Spend	0.013779
dtype: float64	

默认情况下，对于离散变量State而言，模型选择California值作为对照组。

对比预测值和实际值的差异：

	<b>Prediction</b>	<b>Real</b>
8	150621.345802	152211.77
48	55513.218079	35673.41
14	150369.022458	132602.65
42	74057.015562	71498.49
29	103413.378282	101004.64
44	67844.850378	65200.33
4	173454.059692	166187.94
31	99580.888894	97483.56
13	128147.138397	134307.35
18	130693.433835	124266.90

## 多元线性回归模型的预测

```
# 生成由State变量衍生的哑变量
dummies = pd.get_dummies(Profit.State)
# 将哑变量与原始数据集水平合并
Profit_New = pd.concat([Profit,dummies], axis = 1)
# 删除State变量和California变量 ( 因为State变量已被分解为哑变量 , New York变量需要作为参照组 )
Profit_New.drop(labels = ['State','New York'], axis = 1, inplace = True)

# 拆分数据集Profit_New
train, test = model_selection.train_test_split(Profit_New, test_size = 0.2, random_state=1234)
# 建模
model2 =
sm.formula.ols('Profit~RD_Spend+Administration+Marketing_Spend+Florida+California', data
= train).fit()
print('模型的偏回归系数分别为 : \n', model2.params)
```

## 多元线性回归模型的应用

模型的偏回归系数分别为：

Intercept	58068.048193
RD_Spend	0.803487
Administration	-0.057792
Marketing_Spend	0.013779
Florida	1440.862734
California	513.468310

dtype: float64

左侧模型是通过哑变量的方式，自定义离散变量State  
中New York值作为参照组。

Profit  
= 58068.05 + 0.80RD\_Spend - 0.06Administation  
+ 0.01Marketing\_Spend + 1440.86Florida  
+ 513.47California

## 模型的F检验

- ✦ 提出问题的原假设和备择假设
- ✦ 在原假设的条件下，构造统计量F
- ✦ 根据样本信息，计算统计量的值
- ✦ 对比统计量的值和理论F分布的值，当统计量值超过理论值时，拒绝原假设，否则接受原假设

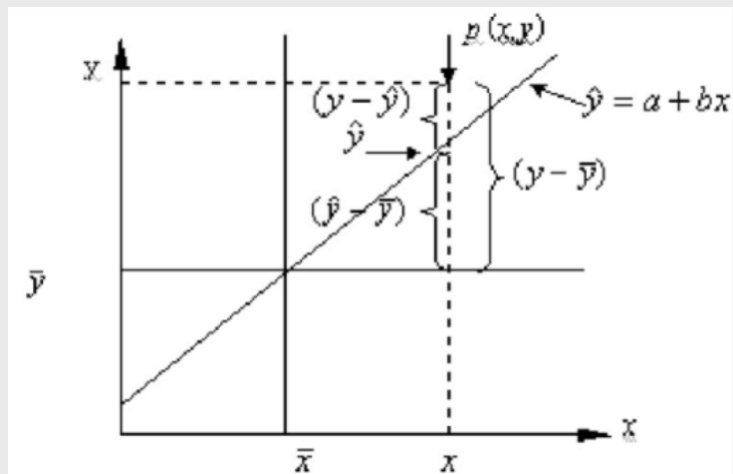
## 模型的F检验

### ✦ 提出假设

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

### ✦ 构造统计量


$$F = \frac{RSS/p}{ESS/(n-p-1)} \sim F(p, n-p-1)$$



$$\begin{cases} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = ESS \\ \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = RSS \\ \sum_{i=1}^n (y_i - \bar{y})^2 = TSS \end{cases}$$



## 模型的F检验

 计算统计量

```
# 导入第三方模块
import numpy as np

# 计算建模数据中因变量的均值
ybar = train.Profit.mean()
# 统计变量个数和观测个数
p = model2.df_model
n = train.shape[0]
# 计算回归离差平方和
RSS = np.sum((model2.fittedvalues-ybar) ** 2)
```

```
# 计算误差平方和
ESS = np.sum(model2.resid ** 2)
# 计算F统计量的值
F = (RSS/p)/(ESS/(n-p-1))
print('F统计量的值：',F)
```

**out:**  
F统计量的值： 174.6372

## 模型的F检验

✈ 对比结果下结论

```
# 导入模块
from scipy.stats import f

# 计算F分布的理论值
F_Theroy = f.ppf(q=0.95, dfn = p,dfd = n-p-1)
print('F分布的理论值为：',F_Theroy)

out:
F分布的理论值为： 2.5026
```

计算出来的F统计量值174.64远远大于F分布的理论值2.50，所以应当**拒绝原假设**，即认为多元线性回归模型是显著的，也就是说回归模型的偏回归系数都不全为0。

## 参数的t检验

✦ 提出假设

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

✦ 构造统计量

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

$$\text{其中, } se(\hat{\beta}_j) = \sqrt{c_{jj} \frac{\sum \varepsilon_i^2}{n - p - 1}}$$

## 参数的t检验

✦ 计算统计量

```
# 有关模型的概览信息  
model2.summary()
```

✦ 对比结果下结论

从右侧返回的结果可知，只有截距项Intercept和研发成本RD\_Spend对应的 $p$ 值小于0.05，才说明**其余变量都没有通过系数的显著性检验**，即在模型中这些变量不是影响利润的重要因素。

OLS Regression Results

Dep. Variable:	Profit	R-squared:	0.964
Model:	OLS	Adj. R-squared:	0.958
Method:	Least Squares	F-statistic:	174.6
Date:	Sat, 03 Mar 2018	Prob (F-statistic):	9.74e-23
Time:	11:47:12	Log-Likelihood:	-401.20
No. Observations:	39	AIC:	814.4
Df Residuals:	33	BIC:	824.4
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P> t	[0.025	0.975]
Intercept	5.807e+04	6846.309	8.482	0.000	4.41e+04	7.2e+04
RD_Spend	0.8035	0.040	19.988	0.000	0.722	0.885
Administration	-0.0578	0.051	-1.133	0.265	-0.162	0.046
Marketing_Spend	0.0138	0.015	0.930	0.359	-0.016	0.044
Florida	1440.8627	3059.931	0.471	0.641	-4784.615	7666.340
California	513.4683	3043.160	0.169	0.867	-5677.887	6704.824

Omnibus:	1.721	Durbin-Watson:	1.896
Prob(Omnibus):	0.423	Jarque-Bera (JB):	1.148
Skew:	0.096	Prob(JB):	0.563
Kurtosis:	2.182	Cond. No.	1.60e+06

# EDU

CSDN学院 IT实战派

