

线性回归模型代码实现

Statsmodels

小胖

目录

ONE 模型实现

Statsmodels、置信区间、假设检验

TWO 预测结果的置信区间

自信的预测

THREE 参数的置信区间

正确理解置信区间的含义

模型实现

Statsmodels

生产记事本

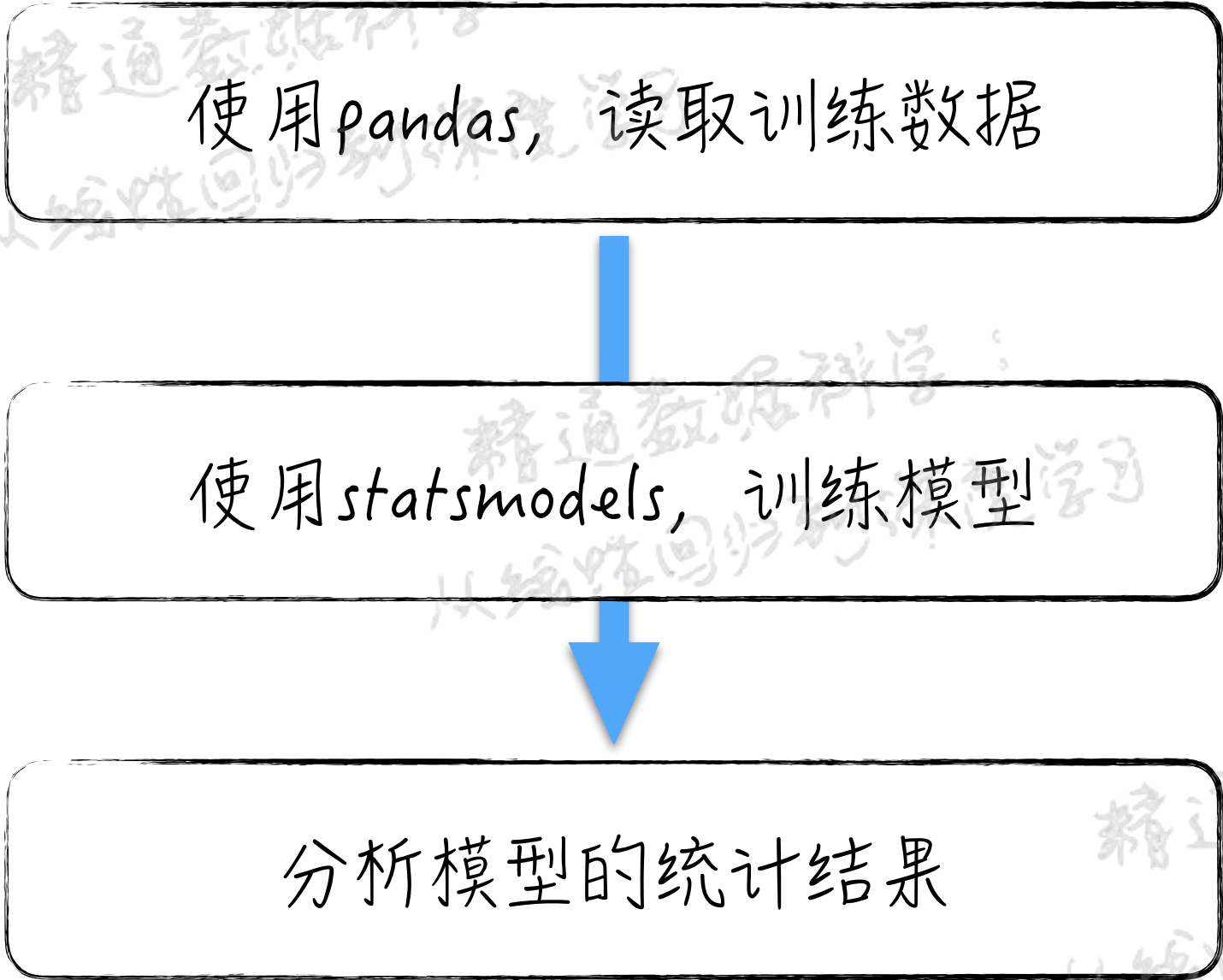
| 日期 | 玩偶个数 | 成本 | 第几天 |
|-------|------|-------|-----|
| 04/01 | 10 | 7.7 | 1 |
| 04/02 | 10 | 9.87 | 2 |
| 04/03 | 11 | 10.87 | 3 |
| 04/04 | 12 | 12.18 | 4 |
| 04/05 | 13 | 11.43 | 5 |
| 04/06 | 14 | 13.36 | 6 |
| 04/07 | 15 | 15.15 | 7 |
| 04/08 | 16 | 16.73 | 8 |
| 04/09 | 17 | 17.4 | 9 |
| ... | ... | ... | ... |

x y

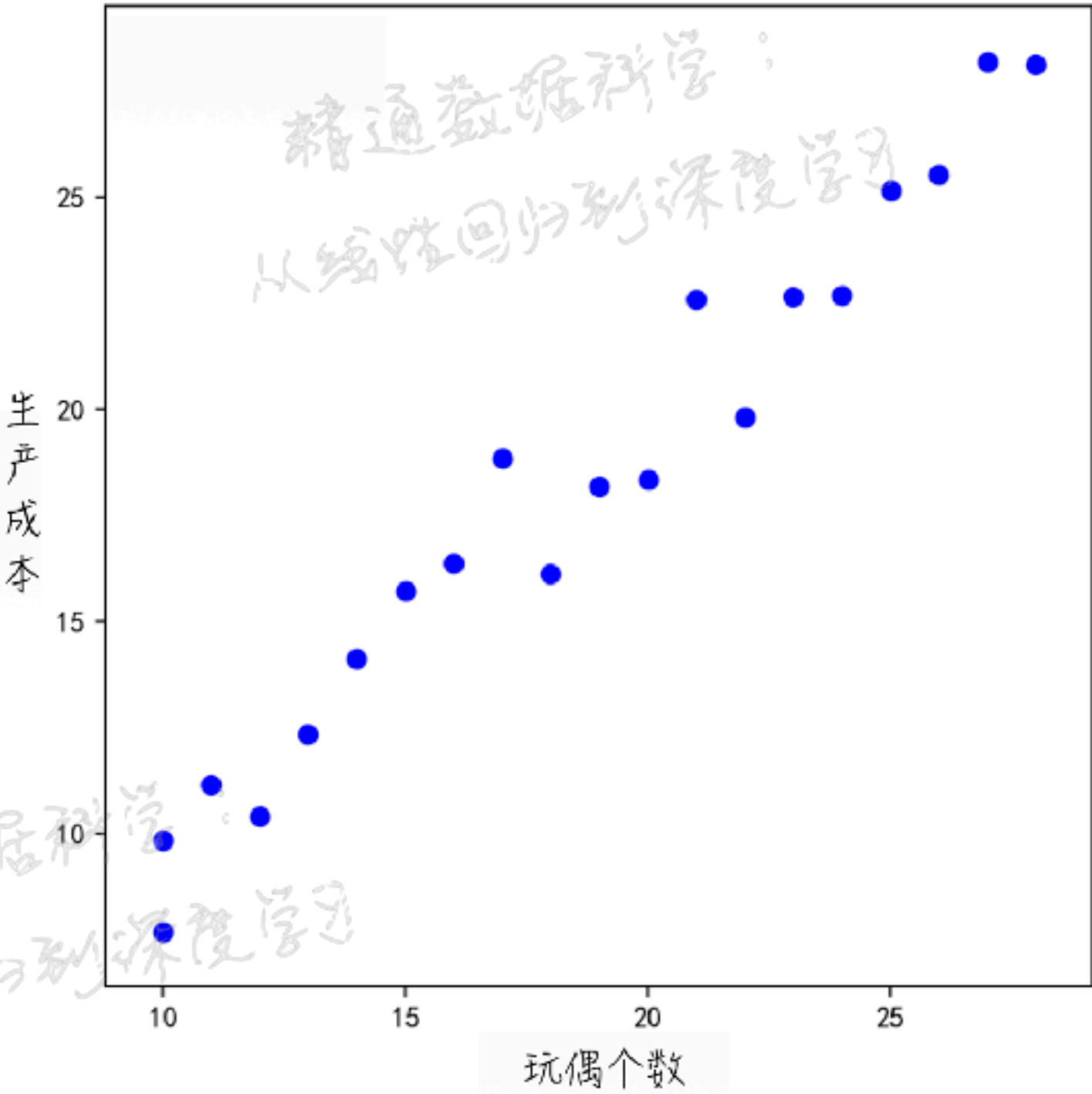
$$y = x + \varepsilon$$

模型假设：

$$y_i = ax_i + b + \varepsilon_i$$



生产数据



模型实现

统计分析结果

模型假设:

$$y_i = ax_i + b + \varepsilon_i$$

• a的真实值等于1

• b的真实值等于0

OLS Regression Results

| | | | |
|-------------------|------------------|---------------------|----------|
| Dep. Variable: | y | R-squared: | 0.962 |
| Model: | OLS | Adj. R-squared: | 0.960 |
| Method: | Least Squares | F-statistic: | 460.5 |
| Date: | Sun, 16 Dec 2018 | Prob (F-statistic): | 2.85e-14 |
| Time: | 15:46:12 | Log-Likelihood: | -31.374 |
| No. Observations: | 20 | AIC: | 66.75 |
| Df Residuals: | 18 | BIC: | 68.74 |
| Df Model: | 1 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P> t | [0.025 | 0.975] |
|-------|---------|---------|--------|-------|--------|--------|
| const | -0.9495 | 0.934 | -1.017 | 0.323 | -2.912 | 1.013 |
| x | 1.0330 | 0.048 | 21.458 | 0.000 | 0.932 | 1.134 |

| | | | |
|----------------|-------|-------------------|-------|
| Omnibus: | 0.745 | Durbin-Watson: | 2.345 |
| Prob(Omnibus): | 0.689 | Jarque-Bera (JB): | 0.673 |
| Skew: | 0.074 | Prob(JB): | 0.714 |
| Kurtosis: | 2.113 | Cond. No. | 66.3 |

参数b

参数a

参数估计值

估计值的标准差

2

数值大于0.05, 变量应该被舍弃

1

95%的可能, 参数所在区间

模型实现

假设检验

模型假设:

$$y_i = ax_i + b + \varepsilon_i$$

• a的真实值等于1

• b的真实值等于0

这行代码表示, 检验的假设为: x变量
的系数等于0 (即 $a=0$) ; 并非 $x=0$

```
# 用f test检测x对应的系数a是否显著  
print res.f_test("x=0")
```

```
<F test: F=array([[ 460.4584822]]), p=2.8484654145e-14, df_denom=18, df_num=1>
```

```
# 用f test检测常量b是否显著
```

```
print res.f_test("const=0")
```

```
<F test: F=array([[ 1.03355794]]), p=0.322795640083, df_denom=18, df_num=1>
```

```
# 用f test检测a=1, b=0同时成立的显著性
```

```
print res.f_test(["x=1", "const=0"])
```

```
<F test: F=array([[ 0.99654631]]), p=0.388626797606, df_denom=18, df_num=2>
```

1

P-value小于0.05。拒绝 $a=0$
这个假设, 即a是显著的

2

P-value大于0.05。不能拒绝 $b=0$
这个假设, 即b是不显著的

3

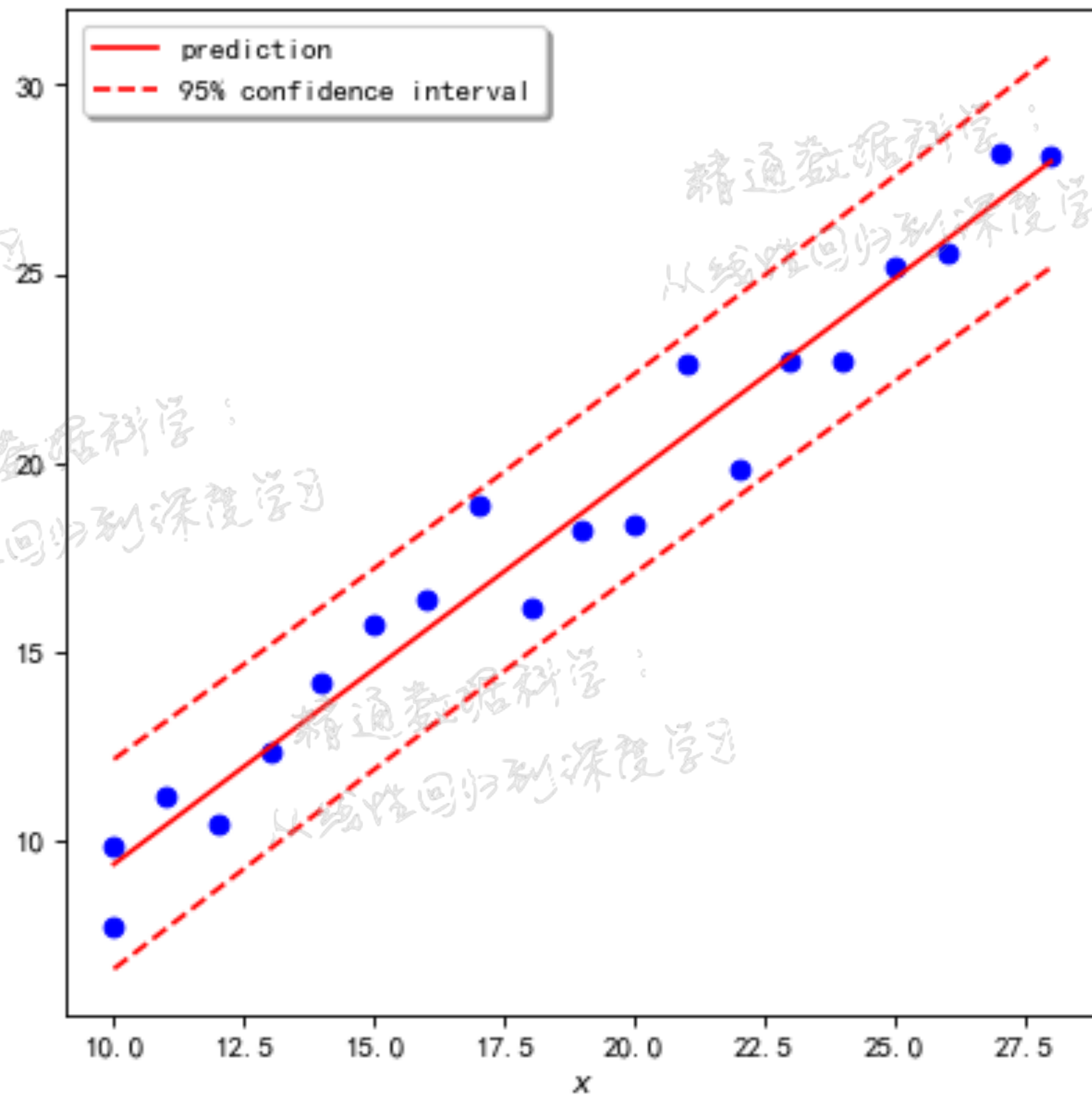
P-value大于0.05。不能拒绝 $b=0, a=1$
这个两个假设同时成立

预测结果的置信区间

自信的预测

模型的预测公式：

$$\hat{y}_i = \hat{a}x_i + \hat{b}$$



参数的置信区间

正确理解置信区间的含义

随机生成训练数据, x 和 y

使用statsmodels, 训练模型

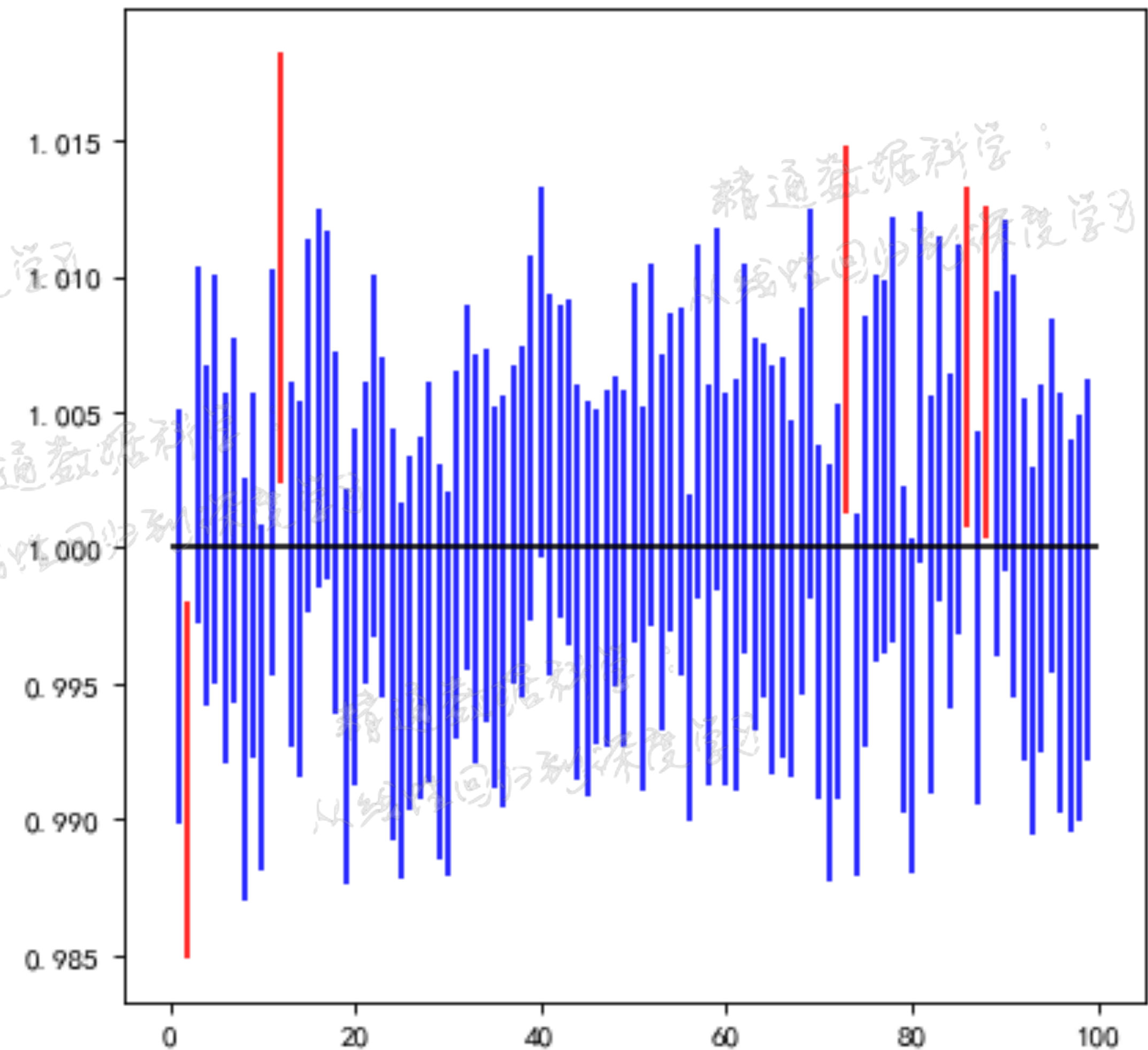
循环100次

记录参数 a 估计值的置信区间

使用matplotlib, 将模型结果可视化

红色竖线表示不包含1的95%置信区间

蓝色竖线表示包含1的95%置信区间



精通数据科学：
从线性回归到深度学习

THANK YOU