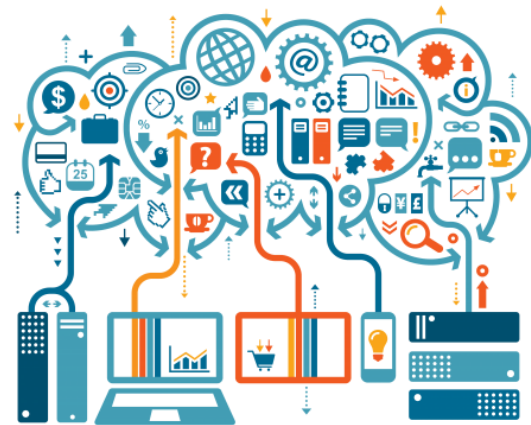


Linear Regression



目录

Contents



01



线性回归模型

02



回归诊断

03



交互效应

04



变量选择

05



回归预测

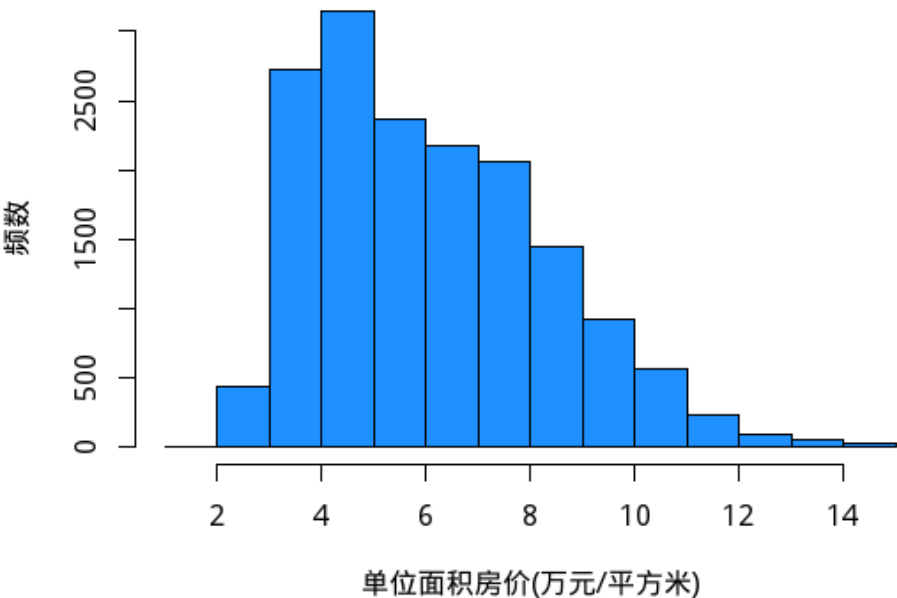


数据说明

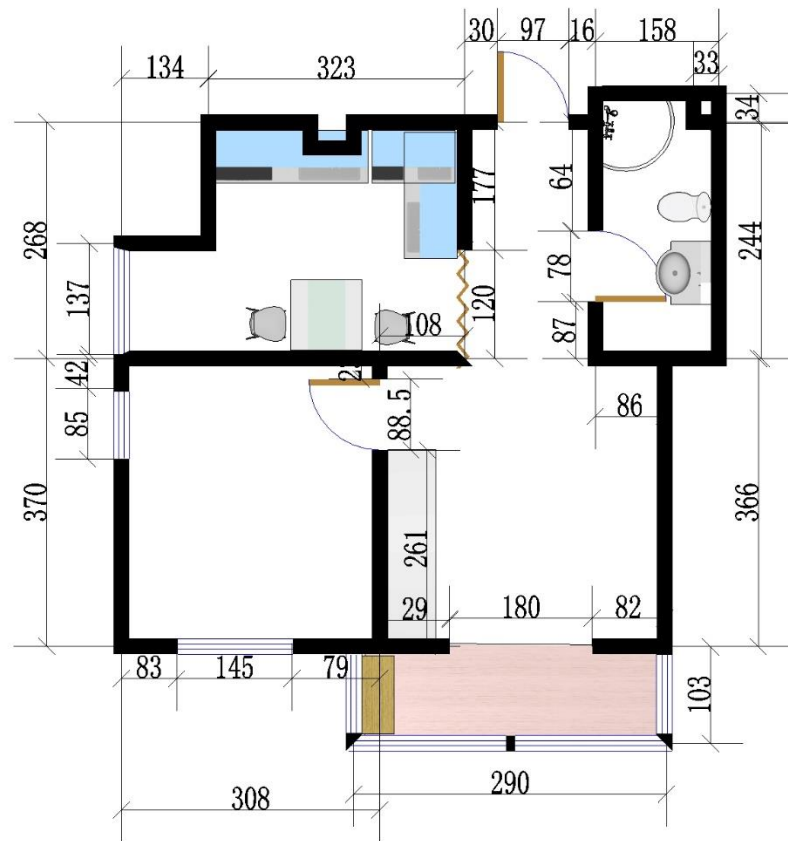
变量类型		变量名	详细说明	取值范围	备注
因变量		单位面积房价	单位：万元/平方米	1.83~14.98	
自变量	内部因素	房屋面积	单位：平方米	30.06~299.00	
		卧室数	单位：个	1~5	
		厅数	单位：个	0~3	
		所属楼层	定性变量 共3个水平	低楼层、中楼层、高楼层	相对楼层
	区位因素	所属城区	定性变量 共6个水平	朝阳区、东城区、丰台区、 海淀区、石景山区、西城区	
		是否邻近地铁	定性变量 共2个水平	1代表邻近地铁 0代表不邻近地铁	82.89% 邻近地铁
		是否学区房	定性变量 共2个水平	1代表学区房 0代表非学区房	30.22% 是学区房



北京二手房单位房价 (N=16210)



- **均 值: 6.12万元/平方米**
- **中位数: 5.74万元/平方米**
- **最小值: 1.40万元/平方米**
 - 丰台区东山坡三里的一间两居室
 - 总面积100.83平米
- **最大值: 14.99万元/平方米**
 - 西城区金融街的一套三室一厅
 - 总面积77.40平米





探秘北京金融街天价学区房：一平米要40万(图)

2016年01月14日 13:19

来源：环球时报

倪浩】

[打印本稿] [字号 大 中 小] [手机看新闻]





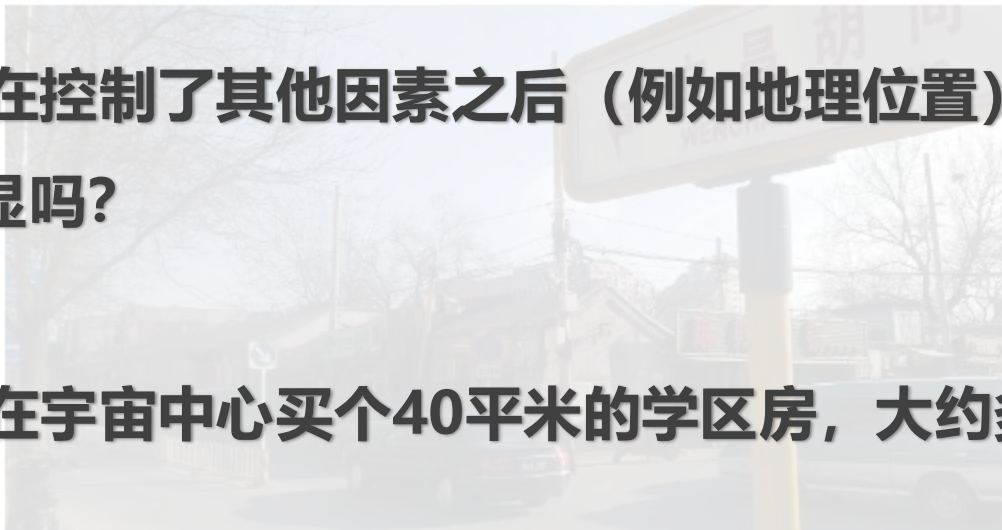
探秘北京金融街天价学区房：一平米要40万(图)

1. 单位房价跟面积、学区房相关吗？有多相关？

[打印本稿] [字号 大 中 小] [手机看新闻]

2. 在控制了其他因素之后（例如地理位置），这种相关性还明显吗？

3. 在宇宙中心买个40平米的学区房，大约多少钱？





简单线性回归



简单线性回归

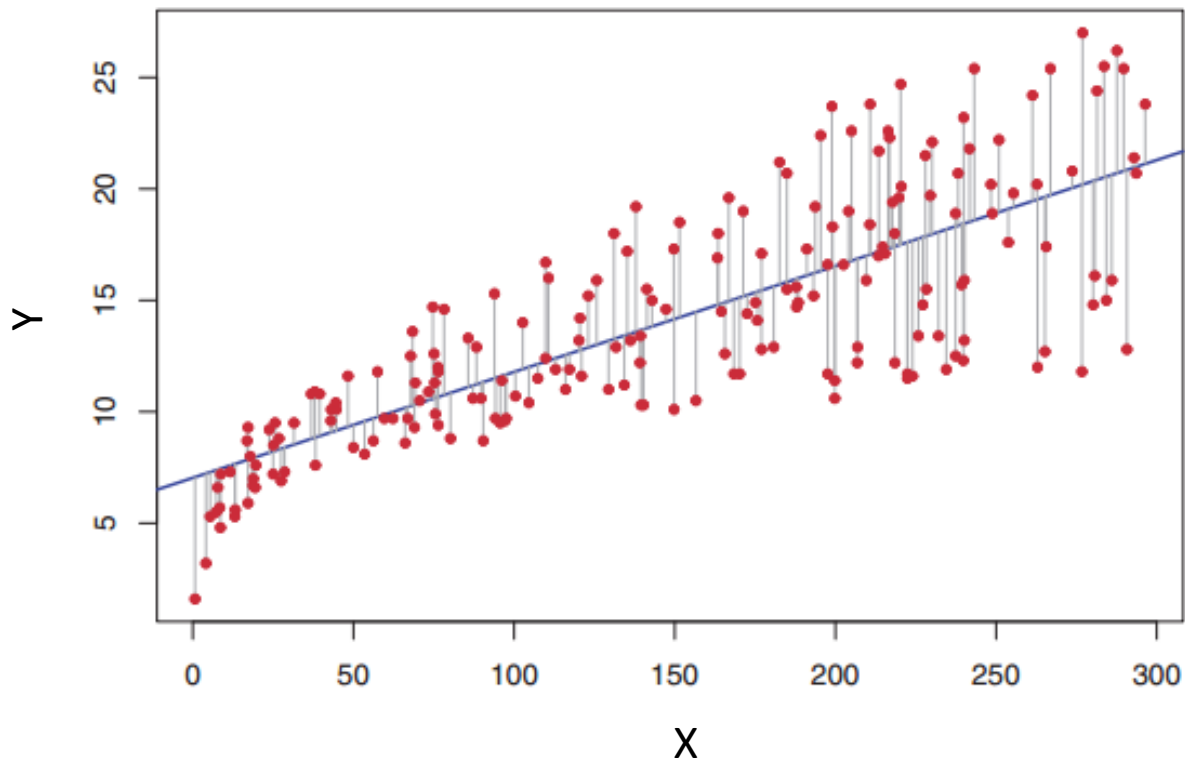
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Diagram illustrating the components of the simple linear regression equation:

- β_0 is labeled as 截距项 (Intercept term) with an upward red arrow.
- X is labeled as 房屋面积 (House area) with a downward red arrow.
- ε is labeled as 误差项 (Error term) with an upward red arrow.



参数估计







最小二乘估计

residual sum of squares (RSS):

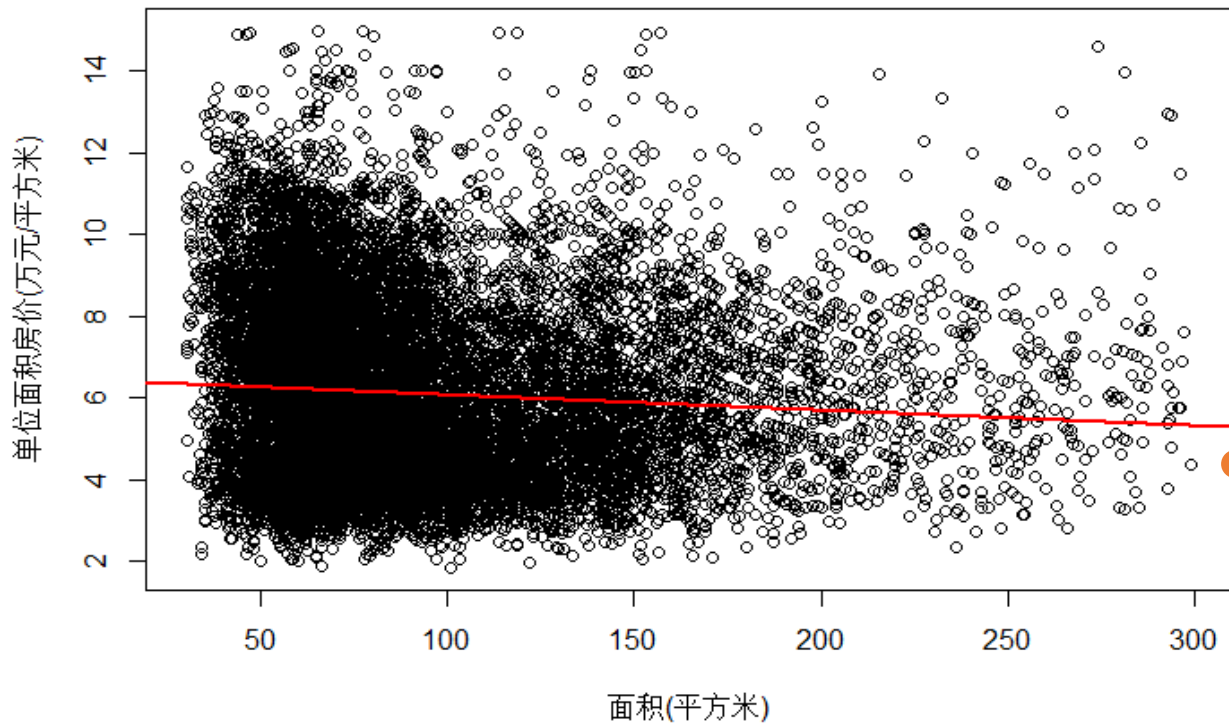
$$\text{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \dots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

min  

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$



单位房价 v.s. 面积



你能看出哪些
信息？

$$\hat{\beta}_0 = 6.459,$$

$$\hat{\beta}_1 = -0.004$$



标准误

- 估计均值 $\hat{\mu} = \frac{\sum_i y_i}{n}$
- 标准误 (Standard Error of $\hat{\mu}$) :

$$Var(\hat{\mu}) = SE(\hat{\mu})^2 = \frac{\sigma^2}{n}$$

→ 与 σ^2 有关
→ 与 n 有关

- 回归系数的标准误 :

$$SE(\widehat{\beta}_0)^2 = \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right), SE(\widehat{\beta}_1)^2 = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}$$



置信区间

- n 足够大时： $\frac{\widehat{\beta}_j - \beta_j}{SE(\widehat{\beta}_j)} \sim N(0,1)$
- 置信水平为 $1 - \alpha$ 的置信区间：

$$\widehat{\beta}_j \pm z_{\alpha/2} SE(\widehat{\beta}_j)$$

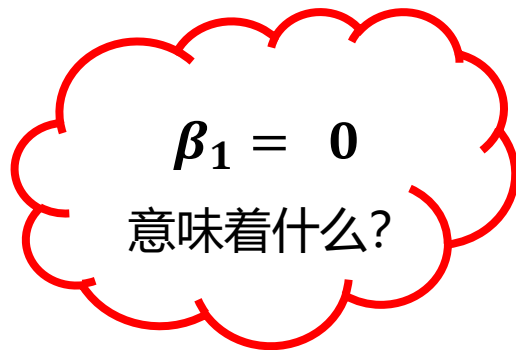
- $P(\widehat{\theta}_L \leq \theta \leq \widehat{\theta}_U) = 1 - \alpha$



回归系数的假设检验

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$





回归系数的假设检验

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

$$\text{检验统计量: } t = \frac{\widehat{\beta}_1 - 0}{SE(\widehat{\beta}_1)}$$

- 在零假设下， t 服从自由度是 $(n-2)$ 的 t 分布。在显著性水平 α 的前提下，如果 $|t| > t_{\alpha/2}(n-2)$ 则拒绝 H_0

$$p\text{-value} = P(t_{n-2} > |t|)$$

- p 值是原假设可被拒绝的最小显著性水平。如果 $p\text{-value} \leq \alpha$ 则在显著性水平 α 下拒绝 H_0



评估回归模型准确性

Residual Standard Error (RSE):

$$RSE = \sqrt{\frac{1}{n-2}RSS} = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \longrightarrow \text{平均偏离程度}$$

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \in [0,1] \longrightarrow \text{回归模型能够解释}\blacktriangledown\text{波动的比例}$$

其中: $TSS = \sum (y_i - \bar{y})^2$, $RSS = \sum (y_i - \hat{y}_i)^2$



单位房价 v.s. 面积: R语言实现

```
> summary(lm(price ~ AREA, data = dat0))

call:
lm(formula = price ~ AREA, data = dat0)

Residuals:
    Min       1Q   Median       3Q      Max
-4.3251 -1.8348 -0.3338  1.5153  9.1663

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.4589537   0.0403832  159.942  <2e-16 ***
AREA        -0.0037470   0.0003969   -9.441  <2e-16 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.223 on 16208 degrees of freedom
Multiple R-squared:  0.005469, Adjusted R-squared:  0.005408
F-statistic: 89.13 on 1 and 16208 DF, p-value: < 2.2e-16

> |
```

β_0 β_1 参数估计



单位房价 v.s. 面积: R语言实现

```
> summary(lm(price ~ AREA, data = dat0))

call:
lm(formula = price ~ AREA, data = dat0)

Residuals:
    Min       1Q   Median       3Q      Max
-4.3251 -1.8348 -0.3338  1.5153  9.1663

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.4589537  0.0403832  159.942  <2e-16 ***
AREA        -0.0037470  0.0003969   -9.441  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.223 on 16208 degrees of freedom
Multiple R-squared:  0.005469, Adjusted R-squared:  0.005408
F-statistic: 89.13 on 1 and 16208 DF, p-value: < 2.2e-16

> |
```

β_0 β_1 参数估计

t统计量



单位房价 v.s. 面积: R语言实现

```
> summary(lm(price ~ AREA, data = dat0))

call:
lm(formula = price ~ AREA, data = dat0)

Residuals:
    Min       1Q   Median       3Q      Max
-4.3251 -1.8348 -0.3338  1.5153  9.1663

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.4589537   0.0403832  159.942  <2e-16 ***
AREA        -0.0037470   0.0003969   -9.441  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.223 on 16208 degrees of freedom
Multiple R-squared:  0.005469, Adjusted R-squared:  0.005408
F-statistic: 89.13 on 1 and 16208 DF, p-value: < 2.2e-16

> |
```

$\beta_0 \beta_1$ 参数估计

t 统计量

p -value



单位房价 v.s. 面积: R语言实现

```
> summary(lm(price ~ AREA, data = dat0))

call:
lm(formula = price ~ AREA, data = dat0)

Residuals:
    Min       1Q   Median       3Q      Max
-4.3251 -1.8348 -0.3338  1.5153  9.1663

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.4589537   0.0403832  159.942  <2e-16 ***
AREA        -0.0037470   0.0003969   -9.441  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.223 on 16208 degrees of freedom
Multiple R-squared:  0.005469,    Adjusted R-squared:  0.005408
F-statistic: 89.13 on 1 and 16208 DF,  p-value: < 2.2e-16

> |
```

$\beta_0 \beta_1$ 参数估计

t统计量

p-value

R^2



自变量离散：0-1自变量

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$X = 1$ 表示是学区房;

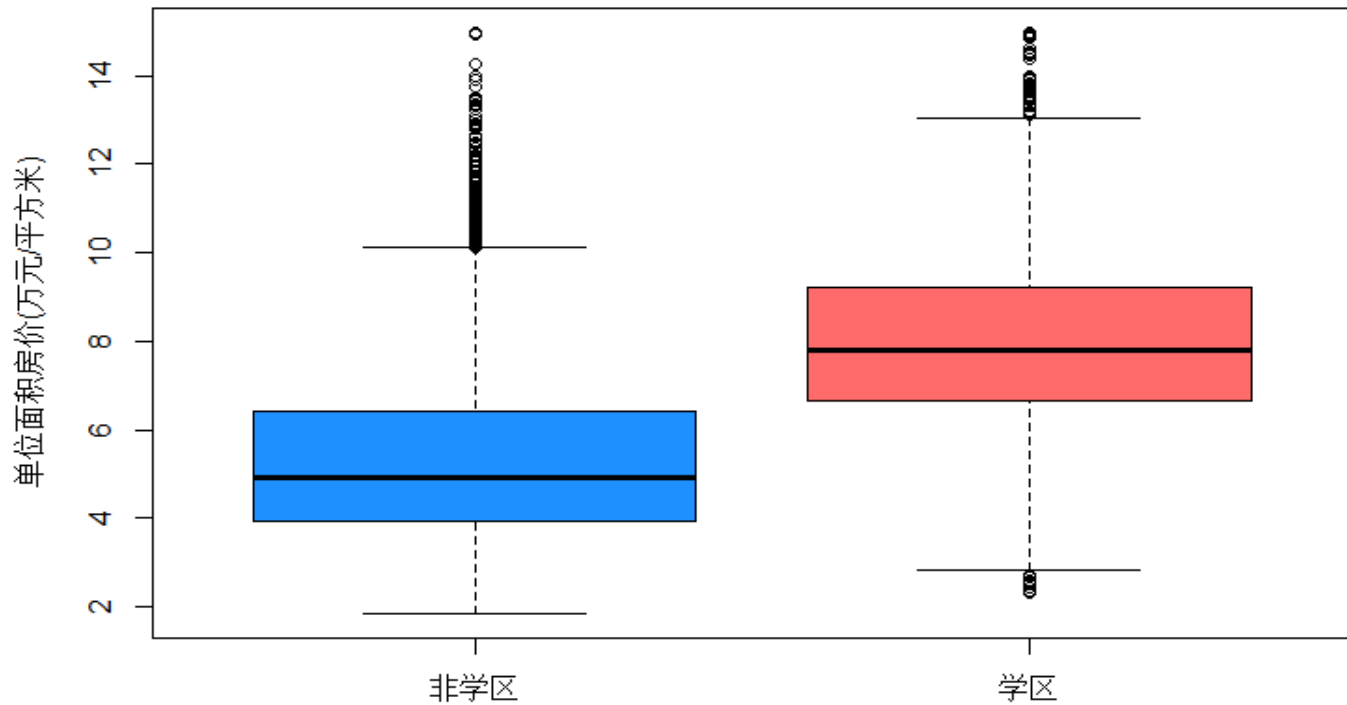
$X = 0$ 表示非学区房



β_1 如何解读?



自变量离散：0-1自变量





自变量离散：0-1自变量

```
> summary(lm(price~school, data=dat0))

Call:
lm(formula = price ~ school, data = dat0)

Residuals:
    Min       1Q   Median       3Q      Max
-5.5957 -1.3670 -0.3366  1.1542  9.6545

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.33262    0.01776   300.32  <2e-16 ***
school         2.58199    0.03225    80.05  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.887 on 16208 degrees of freedom
Multiple R-squared:  0.2834,    Adjusted R-squared:  0.2833
F-statistic: 6408 on 1 and 16208 DF,  p-value: < 2.2e-16
```




自变量离散：多水平（例如：多城区）

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

X = 海淀区、朝阳区、东城区



自变量离散：多水平（例如：多城区）

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

解决方案：

$X_1 = 1$ (海淀区) ; $= 0$ (非海淀区)

$X_2 = 1$ (朝阳区) ; $= 0$ (非朝阳区)



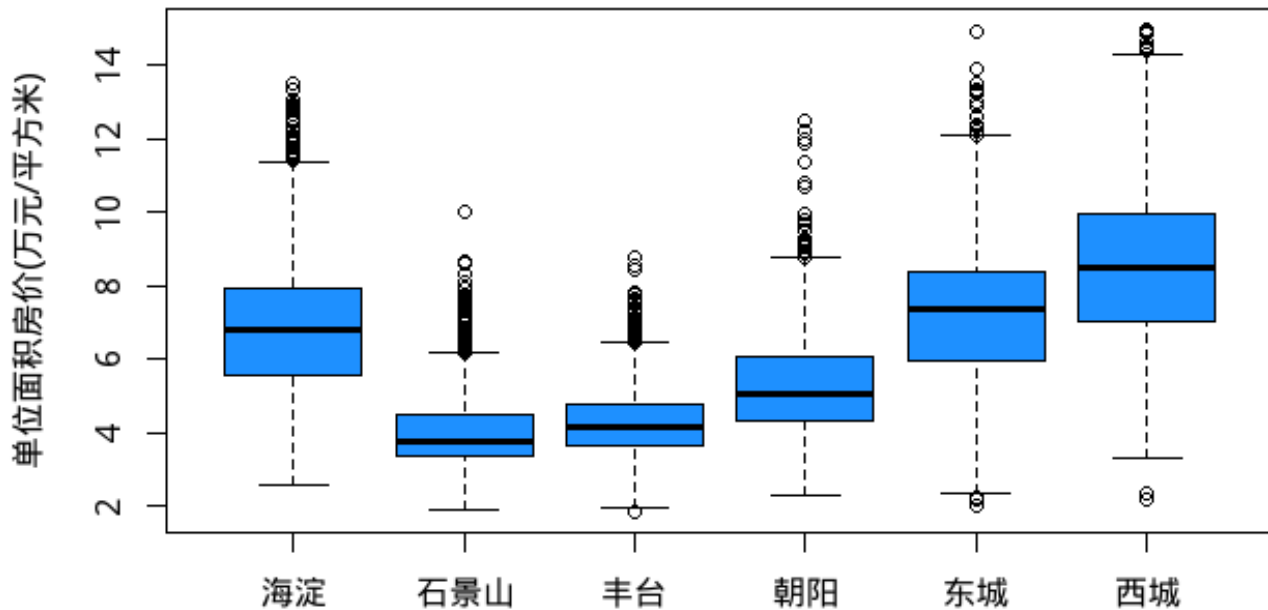
问题：

(1) 是否需要 X_3 ?

(2) 如果有 p 个水平，需要几个0-1变量?



自变量离散：多水平（例如：多城区）





自变量离散：多水平（例如：多城区）

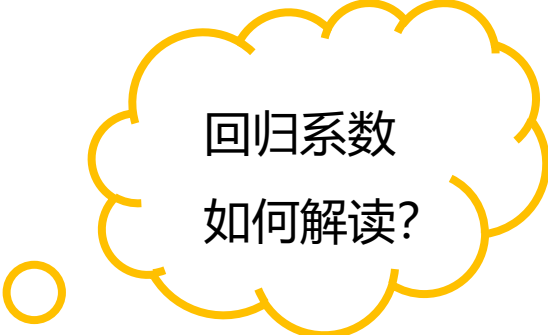
```
> summary(lm(price~CATE, data=dat0))

Call:
lm(formula = price ~ CATE, data = dat0)

Residuals:
    Min       1Q   Median       3Q      Max
-6.3757 -0.9465 -0.1071  0.8932  7.7370

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.02869    0.03489  115.474 < 2e-16 ***
CATE丰台      0.22140    0.04496   4.924 8.54e-07 ***
CATE朝阳      1.25137    0.04522  27.674 < 2e-16 ***
CATE东城      3.15967    0.04548  69.469 < 2e-16 ***
CATE海淀      2.84707    0.04505  63.205 < 2e-16 ***
CATE西城      4.53879    0.04560  99.545 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.539 on 16204 degrees of freedom
Multiple R-squared:  0.5233,    Adjusted R-squared:  0.5232
F-statistic: 3558 on 5 and 16204 DF,  p-value: < 2.2e-16
```



回归系数
如何解读？



自变量离散：多水平（例如：多城区）

```
> dat0$CATE = factor(dat0$CATE,  
+                     levels=c("海淀","石景山","丰台","朝阳","东城","西城"))  
> summary(lm(price~CATE, data=dat0))
```

```
Call:  
lm(formula = price ~ CATE, data = dat0)
```

```
Residuals:  
    Min       1Q   Median       3Q      Max  
-6.3757 -0.9465 -0.1071  0.8932  7.7370
```

```
Coefficients:  
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)   6.87576    0.02849  241.311 < 2e-16 ***  
CATE石景山   -2.84707    0.04505  -63.205 < 2e-16 ***  
CATE丰台     -2.62567    0.04020  -65.315 < 2e-16 ***  
CATE朝阳     -1.59570    0.04049  -39.411 < 2e-16 ***  
CATE东城      0.31260    0.04079   7.665  1.9e-14 ***  
CATE西城      1.69172    0.04091  41.352 < 2e-16 ***  
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.539 on 16204 degrees of freedom  
Multiple R-squared:  0.5233,    Adjusted R-squared:  0.5232  
F-statistic: 3558 on 5 and 16204 DF, p-value: < 2.2e-16
```

将海淀区换
为基础？



多元线性回归

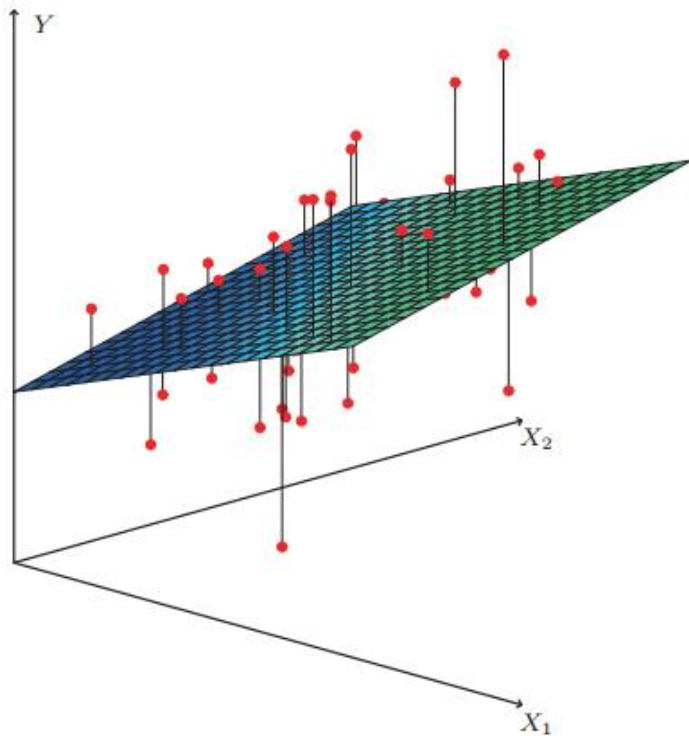


多元线性回归

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$



多元线性回归



$$\begin{aligned}\text{RSS} &= \sum_{i=1}^n (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2.\end{aligned}$$

$$\hat{\beta} = (X'X)^{-1}(XY)$$



评估：是否存在线性关系？

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_1: \text{至少有一个 } \beta_j \neq 0$$

在零假设下，F统计量：
$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n - p - 1)},$$

服从自由度是 $(p, n - p - 1)$ 的 F 分布。



扩展：是否存在线性关系？

$$H_0: \beta_{p-q+1} = \beta_{p-q+2} = \cdots = \beta_p = 0$$

在零假设下，F统计量

$$F = \frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n - p - 1)}.$$

如果 $q=1$ 呢？

服从自由度是 $(q, n - p - 1)$ 的 F 分布。



多元回归中的 R^2

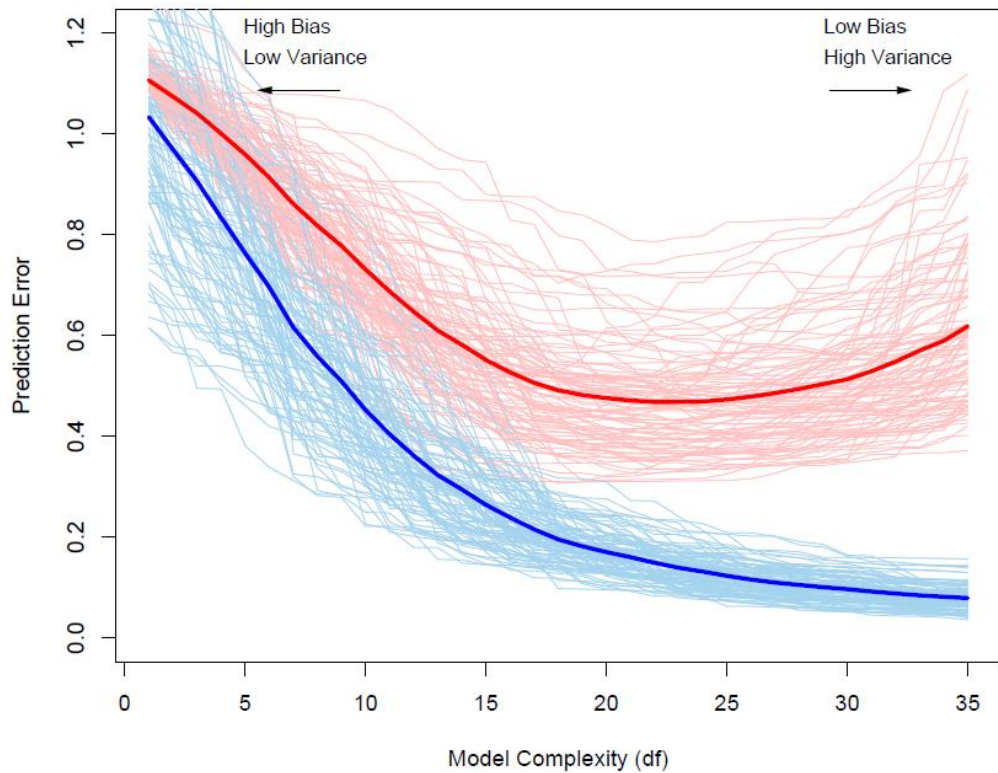
$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- 当加入新变量时，RSS单调下降。因此，更**多**的自变量会得到更**大**的 R^2
- The more, the better?

No No No! 容易造成**过拟合**!



多元回归中的 R^2



训练集

测试集



Adjusted R^2

$$R_a^2 = 1 - \frac{RSS/(n - p - 1)}{TSS/(n - 1)} = 1 - \frac{(n - 1)}{(n - p - 1)} \frac{RSS}{TSS}$$

通过自由度调整，对加入更多变量进行“惩罚”。



扩展：是否存在线性关系？

```
> summary(lm(price~ school + AREA, data = dat0))  
Call:  
lm(formula = price ~ school + AREA, data = dat0)  
Residuals:  
    Min       1Q   Median       3Q      Max   
-5.6618 -1.3745 -0.3286  1.1538  9.6872  
Coefficients:  
            Estimate Std. Error t value Pr(>|t|)      
(Intercept)  5.5881160  0.0359067 155.629  < 2e-16 ***  
school       2.5722336  0.0322103  79.858  < 2e-16 ***  
AREA        -0.0027526  0.0003364  -8.181  3.02e-16 ***  
---  
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 1.883 on 16207 degrees of freedom  
Multiple R-squared:  0.2863,    Adjusted R-squared:  0.2862  
F-statistic: 3251 on 2 and 16207 DF,  p-value: < 2.2e-16
```

与单变量回归
系数是否相同？



扩展：是否存在线性关系？

```
> summary(lm(price~ school + AREA + CATE, data = dat0))

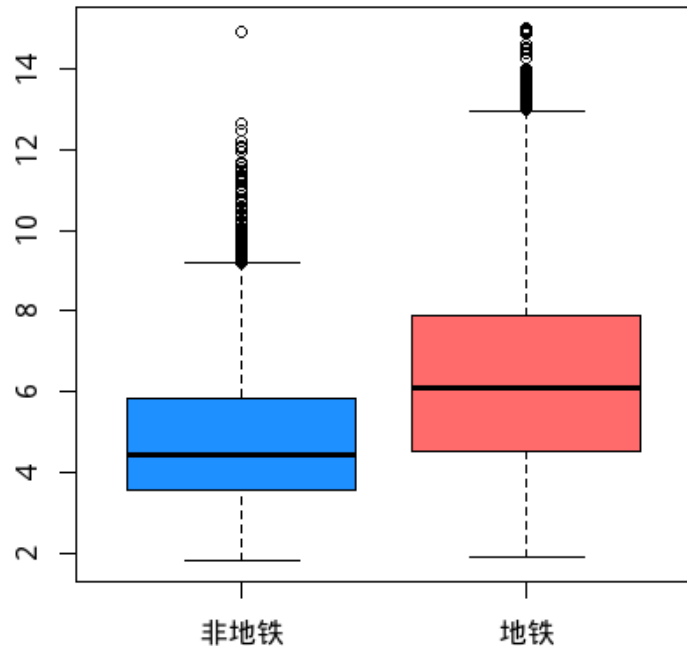
Call:
lm(formula = price ~ school + AREA + CATE, data = dat0)

Residuals:
    Min       1Q   Median       3Q      Max
-5.7442 -0.9176 -0.1425  0.8249  8.2456

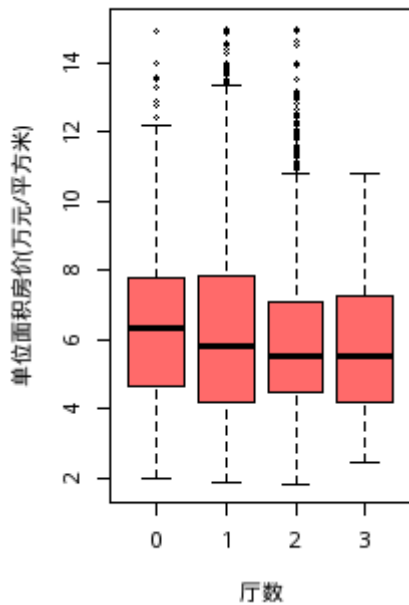
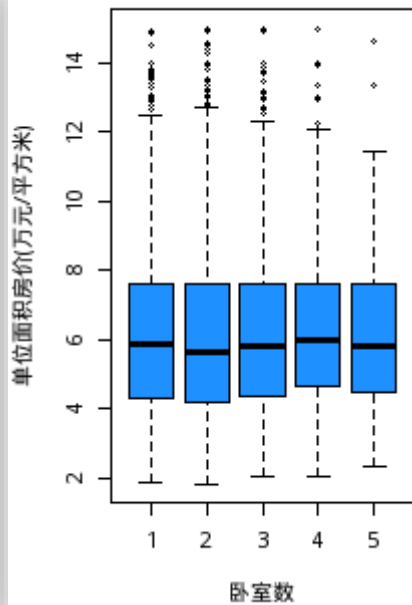
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.4208740  0.0393383  163.222 < 2e-16 ***
school       1.2458616  0.0280696   44.385 < 2e-16 ***
AREA        -0.0014385  0.0002644   -5.440 5.41e-08 ***
CATE石景山  -2.2882854  0.0446501  -51.249 < 2e-16 ***
CATE丰台    -2.0762116  0.0399083  -52.025 < 2e-16 ***
CATE朝阳    -1.2476593  0.0390133  -31.980 < 2e-16 ***
CATE东城     0.3263657  0.0384931   8.479 < 2e-16 ***
CATE西城     1.5643723  0.0388286  40.289 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.452 on 16202 degrees of freedom
Multiple R-squared:  0.576,    Adjusted R-squared:  0.5758
F-statistic: 3144 on 7 and 16202 DF, p-value: < 2.2e-16
```

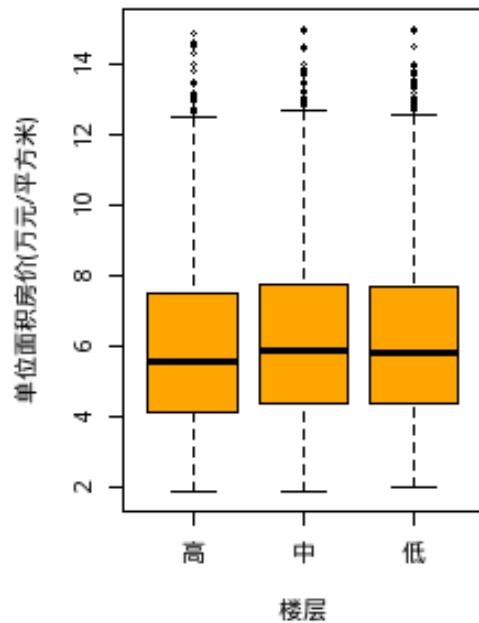
与单变量回归
系数是否相同？



其他因素：房屋室内设计？



其他因素：楼层？





多元线性回归模型（因变量：单位面积房价）

变量	回归系数	p值	备注
截距项	3.315	<.0001	
城区-丰台	0.131	0.001	
城区-朝阳	0.875	<.0001	
城区-东城	2.443	<.0001	基准组：石景山区
城区-海淀	2.191	<.0001	
城区-西城	3.705	<.0001	
学区房	1.183	<.0001	
地铁房	0.672	0.001	
楼层-中层	0.152	<.0001	基准组：高层
楼层-低层	0.198	<.0001	
有客厅	0.163	<.0001	
卧室数	0.111	<.0001	
房间面积	-0.002	<.0001	
F检验	p值<.0001	调整的R2	0.5901



线性回归模型：结果解读

控制其他因素不变时

- **城区**：石景山区单位面积房价最低，西城区单位面积房价最高，比石景山区每平方米平均高出**3.705**万元
- **学区房**比非学区房单位面积房价平均高出**1.18**万元
- **地铁房**比非地铁房单位面积房价平均高出**6720**元
- **高层房屋**单位面积房价最低，其次是中层，低层房屋单位面积房价最高
- **有客厅**的房子单位面积房价更高
- **卧室数**每增加一间，单位面积房价平均增加**1110**元
- **房屋面积**的增加会带来单位面积房价的降低



回归诊断



回归诊断: R语言实现



线性回归模型

```
lm1 = lm(price ~ CATE + school + subway + style + floor + bedrooms + AREA,  
data = dat0)
```

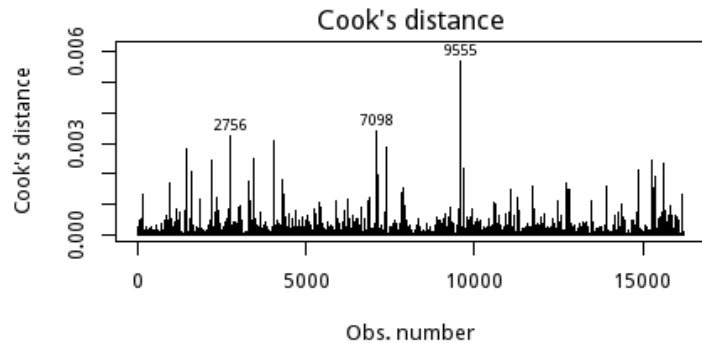
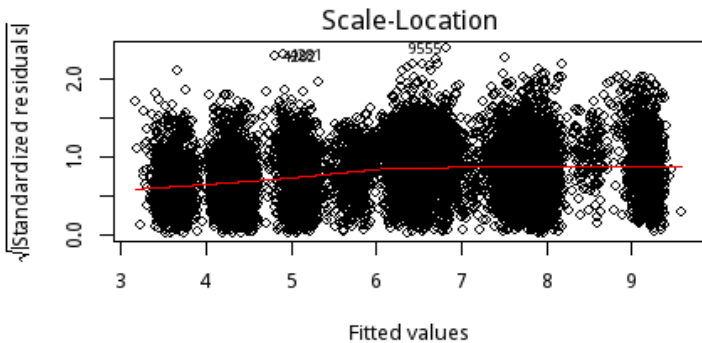
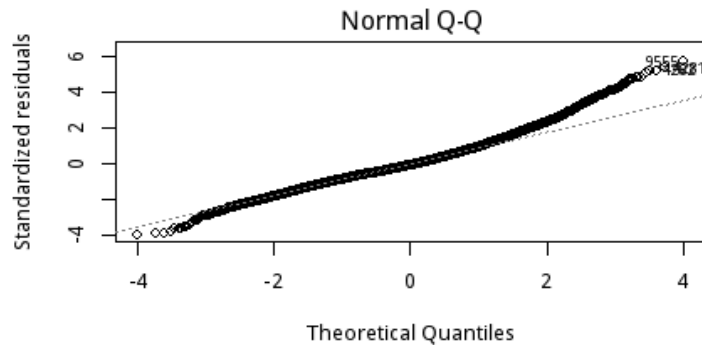
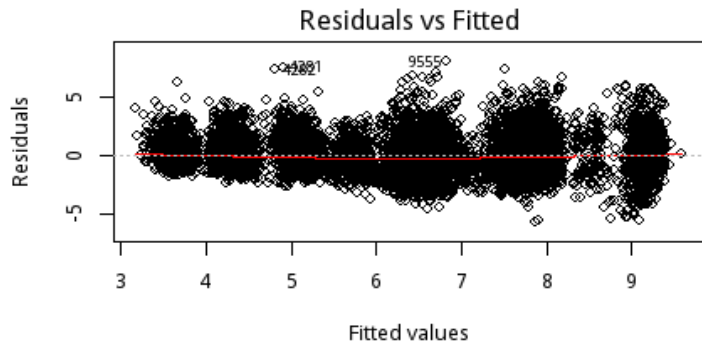
```
summary(lm1) #回归结果展示
```

```
par(mfrow = c(2, 2)) #画2*2的图
```

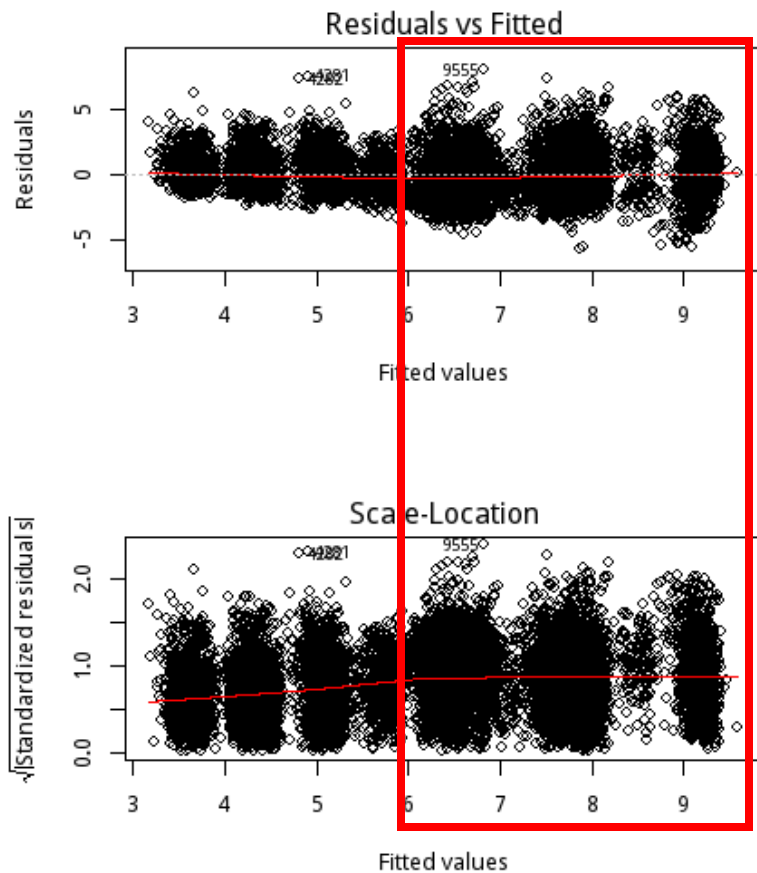
```
plot(lm1, which = c(1:4)) #模型诊断图
```



回归诊断



回归诊断



线性?

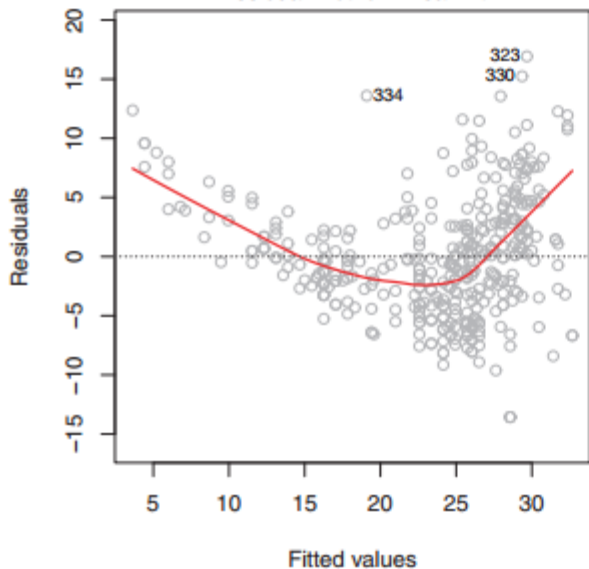


异方差?

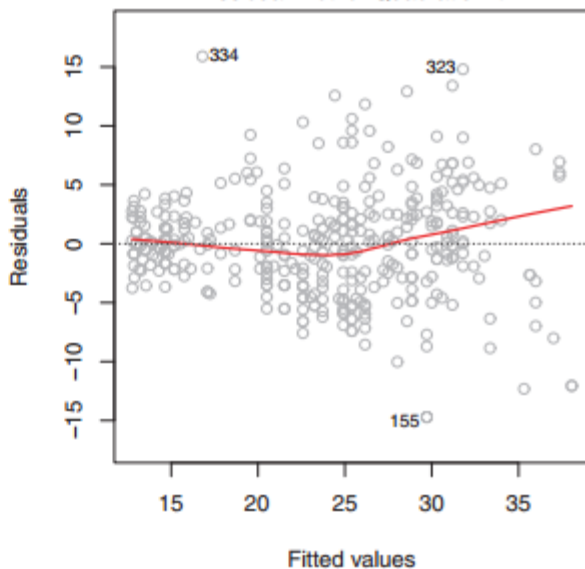


补充：其他常见示例

Residual Plot for Linear Fit



Residual Plot for Quadratic Fit



Problem: 非线性!

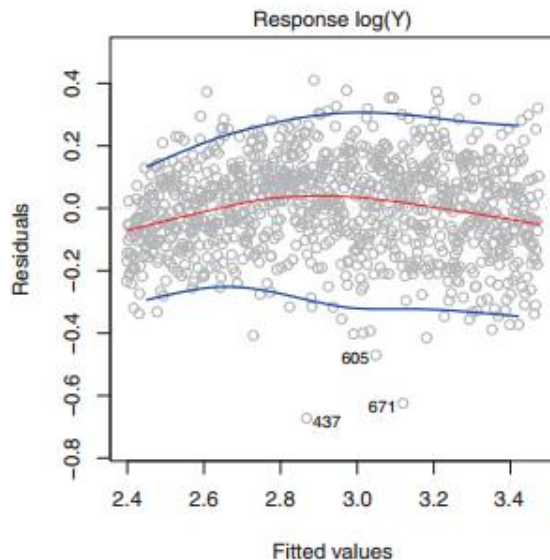
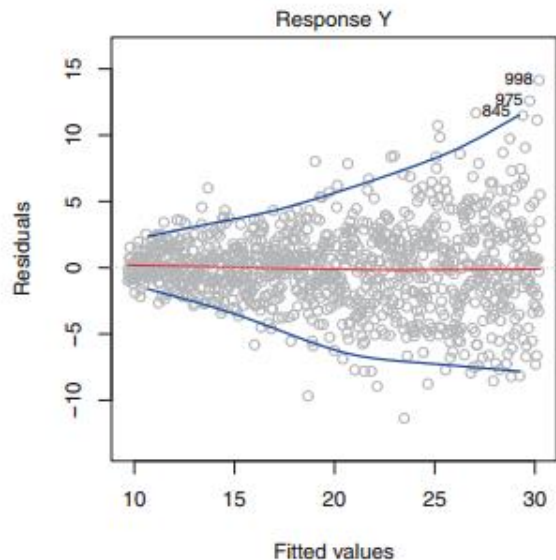
Solution:

加入非线性项:

e.g., X^2 , $\log(X)$, \sqrt{X} 等



补充：其他常见示例



Problem: 异方差!

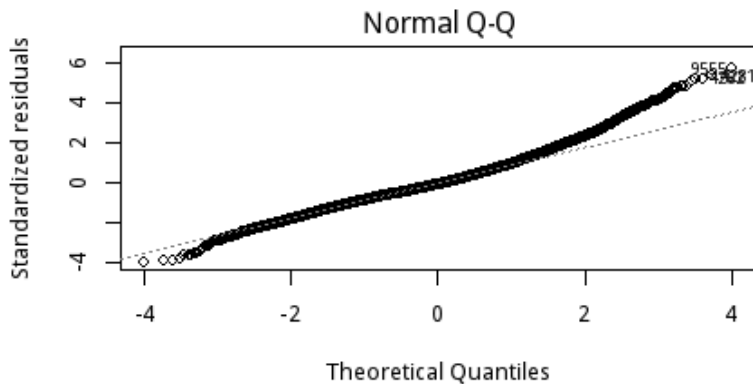
Solution:

对因变量进行变换:

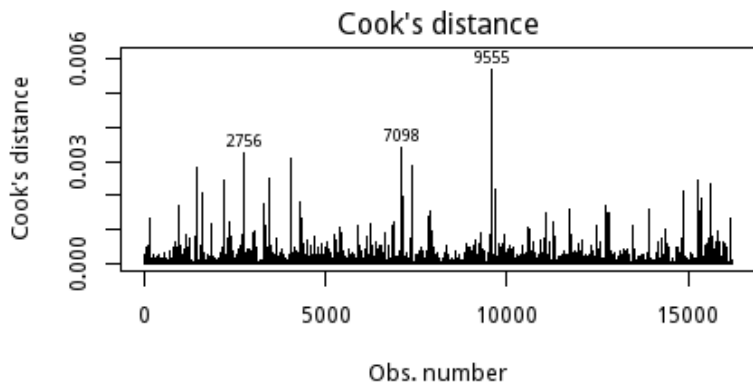
e.g., $\log(Y)$, \sqrt{Y} 等



回归诊断



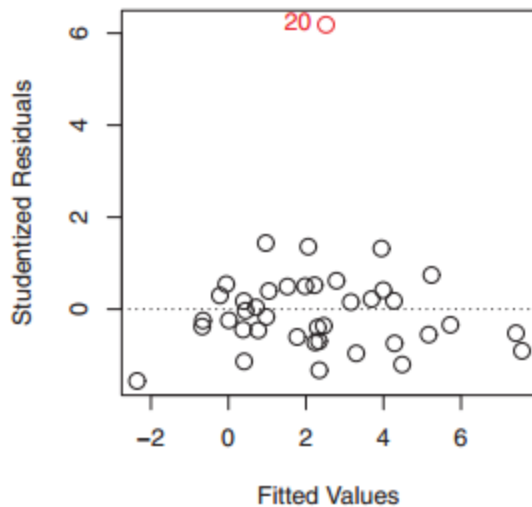
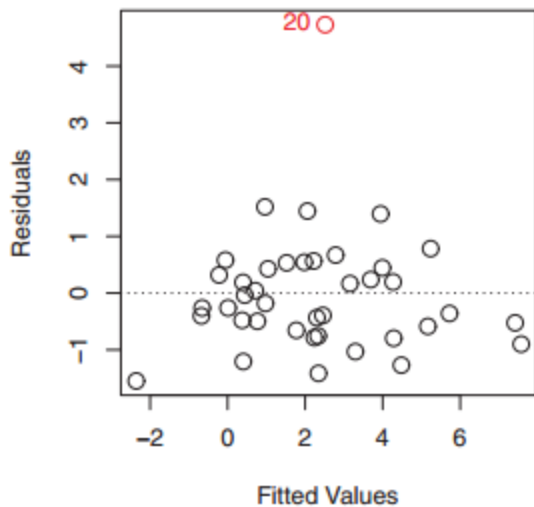
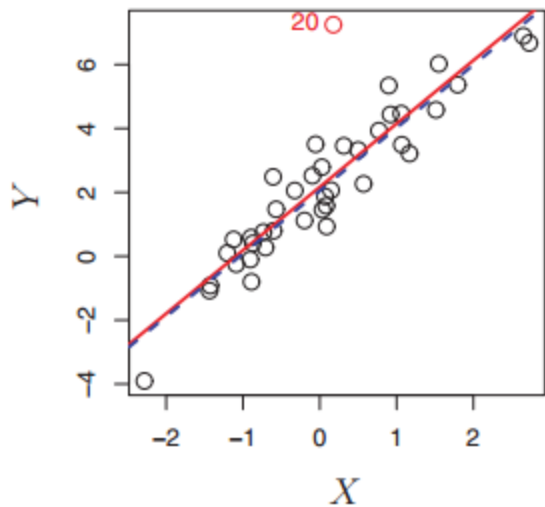
正态性诊断



Outlier: Cook's Distance
一般认为 $>4/n$, 或者 >1



示例：强影响点



强影响点会对回归曲线造成较大影响（对回归方程斜率& R^2 造成较大影响）



共线性

```
> library(car)
> vif(lm1)
```

CATE丰台	CATE朝阳	CATE东城
2.09	2.21	2.31
CATE海淀	CATE西城	school
2.29	2.32	1.30
subway	style有厅	floormiddle
1.08	1.05	1.32
floorlow	bedrooms	AREA
1.32	2.13	2.18

Problem:

回归时，F检验显著，但是单个系数不显著

方差膨胀因子:

$$VIF = (1 - R_i^2)^{-1}$$

一般认为 **$VIF > 10$** ，则存在多重共线性

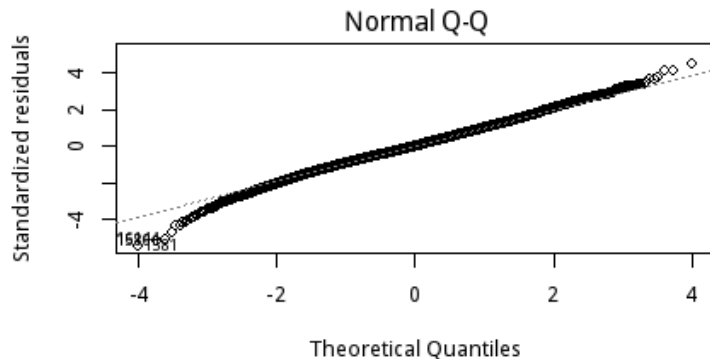
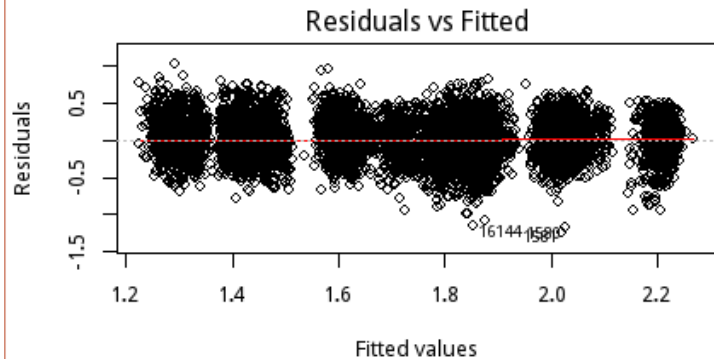


对数线性回归: $\text{Log}(Y)$

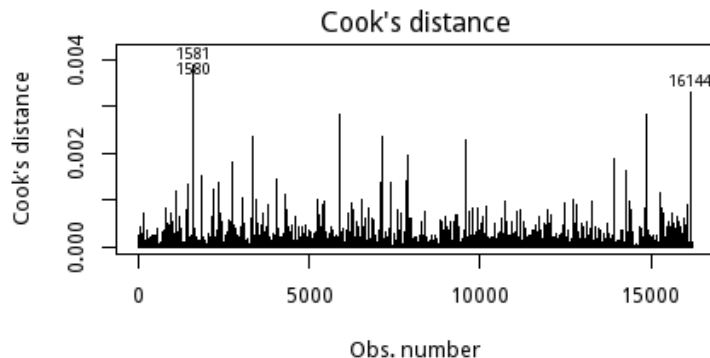
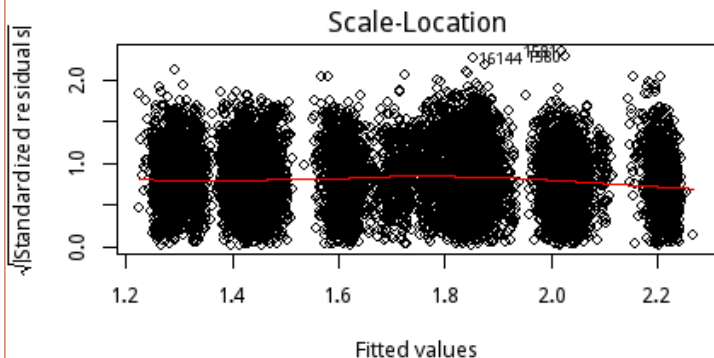
变量	回归系数 ($\times 10^{-1}$)	p 值	备注
截距项	12.360	<0.001	
城区-丰台	0.441	<0.001	基准组: 石景山组
城区-朝阳	2.057	<0.001	
城区-东城	4.577	<0.001	
城区-海淀	4.320	<0.001	
城区-西城	6.270	<0.001	
学区房	1.719	<0.001	
地铁房	1.282	<0.001	
楼层-中层	0.152	<0.001	基准组: 高层
楼层-底层	0.198	<0.001	
有客厅	0.275	0.001	
卧室数	0.140	<0.001	
房间面积	-0.003	<0.001	
F 检验	p 值<0.0001	调整的 R^2	0.6079



对数线性回归: $\text{Log}(Y)$



异方差得
到改善!





对数线性模型：结果解读

$$\beta_j = \frac{d \log(y)}{d x_j} = \frac{dy/y}{dx}$$

- 与线性模型不同，对数线性模型的系数估计解读为“**增长率**”

- 控制其他因素不变时

- 城区**：石景山区单位面积房价最低，西城区单位面积房价最高，比石景山区平均贵**62.70%**

- 学区房**比非学区房单位面积房价平均贵**17.19%**

- 地铁房**比非地铁房单位面积房价平均贵**12.82%**

- 高层房屋**单位面积房价最低，其次是中层，低层房屋单位面积房价最高

- 有客厅**的房子单位面积房价更高，平均贵**2.75%**

- 卧室数**每增加一间，单位面积房价平均增加**1.41%**

- 房屋面积**的增加会带来单位面积房价的降低



交互效应



交互效应

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon.$$



Interaction
Term

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$



$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned}$$



交互效应: 学区房×面积

$$\beta_0 + \beta_1 \times \text{面积} + \beta_2 \times \text{学区} + \beta_3 \times (\text{面积} \cdot \text{学区}) + \varepsilon$$

$$Y = \begin{cases} \beta_0 + \beta_1 \times \text{面积} + \varepsilon & \text{非学区} \\ \beta_0 + \beta_2 + (\beta_1 + \beta_3) \times \text{面积} + \varepsilon' & \text{学区} \end{cases}$$



交互效应: 学区房×面积

```
> summary(lm(log(price)~school+AREA+school*AREA,data=dat0))
```

Call:

```
lm(formula = log(price) ~ school + AREA + school * AREA, data = dat0)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-1.22690	-0.21873	-0.01225	0.21458	1.09244

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	1.6015432	0.0068784	232.84	< 2e-16 ***
school	0.5548057	0.0118299	46.90	< 2e-16 ***
AREA	0.0001936	0.0000672	2.88	0.00398 **
school:AREA	-0.0015622	0.0001171	-13.34	< 2e-16 ***

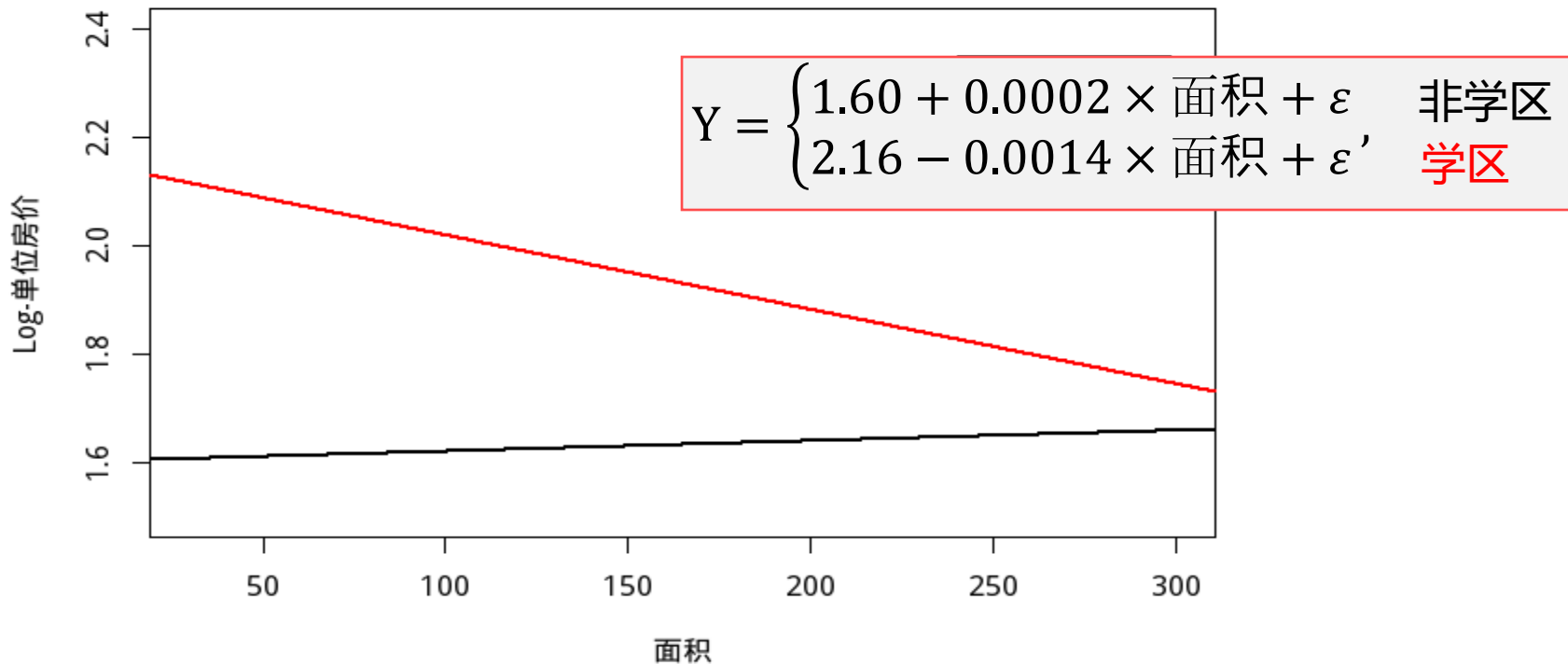
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3081 on 16206 degrees of freedom
Multiple R-squared: 0.2835, Adjusted R-squared: 0.2834
F-statistic: 2138 on 3 and 16206 DF, p-value: < 2.2e-16

如何解释?



交互效应: 学区房×面积

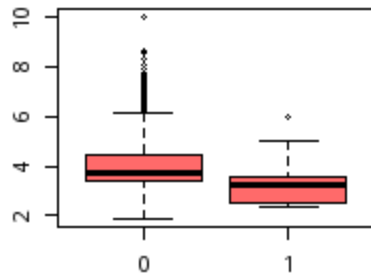




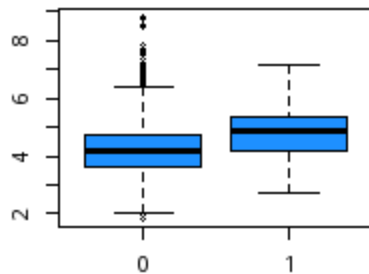


交互效应: 城区×学区房

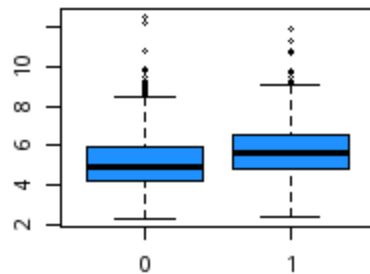
石景山区



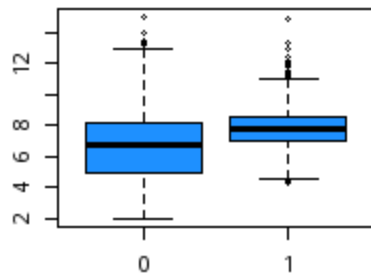
丰台



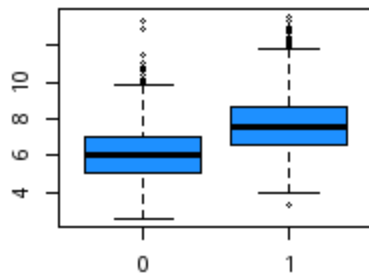
朝阳



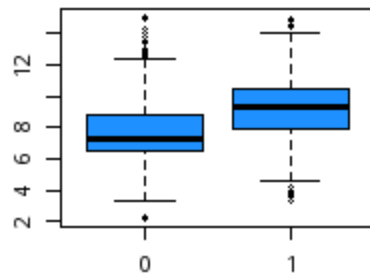
东城



海淀



西城





对数线性交互效应模型

变量	回归系数 ($\times 10^{-1}$)	p 值	备注
截距项	12.410	<0.001	
城区-丰台	0.429	<0.001	
城区-朝阳	2.184	<0.001	
城区-东城	4.467	<0.001	基准组：石景山组
城区-海淀	4.121	<0.001	
城区-西城	6.177	<0.001	
学区房	-1.800	<0.001	
地铁房	1.257	<0.001	
楼层-中层	0.263	<0.001	基准组：高层
楼层-底层	0.343	<0.001	
有客厅	0.270	0.001	
卧室数	0.140	<0.001	
房间面积	-0.003	<0.001	
丰台 \times 学区	2.948	<0.001	基准组： 石景山 \times 学区
朝阳 \times 学区	2.780	<0.001	
东城 \times 学区	3.706	<0.001	
海淀 \times 学区	3.876	<0.001	
西城 \times 学区	3.638	<0.001	
F 检验	p 值<0.0001	调整的 R^2	0.6108



交互模型：结果解读

- 整体而言（不考虑学区交互项），西城区与海淀区的单位面积房价之比
 - 对数线性模型： $e^{0.627-0.432} = e^{0.195} = 1.215$
- **学区房房价哪家强？**
 - 学区房：西城区与海淀区的单位面积房价之比
 - 交互模型： $e^{0.617+0.363-0.412-0.387} = e^{0.181} = 1.198$
 - 非学区房：西城区与海淀区的单位面积房价之比
 - 交互模型： $e^{0.617-0.412} = e^{0.205} = 1.228$



交互效应: 城区×学区房

原因:

1. 石景山区学区资源相对较差
2. 石景山区学区房入样比例较低, 可能存在偏差

城区	样本量	学区房占比 (%)
石景山	1947	0.92
丰台	2947	3.18
朝阳	2864	20.84
东城	2783	45.81
海淀	2919	47.48
西城	2750	56.10



超多水平变量

```
> head(dat0)
  CATE bedrooms halls AREA floor subway school price LONG LAT NAME DISTRICT style
1 朝阳          1    0 46.06 middle      1      0 4.8850 116.4597 39.92835 10AM新坐标 方庄 无厅
2 朝阳          1    1 59.09 middle      1      0 4.6540 116.4597 39.92835 10AM新坐标 方庄 有厅
3 海淀          5    2 278.95 high       1      1 7.1662 116.3036 39.95481 10号名邸 紫竹桥 有厅
4 海淀          3    2 207.00 high       1      1 5.7972 116.3036 39.95481 10号名邸 紫竹桥 有厅
5 丰台          2    1 53.32 low        1      1 7.1268 116.4188 39.94381 17号旁门 蒲黄榆 有厅
6 丰台          2    1 58.00 low        1      1 7.0690 116.4188 39.94381 17号旁门 蒲黄榆 有厅

> tab_dist = table(dat0$DISTRICT)
> length(tab_dist)
[1] 173
> sort(tab_dist, decreasing = T)[1:10]

  鲁谷   马甸   望京  苹果园  广渠门  广安门  崇文门  东直门  清河  六铺炕
   918   418   409   400   366   333   322   288   283   255

> quantile(tab_dist)
0%  25%  50%  75% 100%
 2    29   68  112  918
```

173个水平 = 172个0-1变量!



超多水平变量

1. 保留出现较多的 区域， code成新变量

```
dist_high = names(tab_dist[tab_dist>=100])
dat0$DISTRICT1 = ""
ind = is.element(dat0$DISTRICT, dist_high)
dat0$DISTRICT1[ind] = as.character(dat0$DISTRICT[ind])
dat0$DISTRICT1 = factor(dat0$DISTRICT1)

lm5 = lm(log(price)~CATE*school+subway+style+floor+bedrooms+AREA+DISTRICT1 , data=dat0)
summary(lm5)
```

共72个变量:

Multiple R-squared: 0.6992

Adjusted R-squared: 0.6979



变量选择



给定一个准则: AIC BIC p-value Adjusted R^2

线性回归变量选择:

- 向前选择 (forward selection) :每次添加一个预测变量到模型中,直到添加变量不会使模型有所改进为止。
- 向后选择 (backward selection) :从模型包含所有预测变量开始,一次删除一个变量直到会降低模型质量为止。
- **逐步回归** (stepwise selection) :变量每次进入一个,但是每一步中,变量都会被重新评价,对模型没有贡献的变量将会被删除



变量选择

```
mat_dist = model.matrix(~ DISTRICT1, data = dat0)[,-1]
dat1 = data.frame(dat0[,c("CATE", "school", "price", "subway", "style", "floor", "bedrooms", "AREA")],
                  mat_dist)

lm6 = lm(log(price)~CATE*school+., data=dat1)
reg_table = summary(stepAIC(lm6))
```

R function:

```
step(object, scope, scale = 0,
direction = c("both", "backward",
"forward"), trace = 1, keep = NULL,
steps = 1000, k = 2)
```

共64个变量:

Multiple R-squared: 0.6991

Adjusted R-squared: 0.698

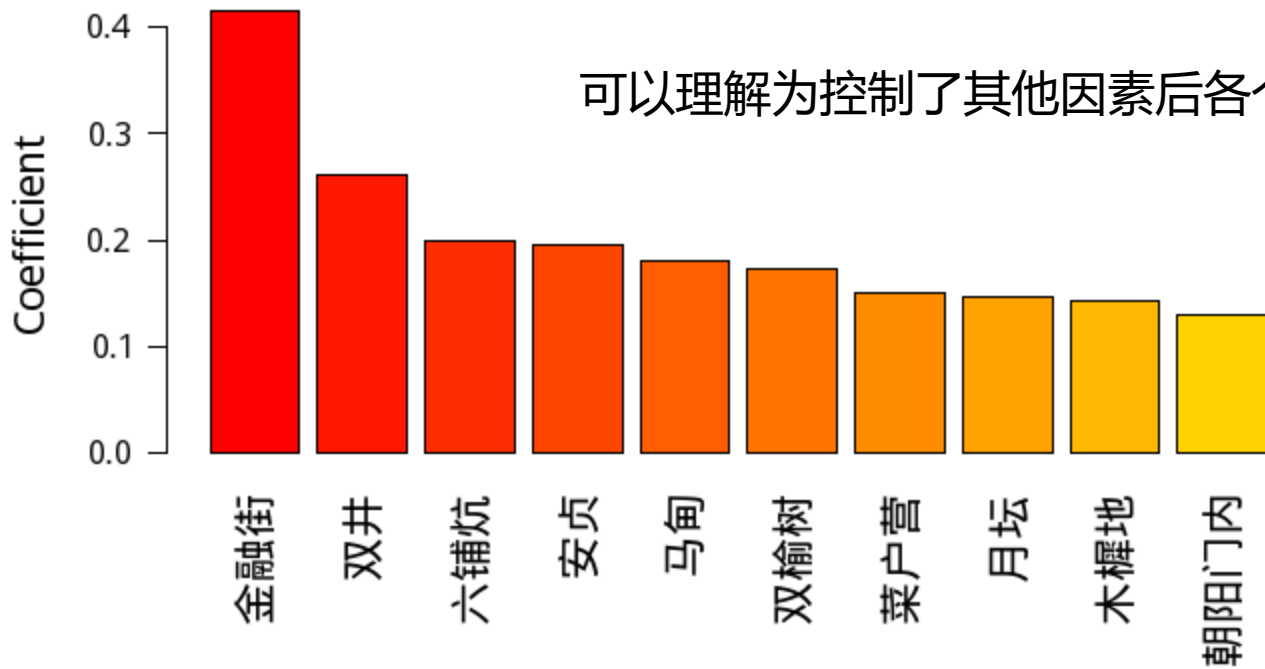
两个重要的参数:

direction: 变量选择方向

k: 可以转换AIC&BIC 准则



十大“寸土寸金”区域







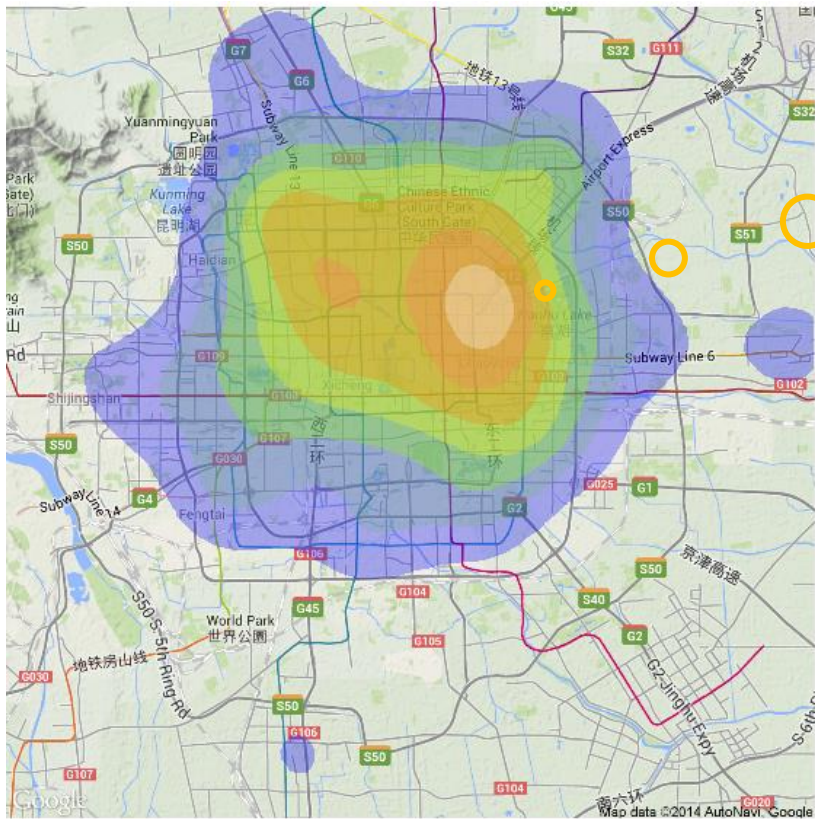
获取外部数据

越繁华的地段
房价越高，是
这样的吗？





朝阳区、
东城区





回归分析结果

Multiple R-squared:

0.7157

Adjusted R-squared:

0.7144

```

DISTRICT1苏州桥      8.570e-02  1.730e-02   4.955 7.30e-07 ***
DISTRICT1太平桥      8.361e-02  1.983e-02   4.217 2.49e-05 ***
DISTRICT1陶然亭      1.574e-01  1.494e-02  10.533 < 2e-16 ***
DISTRICT1天宁寺     -1.485e-01  1.979e-02  -7.507 6.39e-14 ***
DISTRICT1望京       1.278e-01  1.079e-02  11.843 < 2e-16 ***
DISTRICT1西罗园     -2.949e-01  1.424e-02 -20.701 < 2e-16 ***
DISTRICT1西三旗     -1.886e-01  1.636e-02 -11.526 < 2e-16 ***
DISTRICT1西直门     -6.708e-02  1.710e-02  -3.923 8.80e-05 ***
DISTRICT1亚运村      3.258e-02  1.993e-02   1.635 0.102076
DISTRICT1杨庄       -8.339e-04  2.087e-02  -0.040 0.968121
DISTRICT1永定门     -1.971e-01  2.028e-02  -9.717 < 2e-16 ***
DISTRICT1右安门内  -2.466e-02  1.993e-02  -1.237 0.216116
DISTRICT1玉泉路      7.688e-03  2.075e-02   0.371 0.710956
DISTRICT1玉泉营      6.573e-02  1.411e-02   4.658 3.22e-06 ***
DISTRICT1月坛       1.502e-01  1.660e-02   9.048 < 2e-16 ***
DISTRICT1赵公口     -2.800e-02  1.936e-02  -1.446 0.148193
DISTRICT1知春路      3.395e-02  2.053e-02   1.653 0.098261 .
DISTRICT1紫竹桥     -2.561e-02  1.648e-02  -1.554 0.120267
restN                2.399e-03  7.839e-05  30.606 < 2e-16 ***
CATE丰台:school      2.036e-01  5.140e-02   3.961 7.50e-05 ***
CATE朝阳:school      2.062e-01  4.794e-02   4.300 1.72e-05 ***
CATE东城:school      2.544e-01  4.793e-02   5.308 1.12e-07 ***
CATE海淀:school      2.600e-01  4.780e-02   5.440 5.42e-08 ***
CATE西城:school      1.850e-01  4.789e-02   3.864 0.000112 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1945 on 16137 degrees of freedom
Multiple R-squared:  0.7157,    Adjusted R-squared:  0.7144
F-statistic: 564.2 on 72 and 16137 DF,  p-value: < 2.2e-16

```

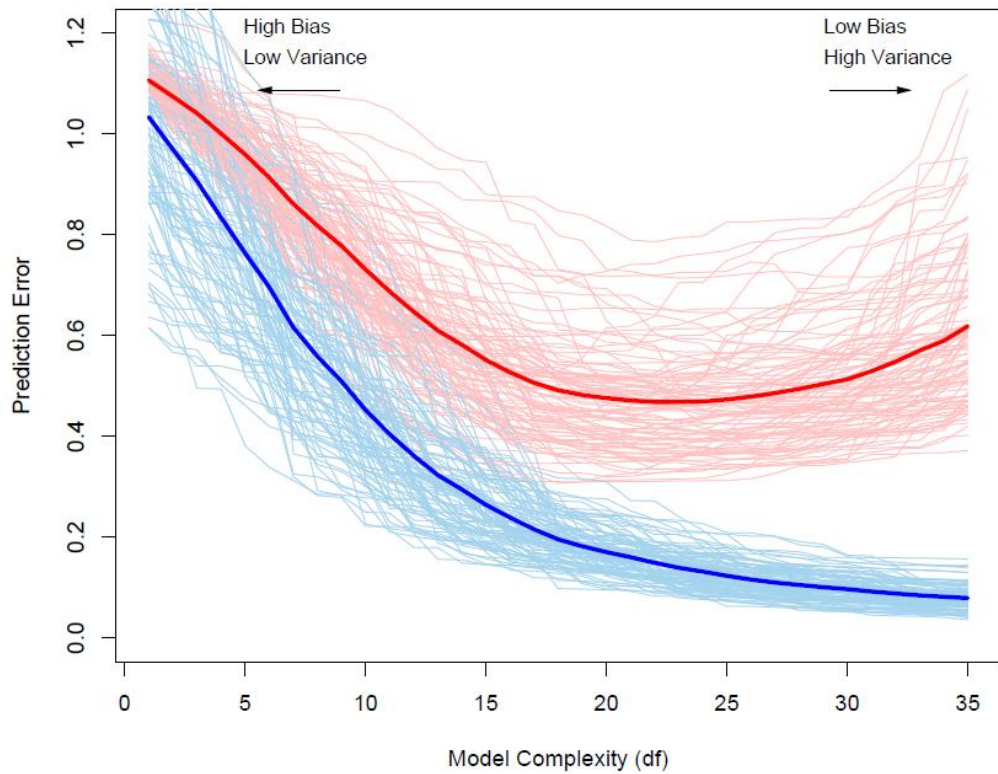
> |



回归预测



预测

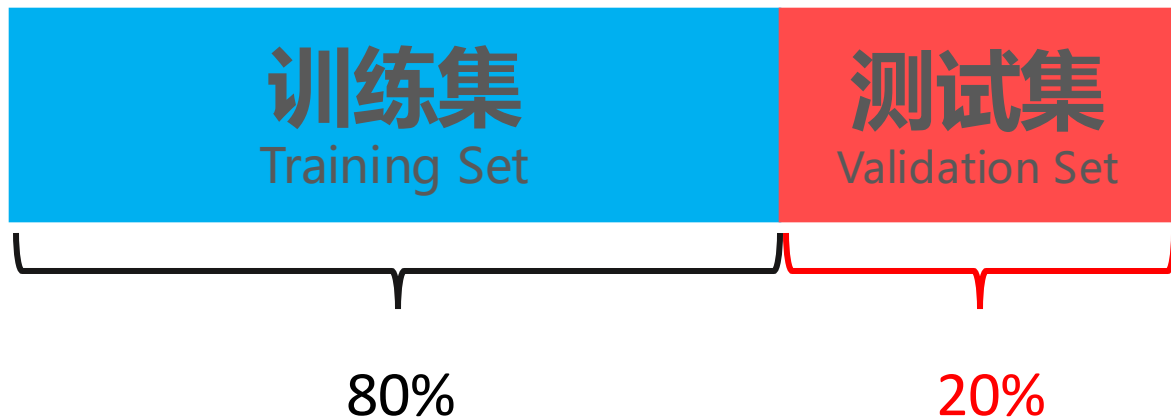


—— 训练集

—— 测试集

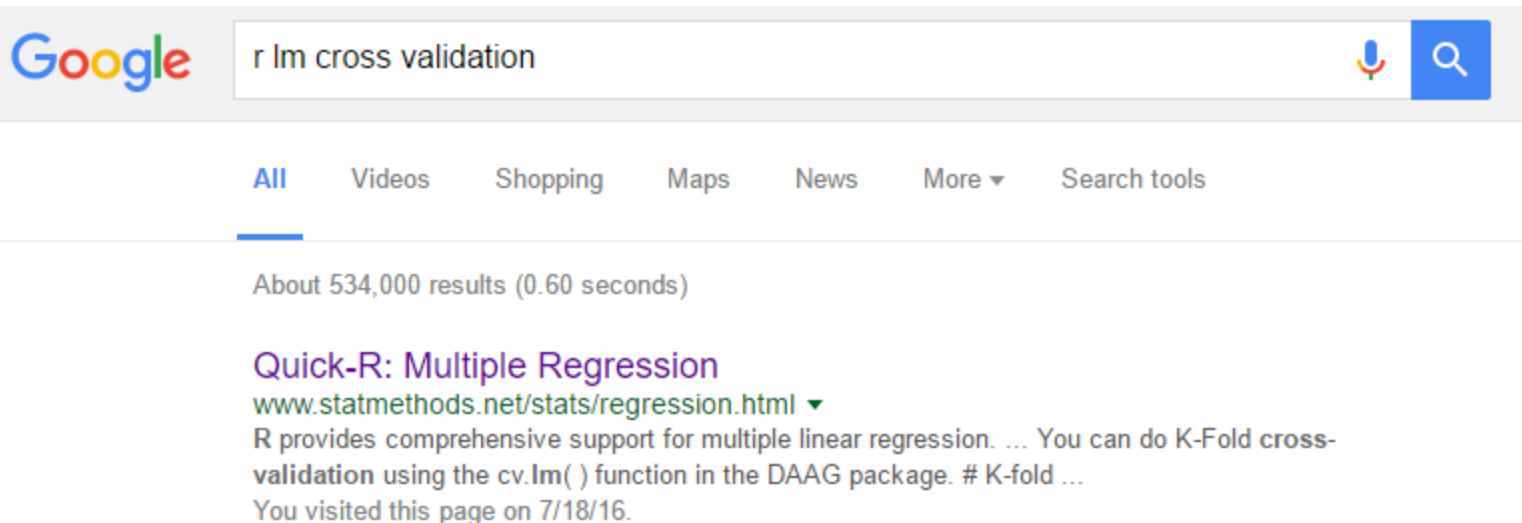


5折交叉验证





5折交叉验证：R语言实现



Google search results for "r lm cross validation". The search bar shows the query "r lm cross validation" and a microphone icon. Below the search bar, the "All" tab is selected. The results show "About 534,000 results (0.60 seconds)". The first result is titled "Quick-R: Multiple Regression" with a link to www.statmethods.net/stats/regression.html. The snippet describes R's support for multiple linear regression and mentions K-Fold cross-validation using the `cv.lm()` function in the DAAG package. It also notes the page was visited on 7/18/16.

Google

r lm cross validation

All Videos Shopping Maps News More ▾ Search tools

About 534,000 results (0.60 seconds)

Quick-R: Multiple Regression
www.statmethods.net/stats/regression.html ▾
R provides comprehensive support for multiple linear regression. ... You can do K-Fold cross-validation using the `cv.lm()` function in the DAAG package. # K-fold ...
You visited this page on 7/18/16.



5折交叉验证：R语言实现

Approach 1：使用DAAG中的cv.lm() 函数

```
> library(DAAG)
> system.time({cvlm4 = cv.lm(data = dat0, lm4, m=5, printit = F, seed = 1234)})
```

用户 系统 流逝

0.72 0.77 1.48

Warning message:

In cv.lm(data = dat0, lm4, m = 5, printit = F, seed = 1234) :

As there is >1 explanatory variable, cross-validation
predicted values for a fold are not a linear function
of corresponding overall predicted values. Lines that
are shown for the different folds are approximate

```
> attr(cvlm4, "ms")
```

```
[1] 0.0516
```

缺点：计算慢，参数较复杂，许多不必要输出



5折交叉验证：R语言实现

Approach 2: 使用bootstrap中的crossval() 函数

```
library(bootstrap)

# define functions
theta.fit = function(x,y){lsfit(x,y)}
theta.predict = function(fit,x){cbind(1,x)%*%fit$coef}

system.time({
# matrix of predictors
mat01 = model.matrix(~ CATE+ floor+style, data = dat0)[,-1]
X = cbind(mat01, dat0[,c(2:4,6:7)])
# vector of predicted values
y = as.matrix(log(dat0$price))

set.seed(1234)
pred_cv = crossval(X,y, theta.fit, theta.predict, ngroup=5)
})
sqrt(mean((pred_cv$cv.fit-log(dat$price))^2))
```

```
用户 系统 流逝
0.08 0.02 0.09
> sqrt(mean((pred_cv$cv.fit-log(dat$price))^2))
[1] 0.226
> |
```

优点：速度快！

缺点：要预处理成矩阵，差评！



5折交叉验证：R语言实现

Approach 3：自己动手，丰衣足食！（示例：如何动手写一个函数）

```
pred.cv<-function(dat, k)
{
  ind = sample(1:k, nrow(dat), replace = T)
  pred_cv = rep(0, nrow(dat))
  for (i in 1:k)
  {
    ii = which(ind==i)
    obj = lm(log(price)~CATE*school+subway+style+floor+bedrooms+AREA,
              data = dat[-ii,])
    pred_cv[ii] = predict(obj, dat0[ii,])
  }
  rmse = sqrt(mean((pred_cv-log(dat$price))^2))
  return(list(pred_cv = pred_cv, rmse = rmse))
}
```

```
set.seed(1234)
system.time({pred_cv = pred.cv(dat = dat0, k = 5)})
pred_cv$rmse
```

```
> system.time({pred_cv = pred.cv(dat = dat0, k = 5)})
用户 系统 流逝
0.24 0.06 0.30
> pred_cv$rmse
[1] 0.227
```



5折交叉验证：R语言实现

重复k次，做评估

```
> set.seed(1234)
> rmsees = rep(0, 50)
> for (i in 1:50)
+ {
+   cat(i, "\r")
+   pred_cv = pred.cv(dat = dat0, k = 5)
+   rmsees[i] = pred_cv$rmse
+ }
50
> mean(rmsees)
[1] 0.227
> |
```

THANKS!