Regression in R

Class 1

Kai Yao

Oct, 2016

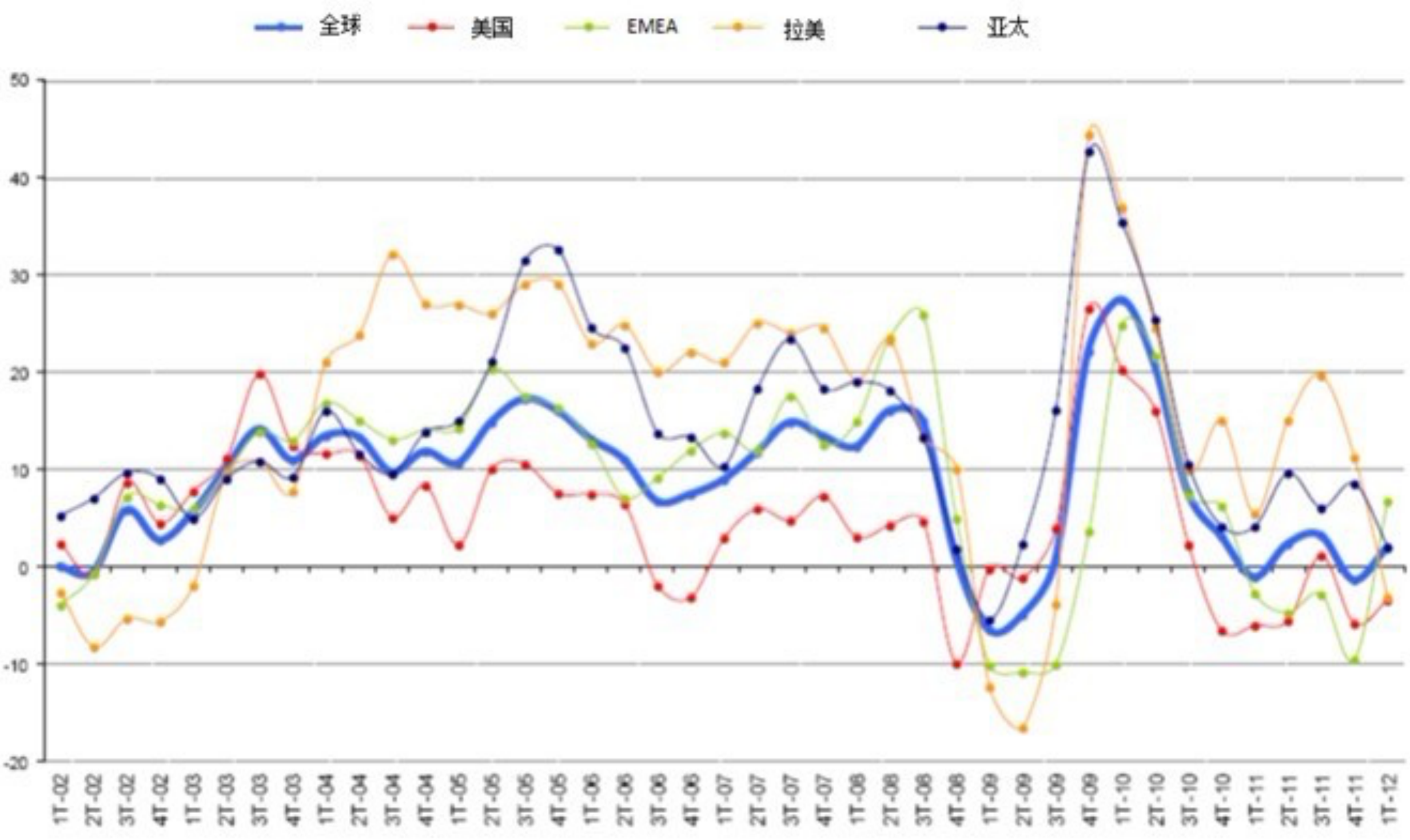
主要内容

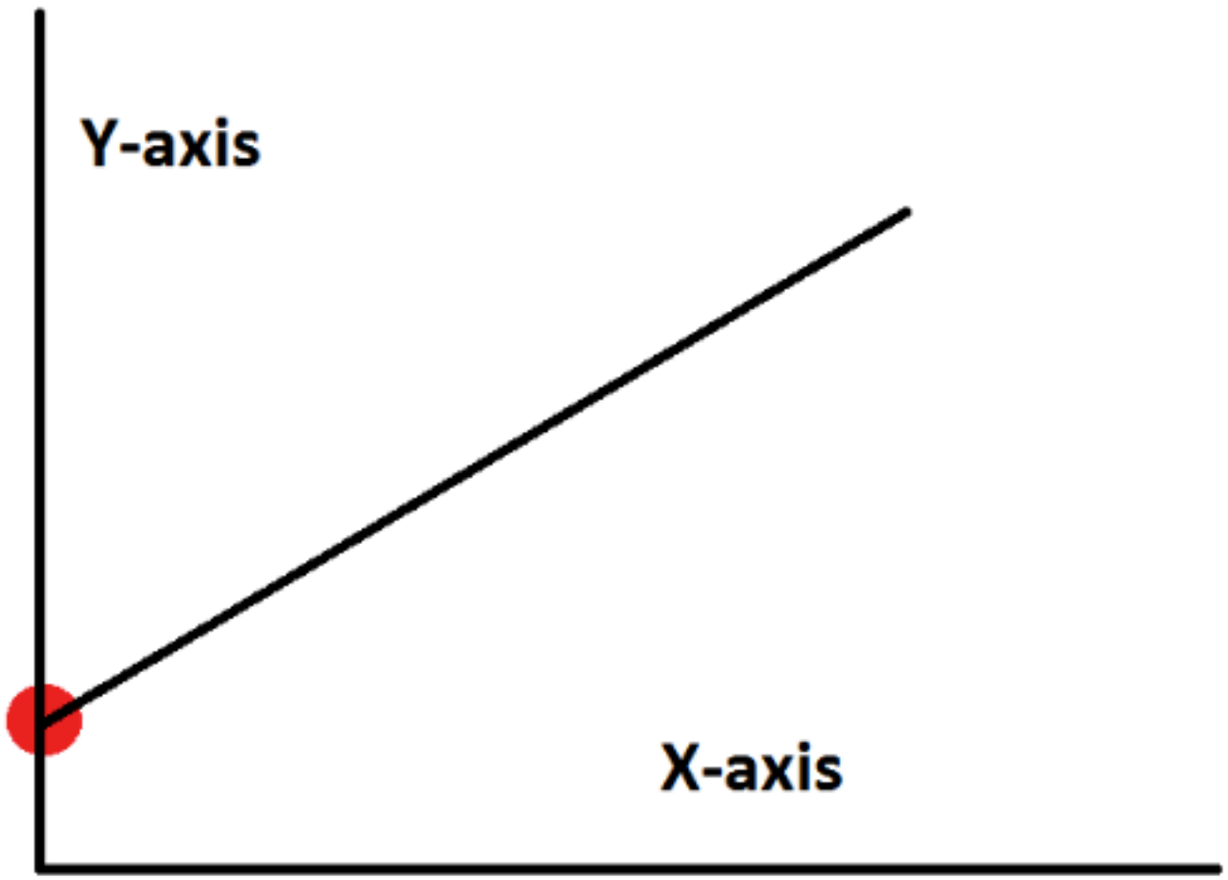
- 线性回归及应用
- 逻辑(0-1)回归及应用
- 作业

线性回归的基本概念

- 线性回归(Linear Regression)是利用称为线性回归方程的最小平方函数对一个或多个自变量和因变量之间关系进行建模的一种回归分析。
- 通常形式为：Y=aX+b，其中Y是连续型变量，X可以是一个或多个解释性变量（自变量）
- 线性回归的作用：
  - 找出X中哪些因素与Y相关
  - 通过a的大小确定单位X对Y的影响程度

线性回归的应用


全球
美国
EMEA
拉美
亚太
50
40
30
20
10
0
-10
-20
1T-02
2T-02
3T-02
4T-02
1T-03
2T-03
3T-03
4T-03
1T-04
2T-04
3T-04
4T-04
1T-05
2T-05
3T-05
4T-05
1T-06
2T-06
3T-06
4T-06
1T-07
2T-07
3T-07
4T-07
1T-08
2T-08
3T-08
4T-08
1T-09
2T-09
3T-09
4T-09
1T-10
2T-10
3T-10
4T-10
1T-11
2T-11
3T-11
4T-11
1T-12

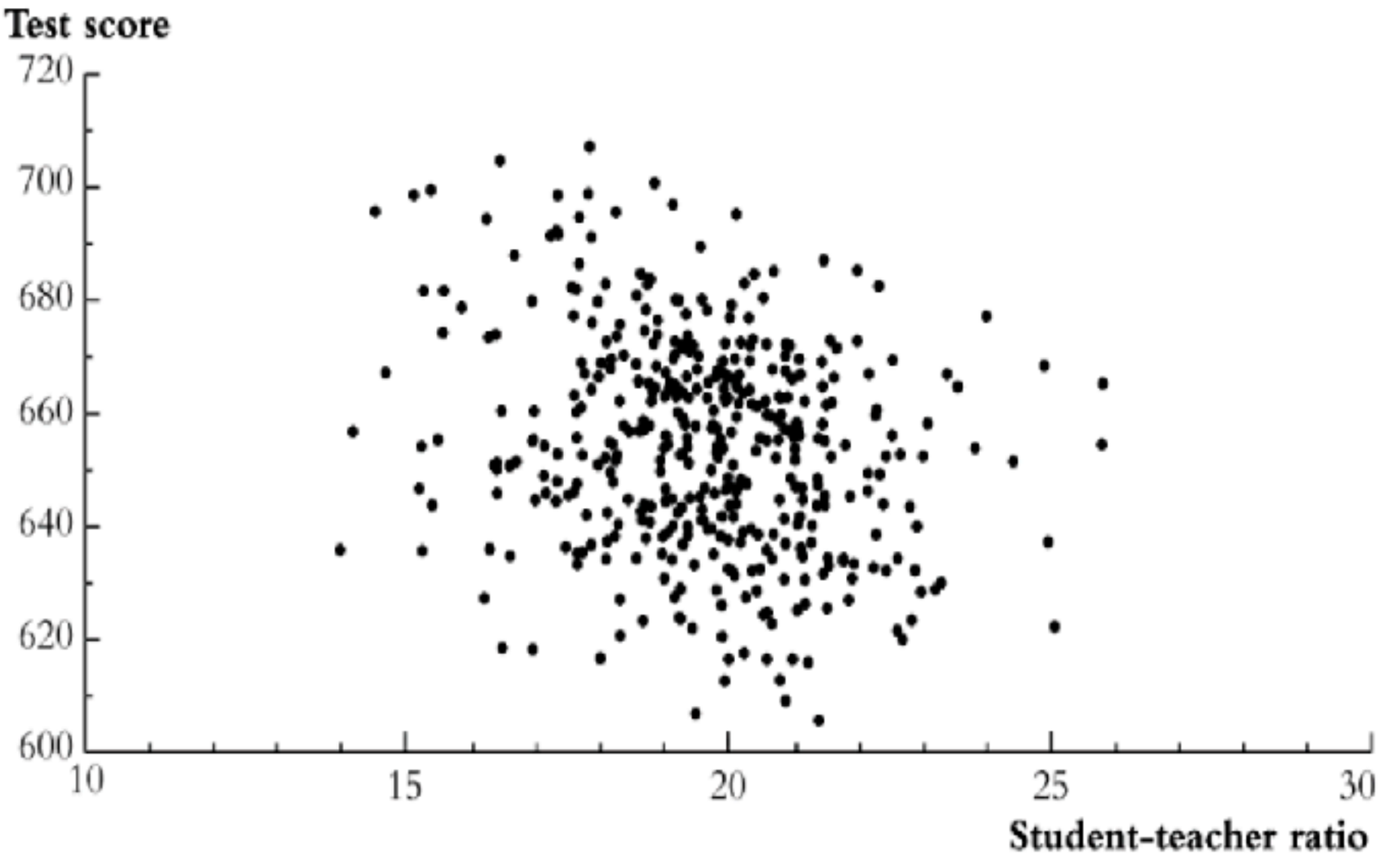


- ε表示误
- β_0表示截
- β表示回归系数，即X对Y的影响程度，每一个回归系数的大小正好是保持其他回归系数对应的自变量不变的情况下，该回归系数对应自变量改变一个单位，因变量对应改变的大小

线性回归示例

FIGURE 4.2 Scatterplot of Test Score vs. Student-Teacher Ratio (California School District Data)

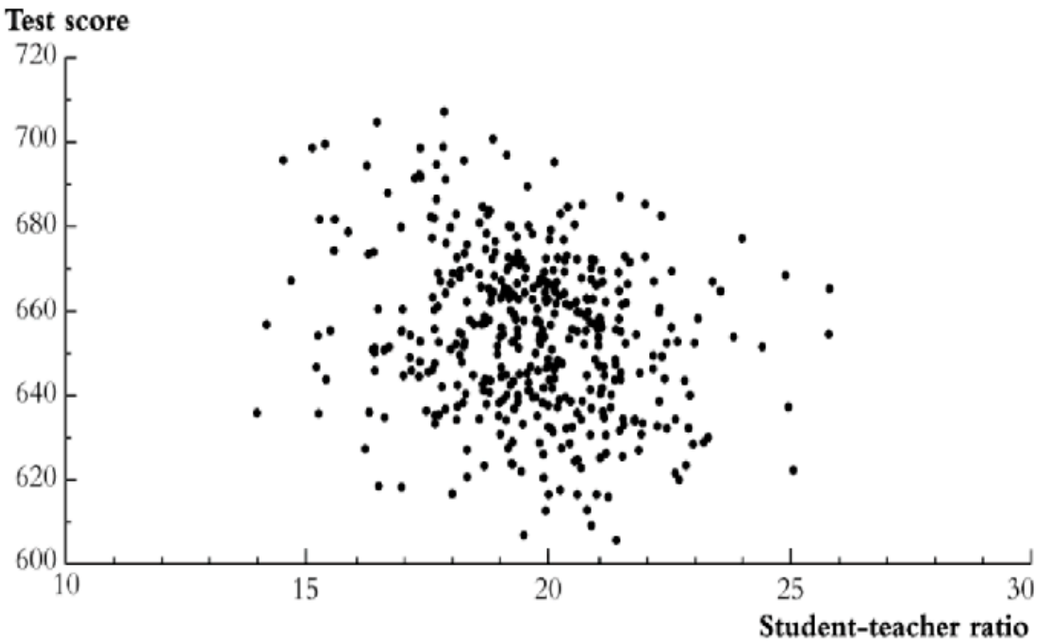
Data from 420 California school districts. There is a weak negative relationship between the student-teacher ratio and test scores: the sample correlation is –0.23.



线性回归示例

FIGURE 4.2 Scatterplot of Test Score vs. Student-Teacher Ratio (California School District Data)

Data from 420 California school districts. There is a weak negative relationship between the student-teacher ratio and test scores: the sample correlation is –0.23.



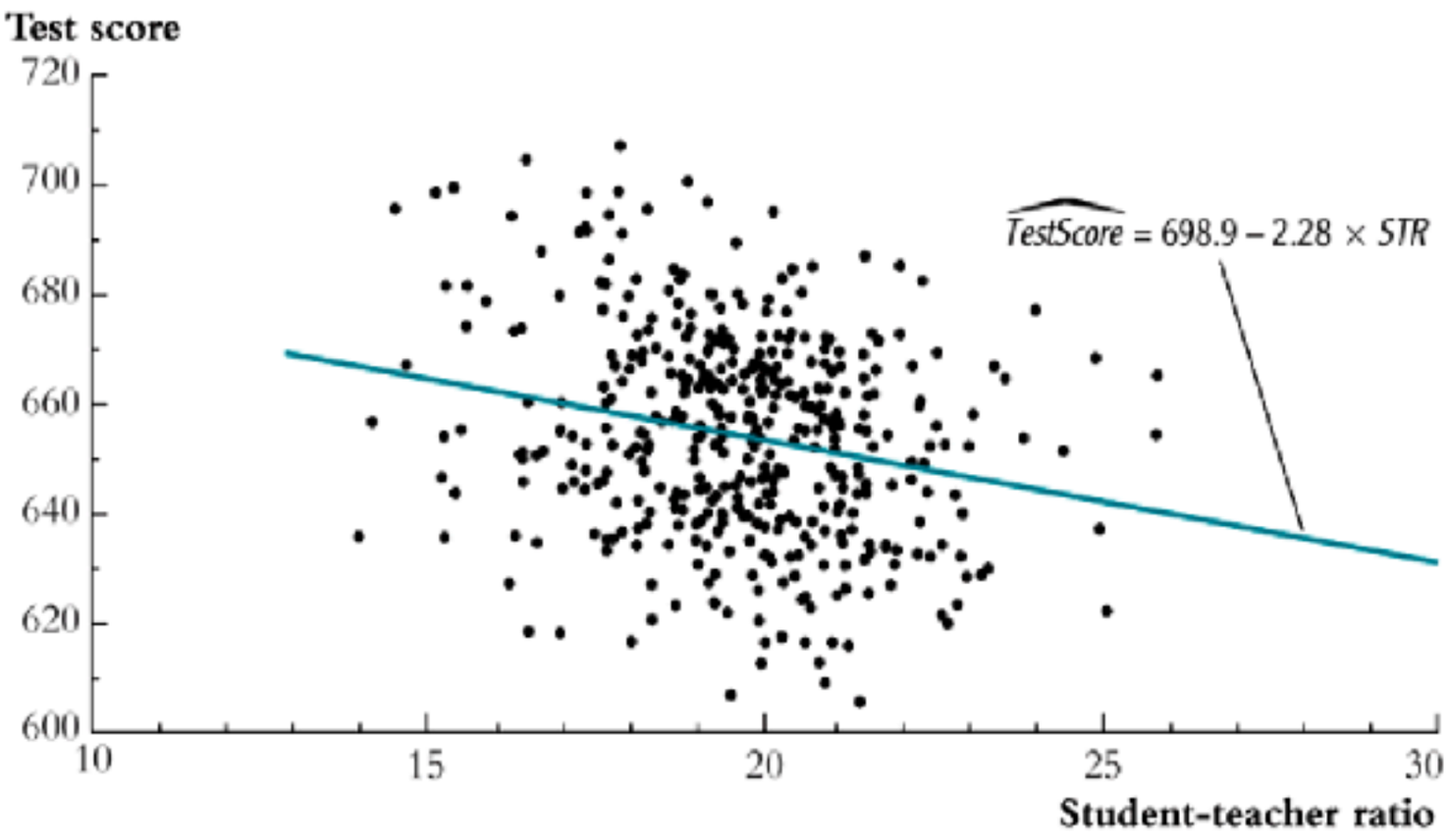
$$Test\ Score = \beta_0 + \beta_1 STR$$

$$\min_{b_0, b_1} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$$

线性回归示例

FIGURE 4.3 The Estimated Regression Line for the California Data

The estimated regression line shows a negative relationship between test scores and the student-teacher ratio. If class sizes fall by 1 student, the estimated regression predicts that test scores will increase by 2.28 points.



Estimated slope $= \hat{\beta}_1 = -2.28$

Estimated intercept $= \hat{\beta}_0 = 698.9$

Estimated regression line: $\widehat{TestScore} = 698.9 - 2.28 \times STR$

判断线性回归模型好坏标准

- R-Square：用来表示线性回归模型匹配的好坏

38%

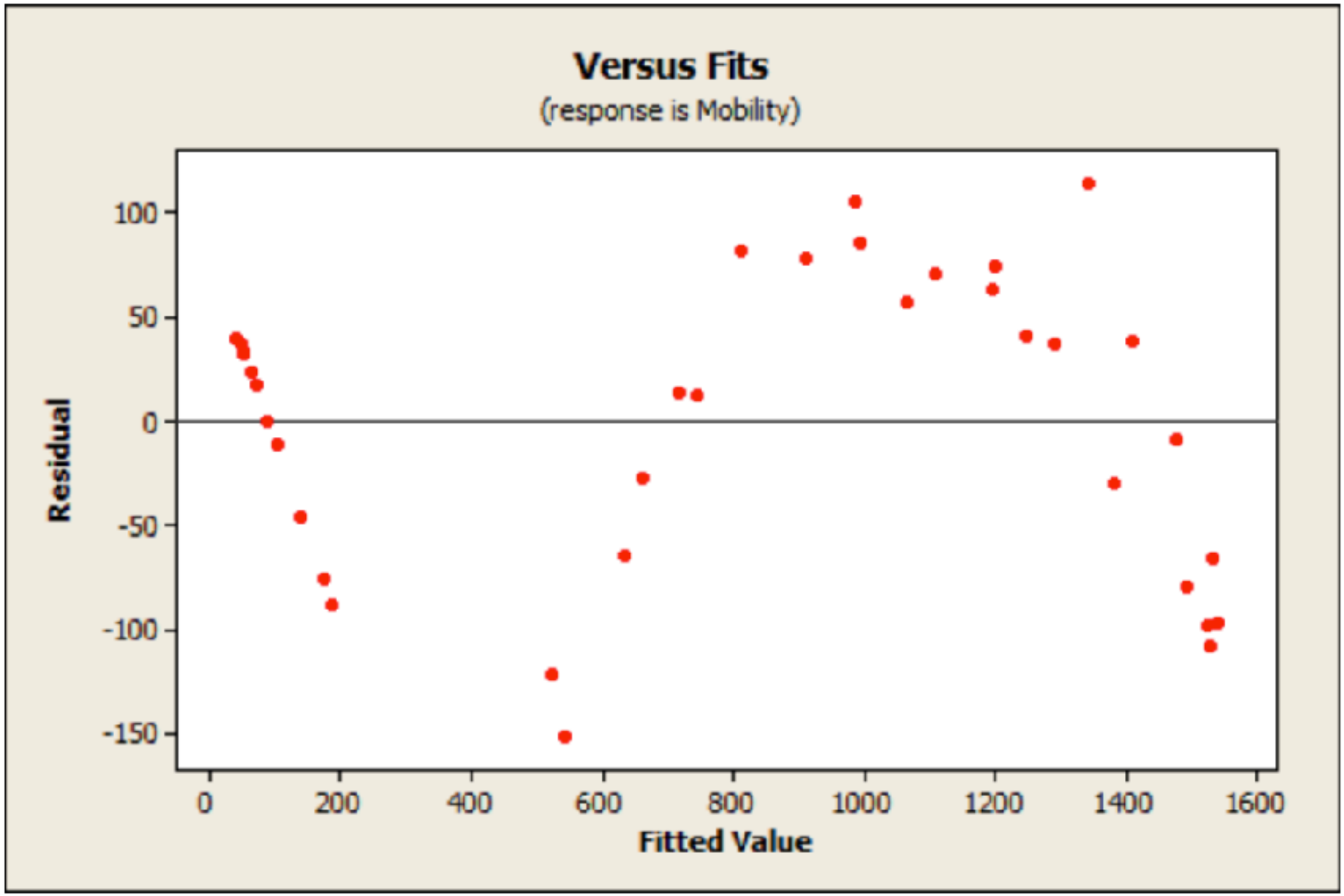
87%

R-Square

$$\sigma_Y = var(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4) + \text{var}(\varepsilon)$$

- 所以R-Square也表示模型解释因变量方差的多少

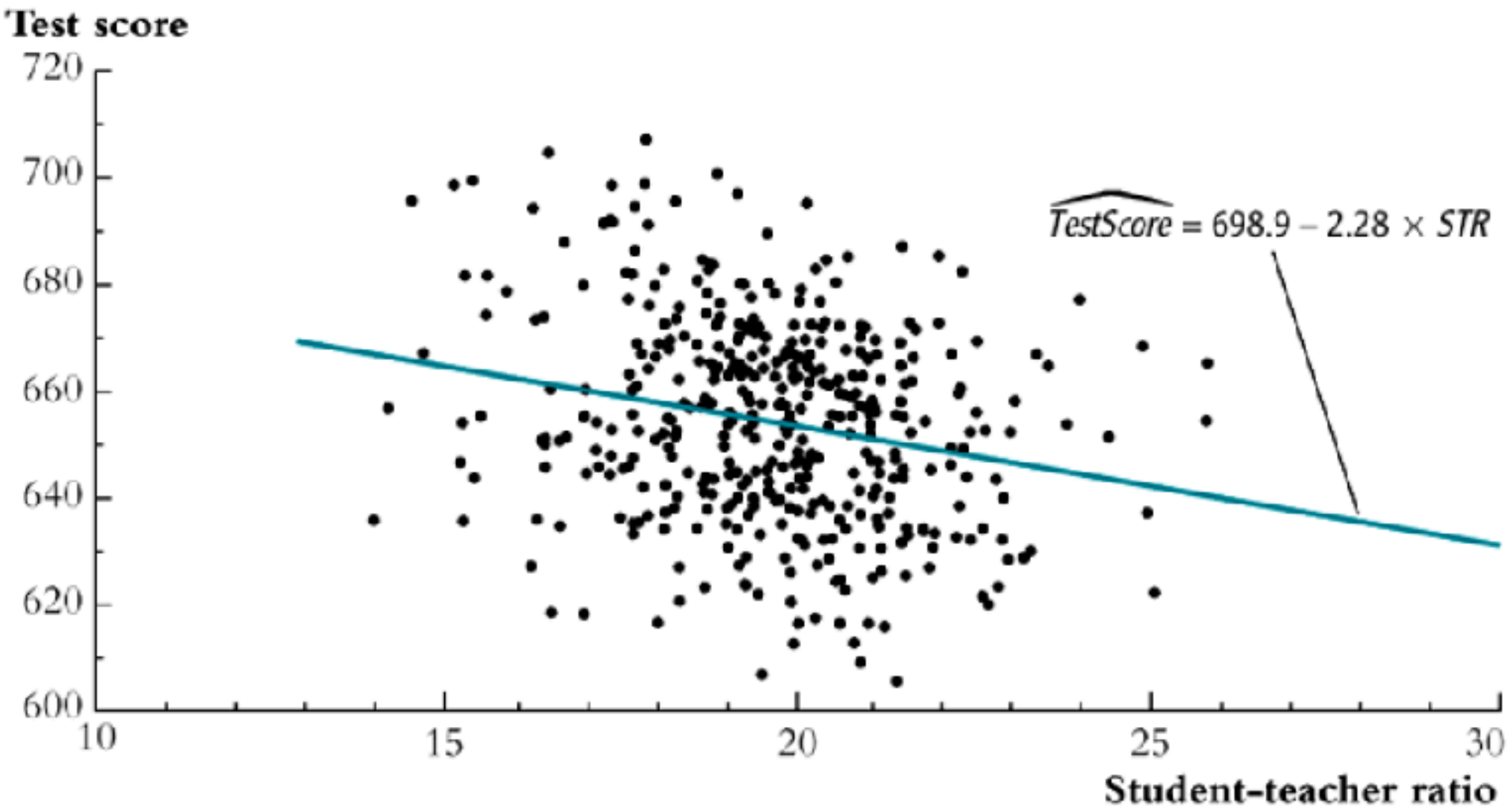
R-Square的高就一定好吗?



判断变量好坏

FIGURE 4.3 The Estimated Regression Line for the California Data

The estimated regression line shows a negative relationship between test scores and the student-teacher ratio. If class sizes fall by 1 student, the estimated regression predicts that test scores will increase by 2.28 points.



假设检验 t-test

- 判断单一变量
- 零假设（Null Hypothesis）：假设样本均值为某值
- 显著水平：一般选择0.05

$$t = \frac{\text{estimator - hypothesized value}}{\text{standard error of the estimator}}$$

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$$

http://www.r-tutor.com/elementary-statistics/hypothesis-testing

单尾检验 VS 双尾检验

Null hypothesis and **two-sided** alternative:

$$H_0: \ \beta_1 = 0 \ \text{ vs. } H_1: \ \beta_1 \neq 0$$

or, more generally,

$$H_0: \ \beta_1 = \beta_{1,0} \ \text{ vs. } H_1: \ \beta_1 \neq \beta_{1,0}$$

where $\beta_{1,0}$ is the hypothesized value under the null.

Null hypothesis and **one-sided** alternative:

$$H_0: \ \beta_1 = \beta_{1,0} \ \text{ vs. } H_1: \ \beta_1 < \beta_{1,0}$$

同时检验多个变量怎么办?

F Test

方差分析

- 方差分析的基本原理是认为不同处理组的均数间的差别基本来源有两个:
- (1) 实验条件，即不同的处理造成的差异，称为组间差异。用变量在各组的均值与总均值之偏差平方和的总和表示，记作SSb，组间自由度dfb。
- (2) 随机误差，如测量误差造成的差异或个体间的差异，称为组内差异，用变量在各组的均值与该组内变量值之偏差平方和的总和表示， 记作SSw，组内自由度dfw。
- 总偏差平方和 SSt = SSb + SSw。
- MSb/MSw比值构成F分布。用F值与其临界值比较，推断各样本是否来自相同的总体。

方差分析示例

- 如某克山病区测得11例克山病患者和13名健康人的血磷值（mmol/L）如下：
  - 患者：0.84 1.05 1.20 1.20 1.39 1.53 1.67 1.80 1.87 2.07 2.11
  - 健康人：0.54 0.64 0.64 0.75 0.76 0.81 1.16 1.20 1.34 1.35 1.48 1.56 1.87

	自由度	离差平方和	均方	F 值	P值
SS组间（处理因素）	1	1.13418185	1.13418185	6.37	0.0193（有统计学意义）
SS组内（抽样误差）	22	3.91761399	0.17807336		
总和	23	5.05179583			

变量显著程度

```
> reg1 = lm(prestige~education+log2(income)+women,data=Prestige)
> summary(reg1)

Call:
lm(formula = prestige ~ education + log2(income) + women, data = Prestige)

Residuals:
    Min      1Q  Median      3Q     Max
-17.364  -4.429  -0.101   4.316  19.179

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -110.9658    14.8429  -7.476 3.27e-11 ***
education      3.7305     0.3544  10.527  < 2e-16 ***
log2(income)   9.3147     1.3265   7.022 2.90e-10 ***
women          0.0469     0.0299   1.568     0.12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.093 on 98 degrees of freedom
Multiple R-squared:  0.8351,    Adjusted R-squared:   0.83
F-statistic: 165.4 on 3 and 98 DF,  p-value: < 2.2e-16
```

变量选择

- RSS：残差平方和
- df：自由度

$$AIC = n\left\{\log\left(\frac{RSS}{n}\right)+1+2\pi\right\}+2\times df$$

$$BIC = n\left\{\log\left(\frac{RSS}{n}\right)+1+2\pi\right\}+\log(n)\times df$$

变量选择

Choose a model by AIC in a Stepwise Algorithm

Description

Performs stepwise model selection by AIC.

Usage

```
stepAIC(object, scope, scale = 0,
        direction = c("both", "backward", "forward"),
        trace = 1, keep = NULL, steps = 1000, use.start = FALSE,
        k = 2, ...)
```

线性回归示例

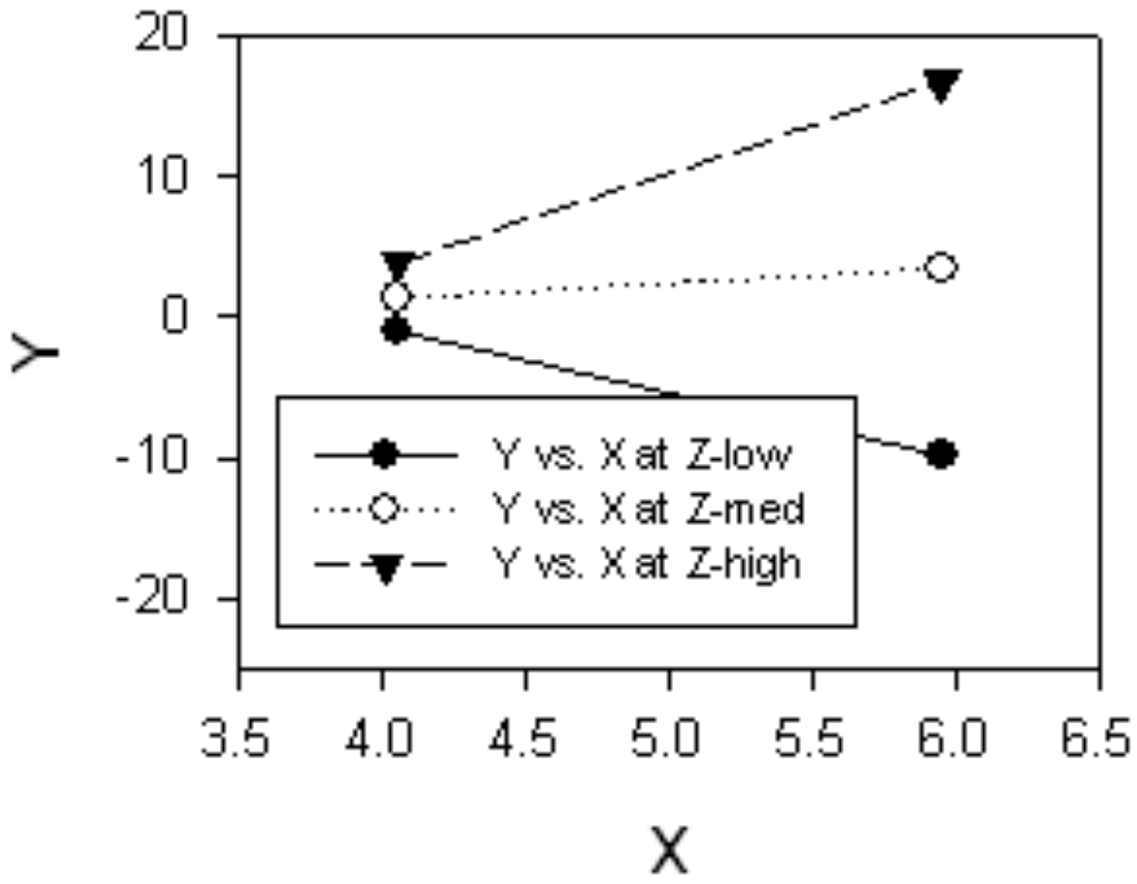
- education
  - Average education of occupational incumbents, years, in 1971.
- income
  - Average income of incumbents, dollars, in 1971.
- women
  - Percentage of incumbents who are women.
- prestige
  - Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s.
- census
  - Canadian Census occupational code.
- type
  - Type of occupation. A factor with levels (note: out of order): bc, Blue Collar; prof, Professional, Managerial, and Technical; wc, White Collar.

线性回归示例

- reg1 = lm(prestige~education+log2(income)+women,data=Prestige)
- prestige_predict = fitted(reg1)
- prestige_resid = residuals(reg1)

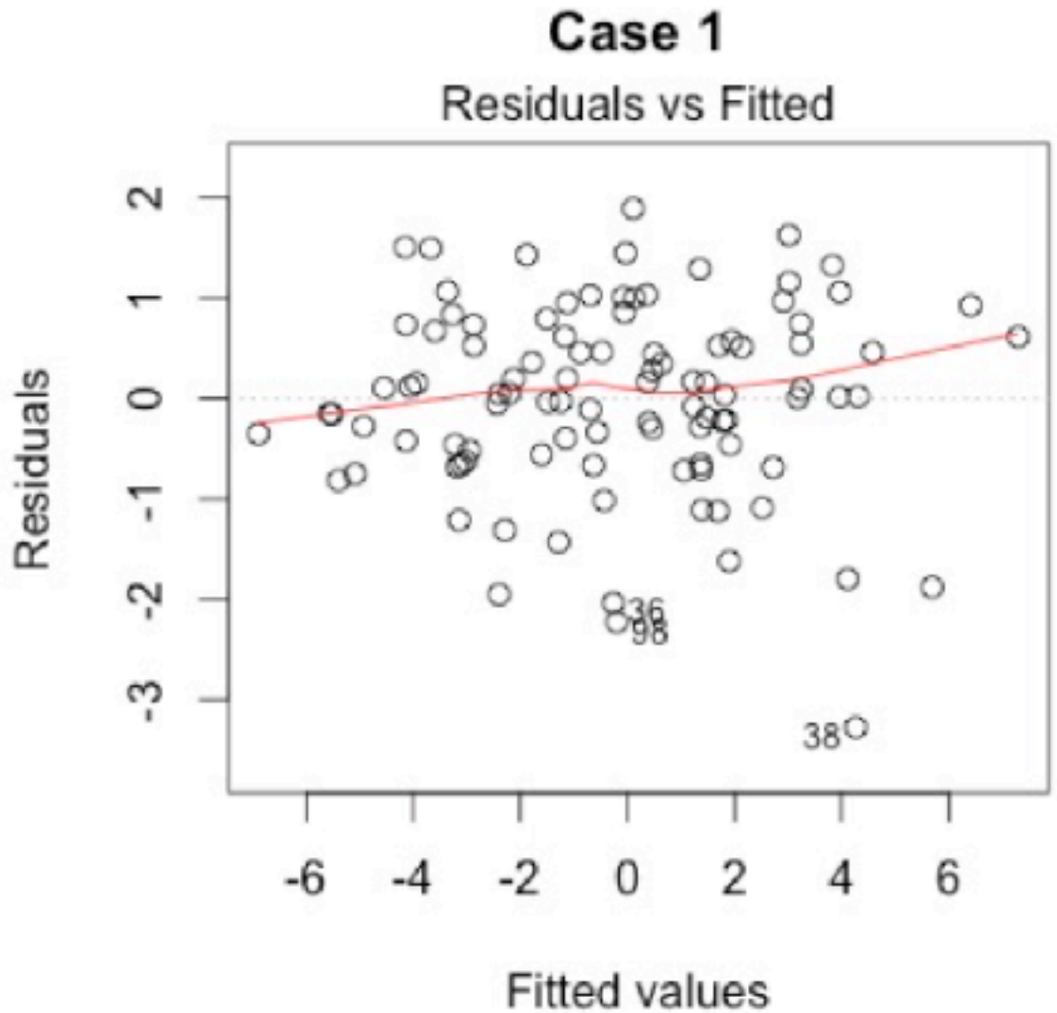
交互项（interaction）

- 模型中的系数在X处于不同水平的值和显著水平不同

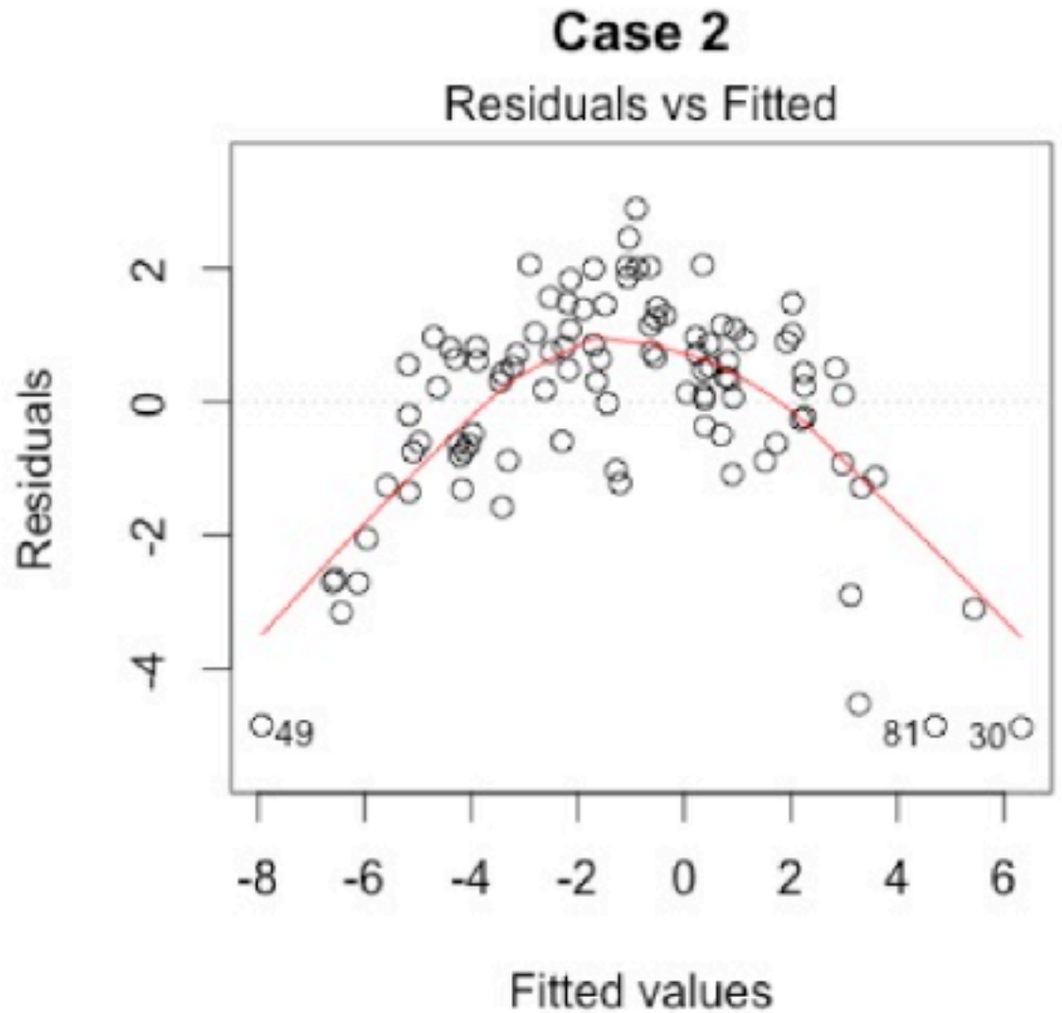


模型诊断

- Residuals vs Fitted：残差是否有非线性模式

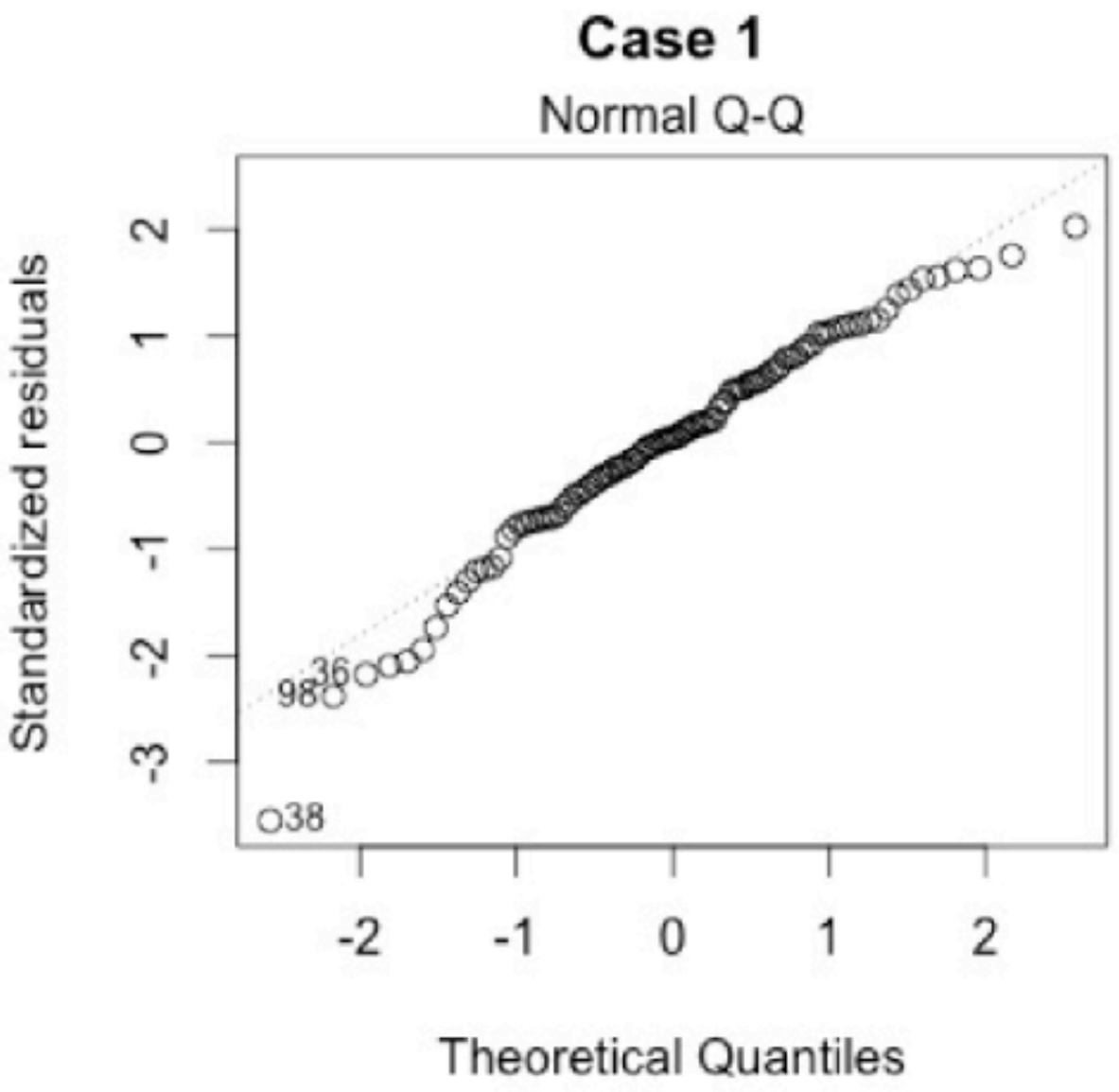


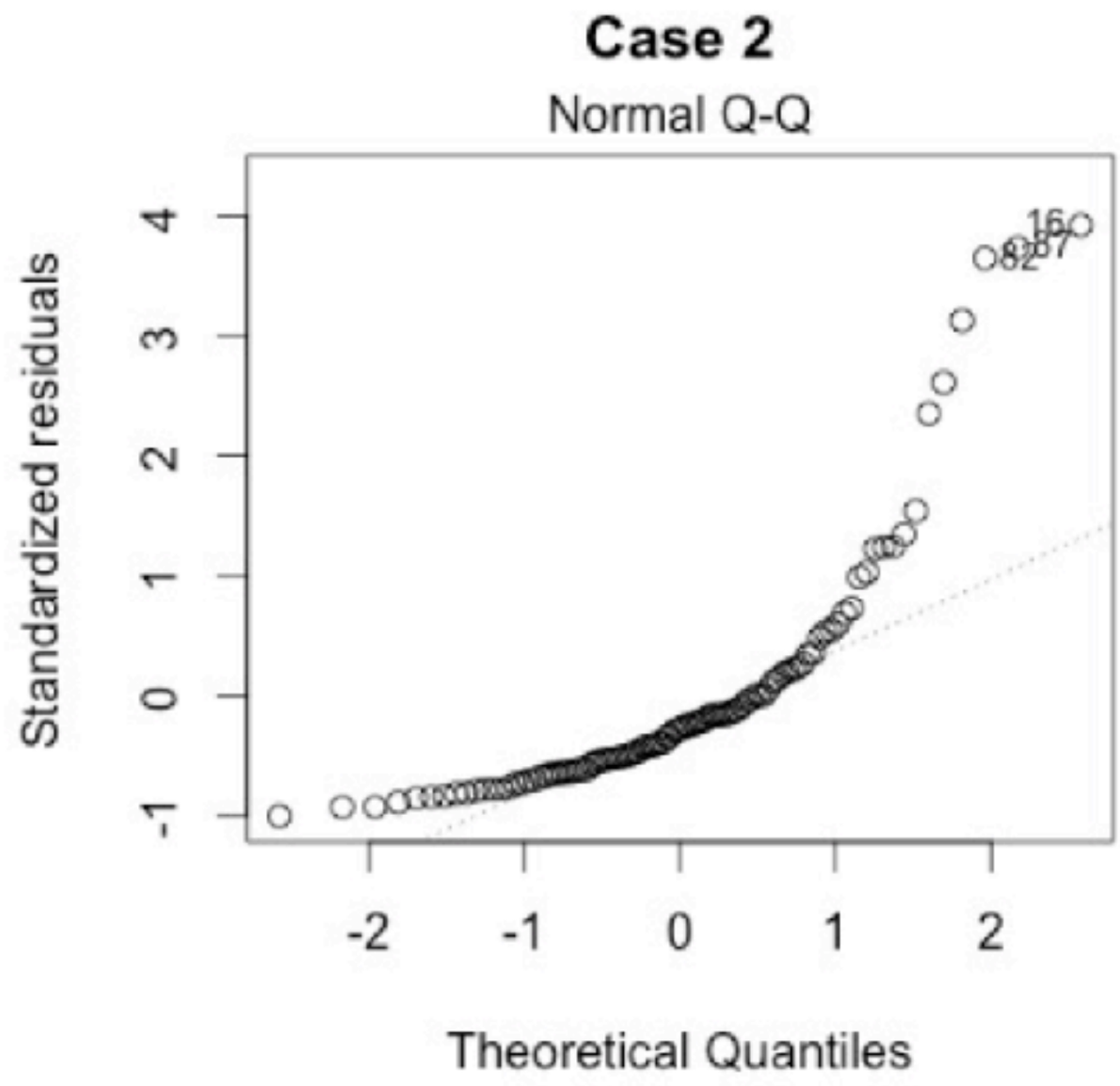
Case 2



模型诊断

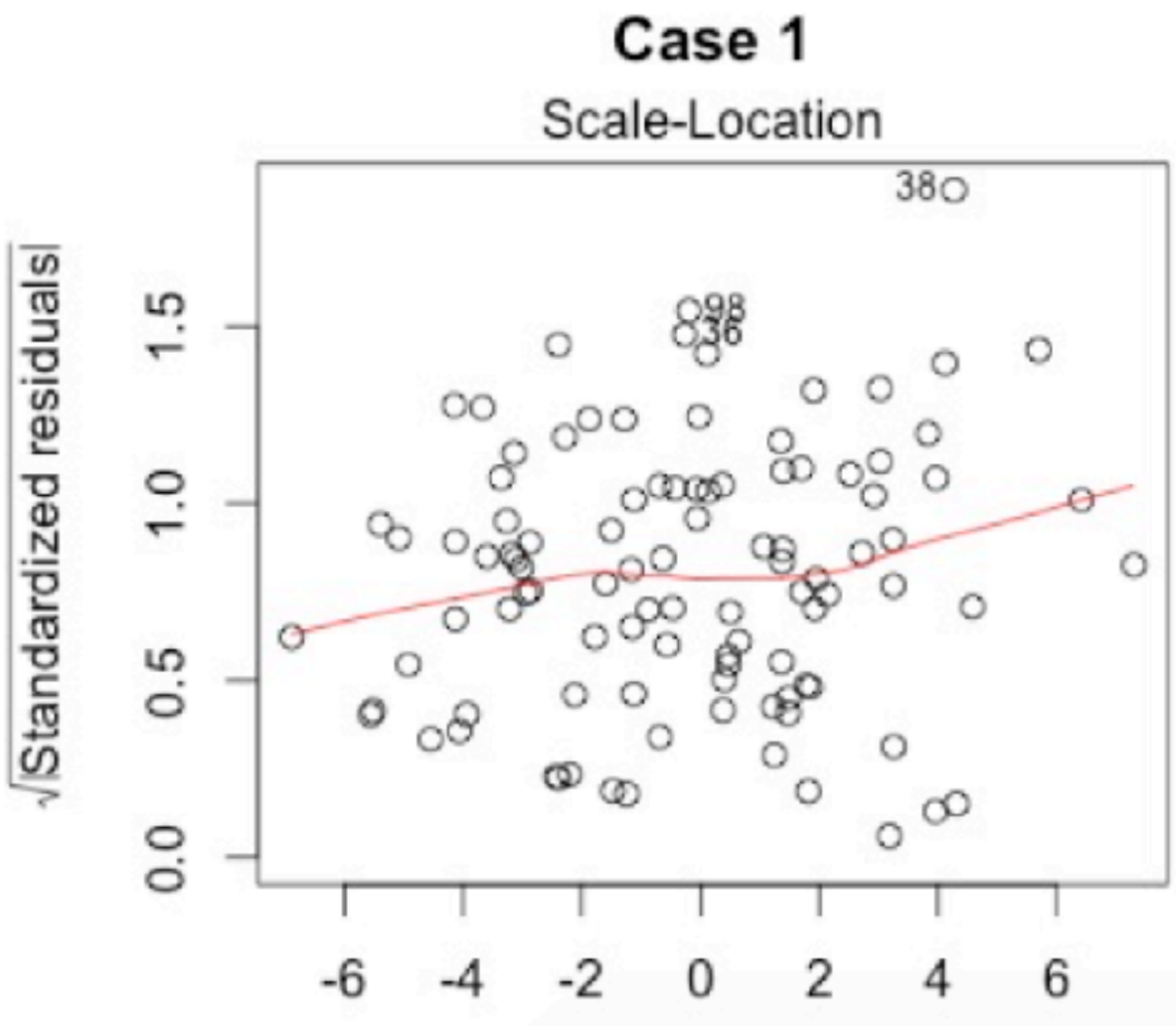
- Normal Q-Q：残差是否正态分布

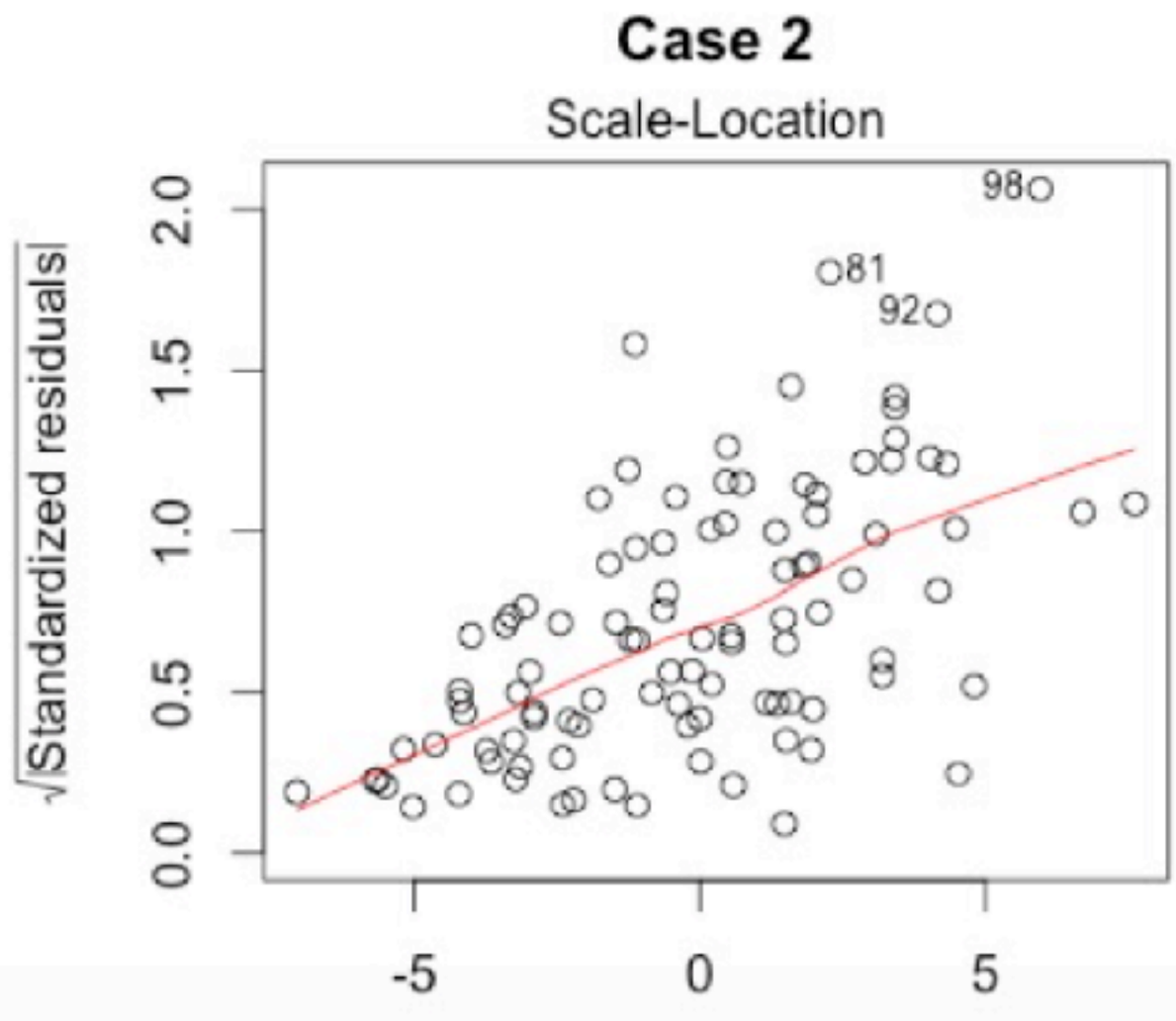




模型诊断

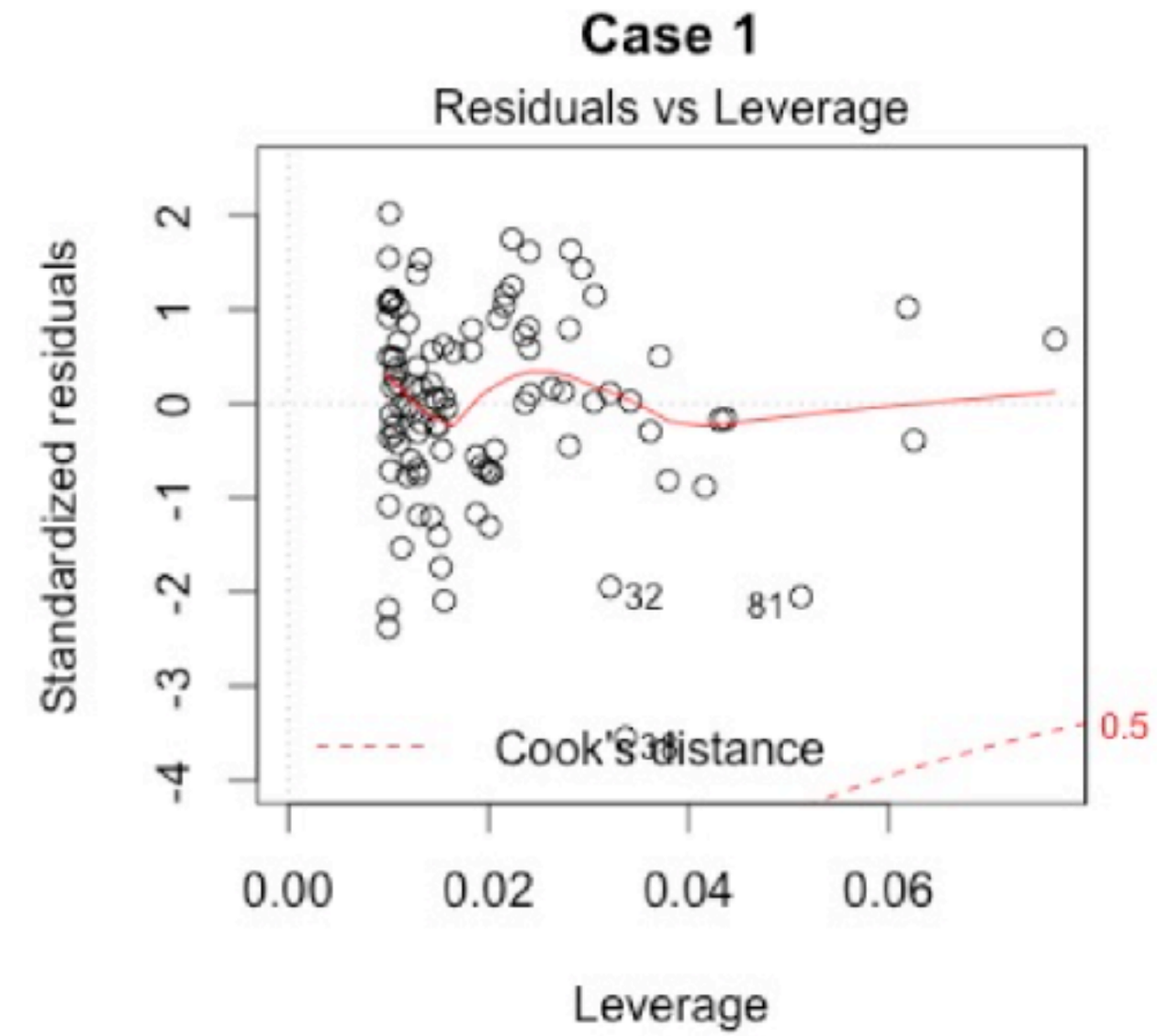
- Scale-Location：残差分布是否平均

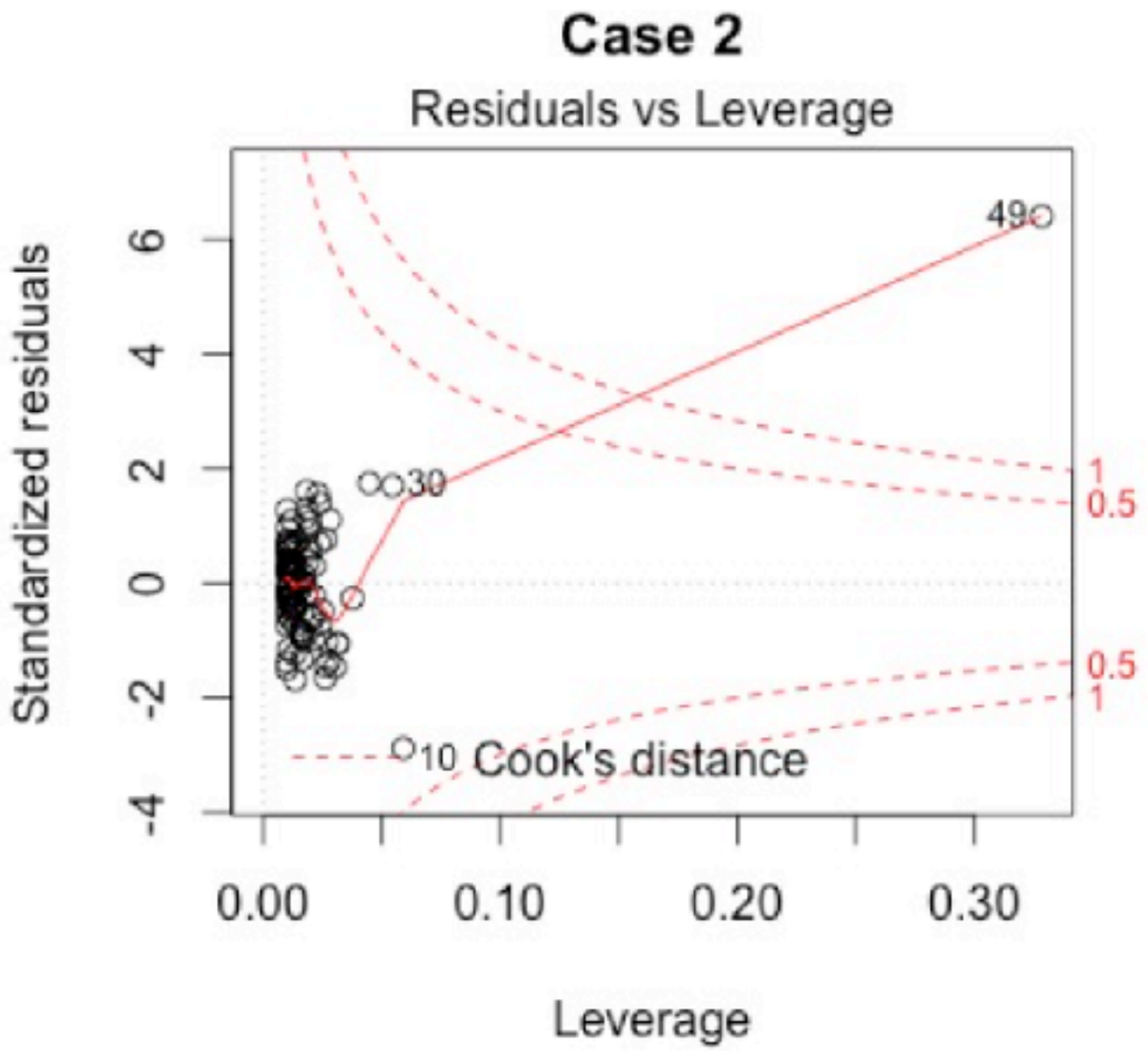




模型诊断

- Residuals vs Leverage：找出影响回归结果的观测





主要内容

- 线性回归及应用
- 逻辑(0-1)回归及应用
- 作业

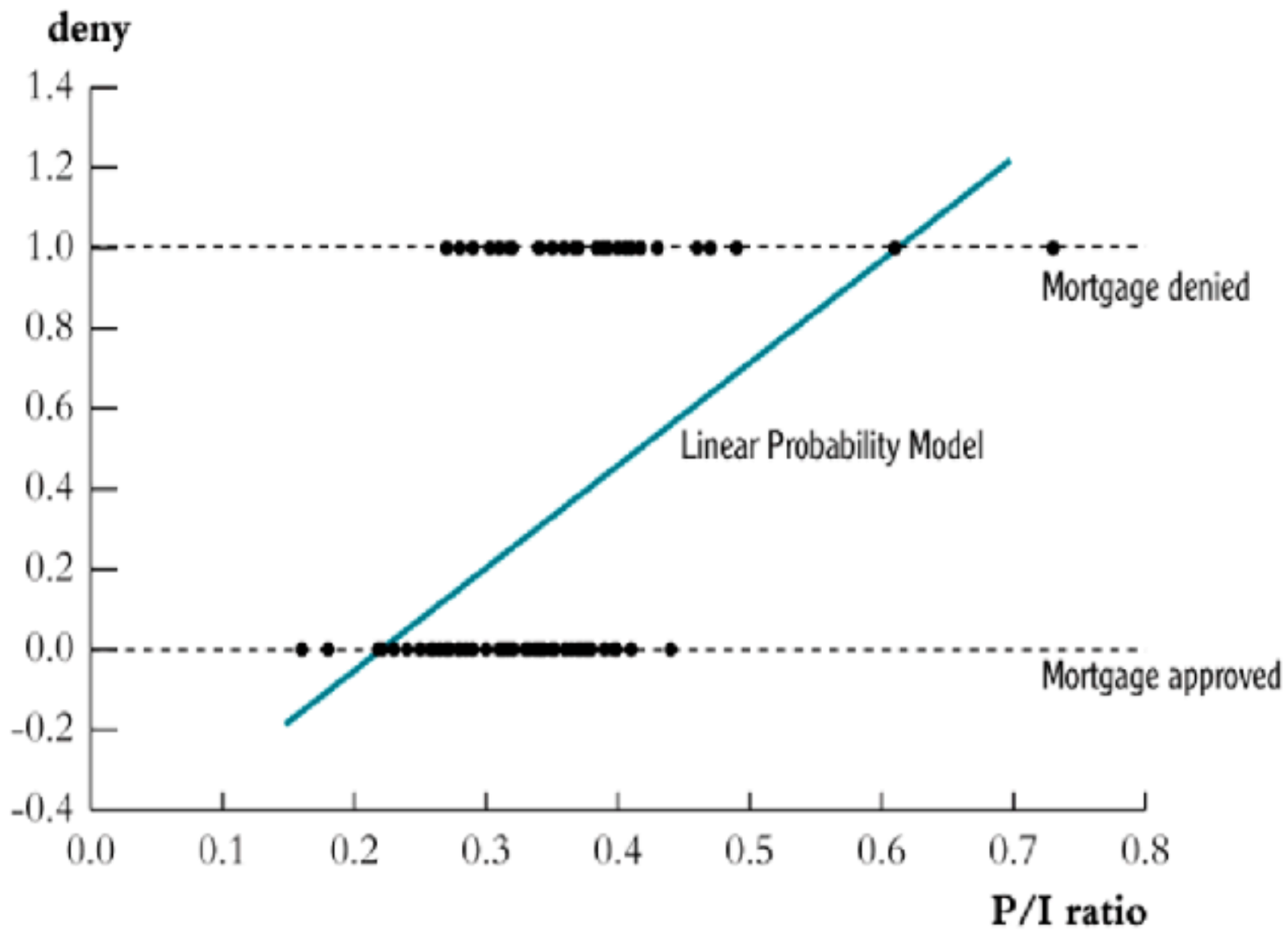
0-1回归

- 特点：因变量是二元变量，取值为两种，通常为0或1
- 本质不是预测问题，而是分类问题
- 比如：
  - 能够考上大学
  - 上课是否会迟到
  - 消费者是否购买

是否可以用线性回归?

FIGURE 9.1 Scatterplot of Mortgage Application Denial and the Payment-to-Income Ratio

Mortgage applicants with a high ratio of debt payments to income (*P/I ratio*) are more likely to have their application denied (*deny* = 1 if denied, *deny* = 0 if approved). The linear probability model uses a straight line to model the probability of denial, conditional on the *P/I ratio*.



$$\widehat{deny} = -.080 + .604 P/I\ ratio \qquad (n = 2380)$$
$$(.032)\ (.098)$$

$$\widehat{\Pr(deny = 1 \mid P/Iratio = .3)} = -.080 + .604 \times .3 = .15$$

$$\widehat{deny} = -.091 + .559 P/I\ ratio + .177 black$$
$$(.032)\ \ (.098) \qquad\qquad (.025)$$

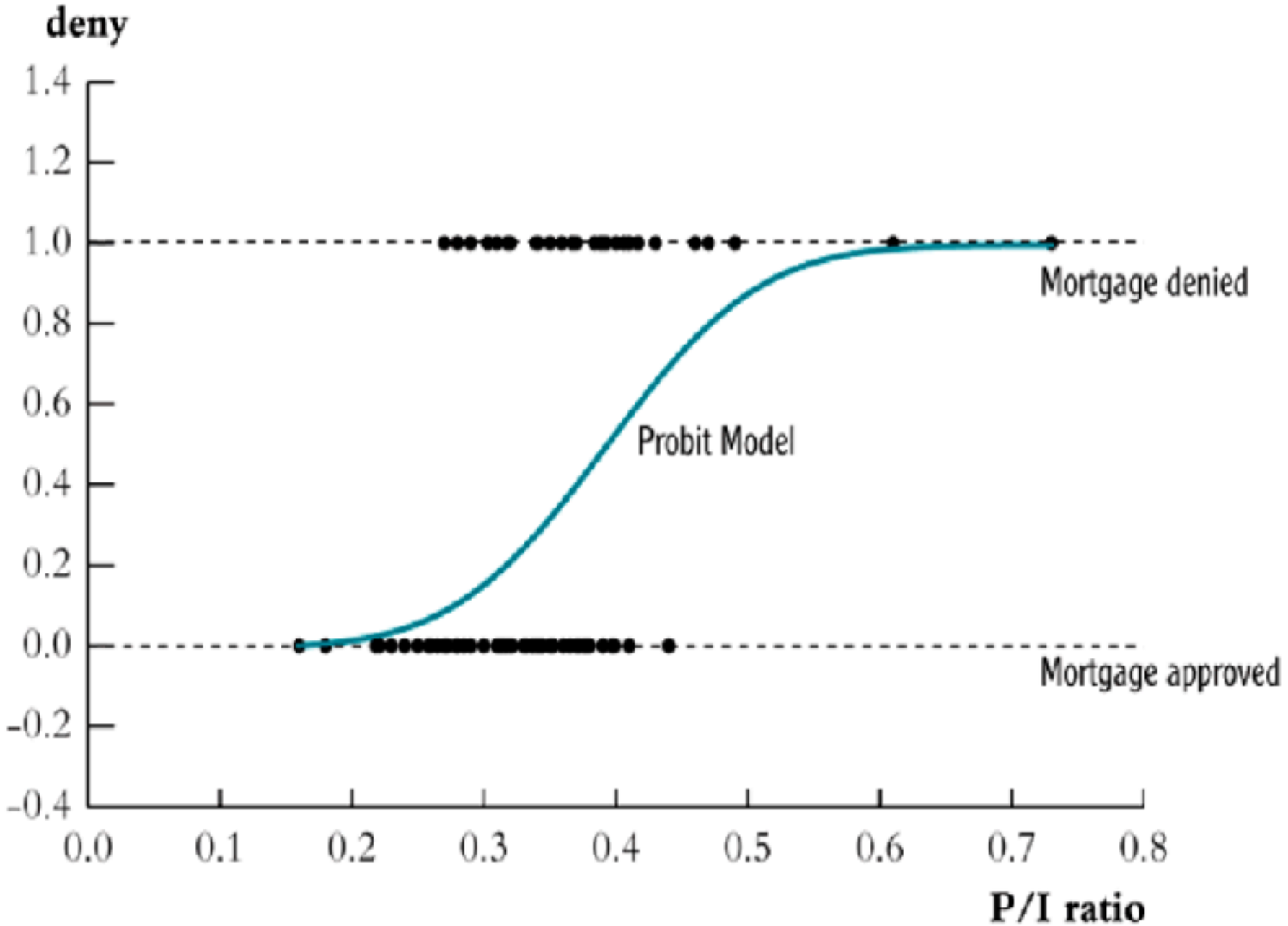
$$\widehat{\Pr(deny = 1)} = -.091 + .559 \times .3 + .177 \times 1 = .254$$

$$\widehat{\Pr(deny = 1)} = -.091 + .559 \times .3 + .177 \times 0 = .077$$

逻辑回归

$$\Pr(Y=1|X)=\beta_0+\beta_1 X$$

The probit model uses the cumulative normal distribution function to model the probability of denial given the payment-to-income ratio or, more generally, to model $\Pr(Y=1|X)$. Unlike the linear probability model, the probit conditional probabilities are always between zero and one.



逻辑回归

$$\Pr(Y = 1|X) = F(\beta_0 + \beta_1 X)$$

$$F(\beta_0 + \beta_1 X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

$\beta_0 = -3, \beta_1 = 2, X = .4,$

so $\beta_0 + \beta_1 X = -3 + 2 \times .4 = -2.2$ so

$\Pr(Y = 1|X = .4) = 1/(1 + e^{-(-2.2)}) = .0998$

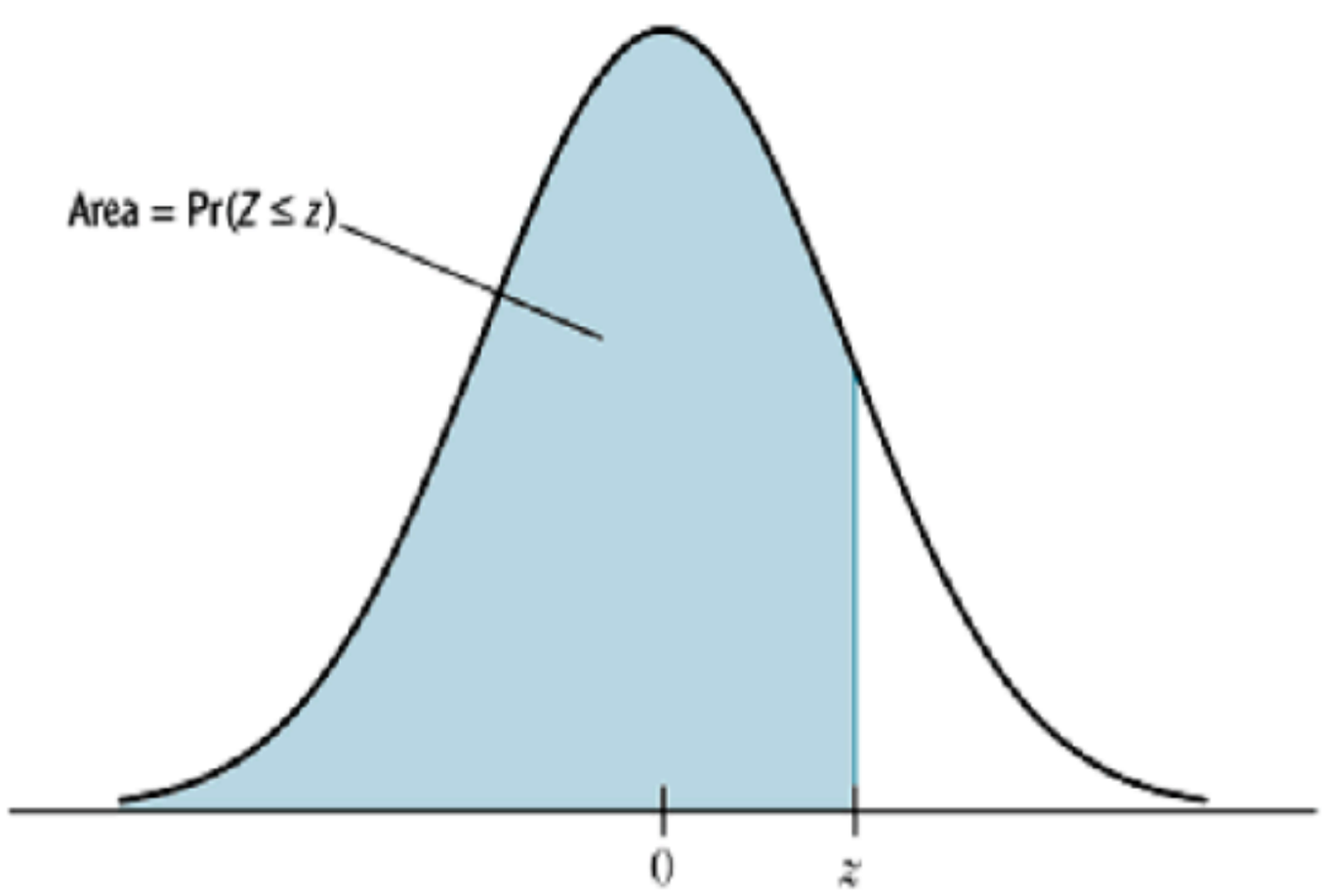
Probit 模型

$$\Pr(Y = 1|X) = \Phi(\beta_0 + \beta_1 X)$$

Φ 表示累积概率密度函数

Example: Suppose $\beta_0 = -2$, $\beta_1 = 3$, $X = .4$, so

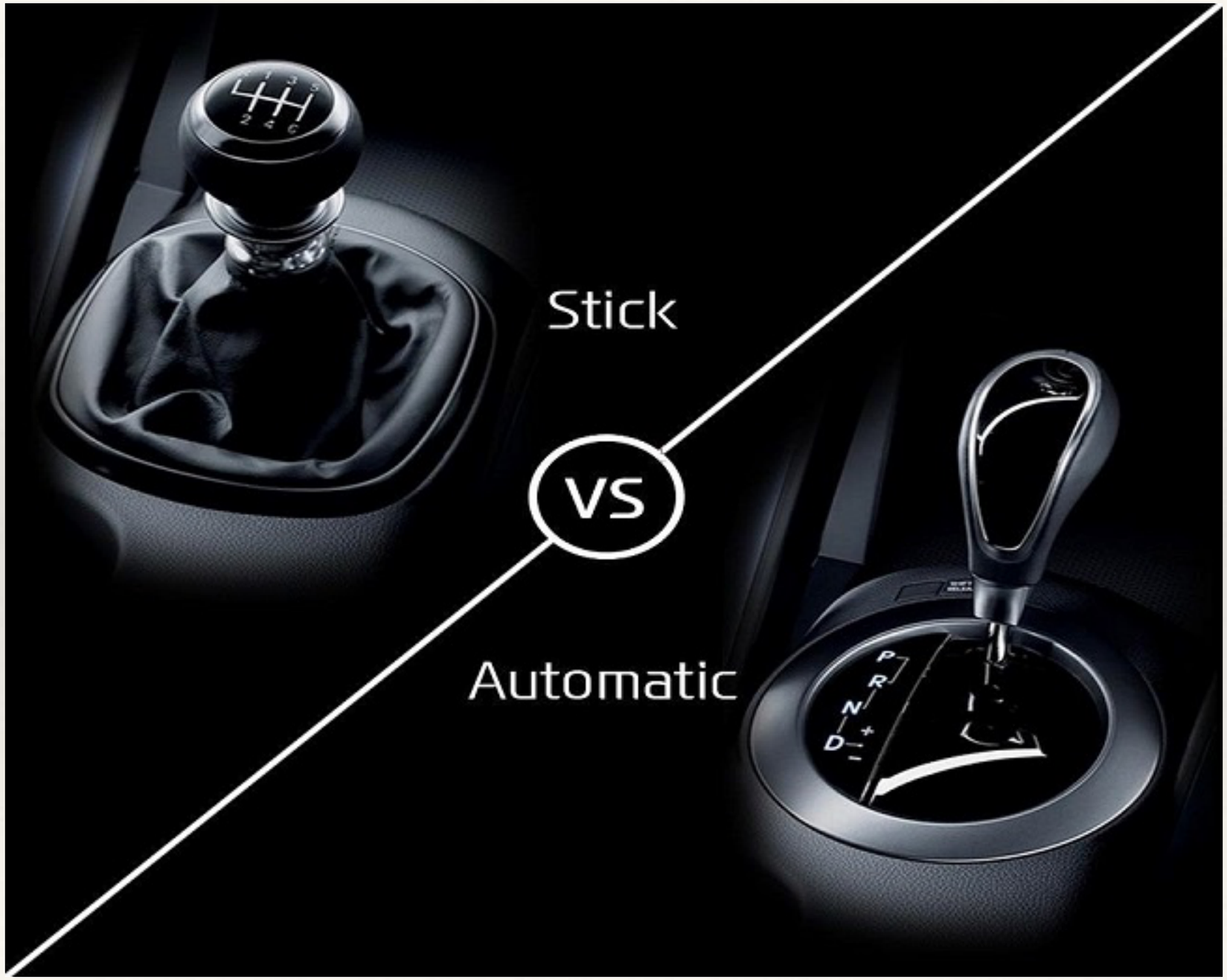
$$\Pr(Y = 1|X=.4) = \Phi(-2 + 3\times.4) = \Phi(-0.8)$$



	Second Decimal Value of z									
z	0	1	2	3	4	5	6	7	8	9
–2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
–2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
–0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
–0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
–0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
–0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
–0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121

$$\Pr(Z \leq -0.8) = .2119$$

0-1回归在R中实现



0-1回归在R中实现

- mpg
  - Miles/(US) gallon
- cyl
  - Number of cylinders
- disp
  - Displacement (cu.in.)
- hp
  - Gross horsepower
- drat
  - Rear axle ratio
- wt
  - Weight (1000 lbs)
- qsec
  - 1/4 mile time
- vs
  - V/S
- am
  - Transmission (0 = automatic, 1 = manual)
- gear
  - Number of forward gears
- carb
  - Number of carburetors

```
> am.glm.logit = glm(formula=am ~ hp + wt,  data=mtcars, family=binomial)
> summary(am.glm.logit)

Call:
glm(formula = am ~ hp + wt, family = binomial, data = mtcars)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2537  -0.1568  -0.0168   0.1543   1.3449

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 18.86630    7.44356   2.535  0.01126 *
hp           0.03626    0.01773   2.044  0.04091 *
wt          -8.08348    3.06868  -2.634  0.00843 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43.230  on 31  degrees of freedom
Residual deviance: 10.059  on 29  degrees of freedom
AIC: 16.059

Number of Fisher Scoring iterations: 8
```

```
> am.glm.probit = glm(formula=am ~ hp + wt,  data=mtcars, family=binomial(link = "probit"))
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(am.glm.probit)

Call:
glm(formula = am ~ hp + wt, family = binomial(link = "probit"),
    data = mtcars)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.13745  -0.10433  -0.00229   0.13473   1.39782

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 10.40552    3.62034   2.874  0.00405 **
hp           0.02126    0.00919   2.313  0.02071 *
wt          -4.54219    1.51205  -3.004  0.00266 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 43.2297  on 31  degrees of freedom
Residual deviance:  9.8605  on 29  degrees of freedom
AIC: 15.861

Number of Fisher Scoring iterations: 9
```

作业提交说明

- 1. 访问github上面课程的文件夹https://github.com/jasonyaopku/Data-Processing-in-R.git，然后进入课程作业的目录Homeworks中下载对应的作业
- 2. 请大家将作业答案保存到word文件中另存为pdf，然后发送到邮箱 yaokai@cufe.edu.cn
- 3. 提交作业的邮件标题和word文件名：DSJJYB_组名_作业1，务必按照这个方式，否则可能会遗漏大家的邮件造成减分。
- 4. 提交报告、数据和代码，报告中标出每个人负责的内容和分工比例