- 在第2-5章中我们已经讨论了线性回归模型与OLS估计值的理论性质
- 本章将讨论实际应用中的几个重要问题，例如：

（1）数据测度单位改变会不会影响OLS标准误及相关统计量的取值？

（2）怎么刻画变化的偏效应？

（3）怎么选择回归元？等

JEFFREY M. WOOLDRIDGE

Introductory Econometrics
A Modern Approach

SIXTH EDITION

Chapter 6

多元回归分析：深入专题

章节框架

- 在这一章中我们介绍回归模型实际应用中的几个重要问题：
- 首先，我们介绍数据测度单位对OLS统计量的影响
- 然后，我们对函数形式进行进一步讨论
- 之后，我们进一步探讨拟合优度和回归元选择
- 最后，我们讨论预测的构造

数据测度单位对OLS统计量的影响

- 在第二章中我们已经讨论了解释变量与被解释变量测度单位对OLS估计值的影响

- 现在我们考虑测度单位对标准误、*t* 统计量，*F* 统计量和置信区间的影响

- 例子：婴儿出生体重与孕妇抽烟量

$y = cigs/20 \cdot (\hat{\beta}_1 \cdot 20)$ t ...

$$\widehat{bwght} = \hat{\beta}_0 + \hat{\beta}_1 cigs + \hat{\beta}_2 faminc$$

$y/16$ $\hat{\beta}_0/16 + \hat{\beta}_1/16\ cigs + \hat{\beta}_2/16\ faminc.$

其中 $bwght$ 表示以盎司为单位的婴儿出生体重， $cigs$ 表示母亲每天抽烟量， $faminc$ 表示以千美元为单位的家庭年收入

以下变化对估计值与统计量有什么影响？

（1）如果$bwght$ 以磅为单位（取值/16）

(2)如果$cigs$ 以包为单位（取值/20）

数据测度单位对OLS统计量的影响

TABLE 6.1 Effects of Data Scaling

Dependent Variable	(1) *bwght*	(2) *bwghtlbs*	(3) *bwght*
Independent Variables			
cigs	−.4634 (.0916)	−.0289 (.0057)	—
packs	—	—	−9.268 (1.832)
faminc	.0927 (.0292)	.0058 (.0018)	.0927 (.0292)
intercept	116.974 (1.049)	7.3109 (.0656)	116.974 (1.049)
Observations	1,388	1,388	1,388
R-Squared	$\boxed{.0298}$	$\boxed{.0298}$	$\boxed{.0298}$
SSR	557,485.51	2,177.6778	557,485.51
SER	20.063	1.2539	20.063

$\frac{-0.4634}{0.0916} = -5.06$

$\frac{-0.4634}{16}$

$-0.4634 \times 20 =$

$0.0916 \times 20 =$

$\frac{557485.51}{16^2}$

$\frac{20.063}{16}$

$\sqrt{SSR/(n-k-1)}$

对函数形式的进一步讨论

- **例子：污染对住房价格的影响**

显著性 ← $\begin{cases} x \to \log x \\ y \to \log y. \end{cases}$

$$\log(\widehat{price}) = \boxed{9.23} - 0.718\log(nox) + 0.306rooms$$

c c

- **对数形式的讨论：**

$\log(price) \to \log(c\cdot price) = \log c + \log price.$

< 0.1% → 0.1%

> 99.9%

Winsorize

trimming.

  - 方便用于百分比或弹性解释
  - 对数变量的斜率系数无关于单位变化
  - 取对数通常可以消除或缓解异常值问题
  - 取对数通常有助于确~~保~~正态性和同方差性
  - 以年等为单位变量不应取对数
  - 百分数为单位的变量也不应取对数
  - 变量有零值或负值都不能取对数

OLS. → Mean "squared" error.
Absolute.

$Var(u|x) = \sigma^2$

Robust.

0.2, 20%

8% → 9%

百分点 变化, 1%

百分比, 12.5%

对函数形式的进一步讨论

- 为了描述递减或递增的边际效应，有时需要二次函数形式
- 例子：工资等式

凹函数

正

负

$$\widehat{wage} = \underset{(.35)}{3.73} \boxed{+ \underset{(.041)}{.298}} exper \boxed{- \underset{(.0009)}{.0061}} exper^2 + IQ + edu$$

$$n = 526, R^2 = .093$$

- 经验的边际效应

第一年的工作经验会增加大约0.30美元的工资，第二年会增加0.298−2(0.0061)=0.29美元，以此类推。

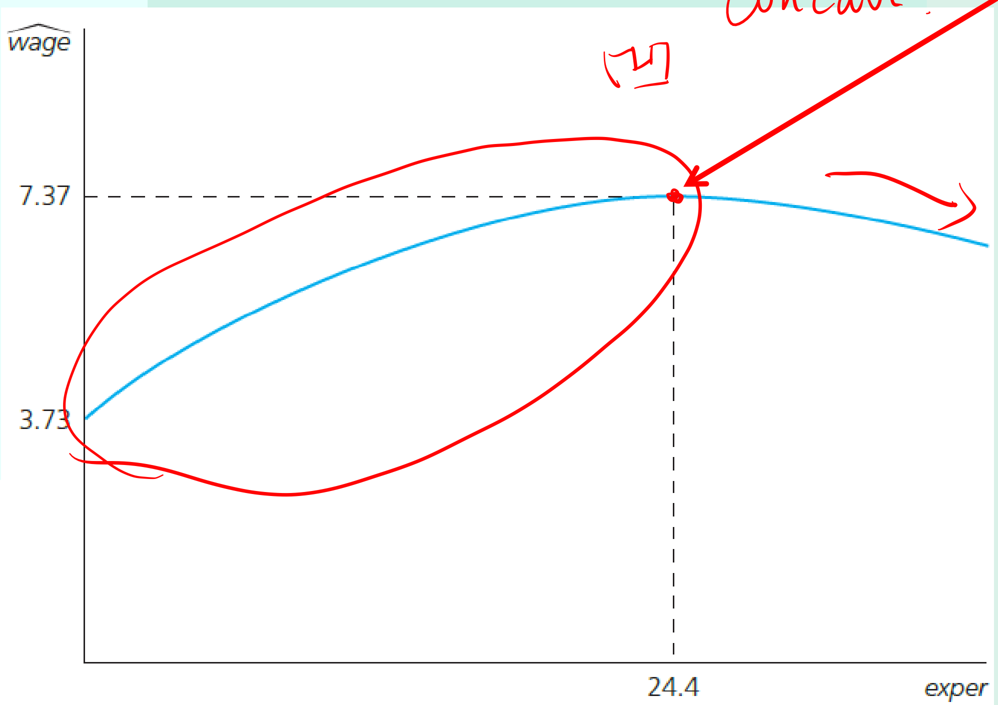
$$\frac{\Delta wage}{\Delta exper} = .298 - 2(.0061)exper = 0$$

对函数形式的进一步讨论

- **工作经验的最高工资**

concave

凹



$\frac{0.298}{2\cdot(0.0061)}$

$$x^* = \left|\frac{\hat{\beta}_1}{2\hat{\beta}_2}\right| = \left|\frac{.298}{2(.0061)}\right| \approx 24.4$$

二次函数的代价：转折点

是否意味着24.4年后，经验的回报为负？
需要考虑以下几个情况：

（1）取决于样本中有多少观测值位于转折点的右侧。如果很少则不需考虑（在给定的例子中，大约28%的观察值在转折点右侧）。

（2）是否存在模型设定问题（例如忽略变量，如年龄）。

对函数形式的进一步讨论

- **例子：污染对房价的影响**

空气中的氮氧化物、与就业中心的距离、
平均学生/教师比率

$$\widehat{\log}(price) = \underset{(.57)}{13.39} - \underset{(.115)}{.902}\ \log(nox) - \underset{(.043)}{.087}\ \log(dist)$$

+ 0.22

$$- \underset{(.165)}{.545}\ rooms + \underset{(.013)}{\boxed{.062}}\ rooms^2 - \underset{(.006)}{.048}\ stratio$$

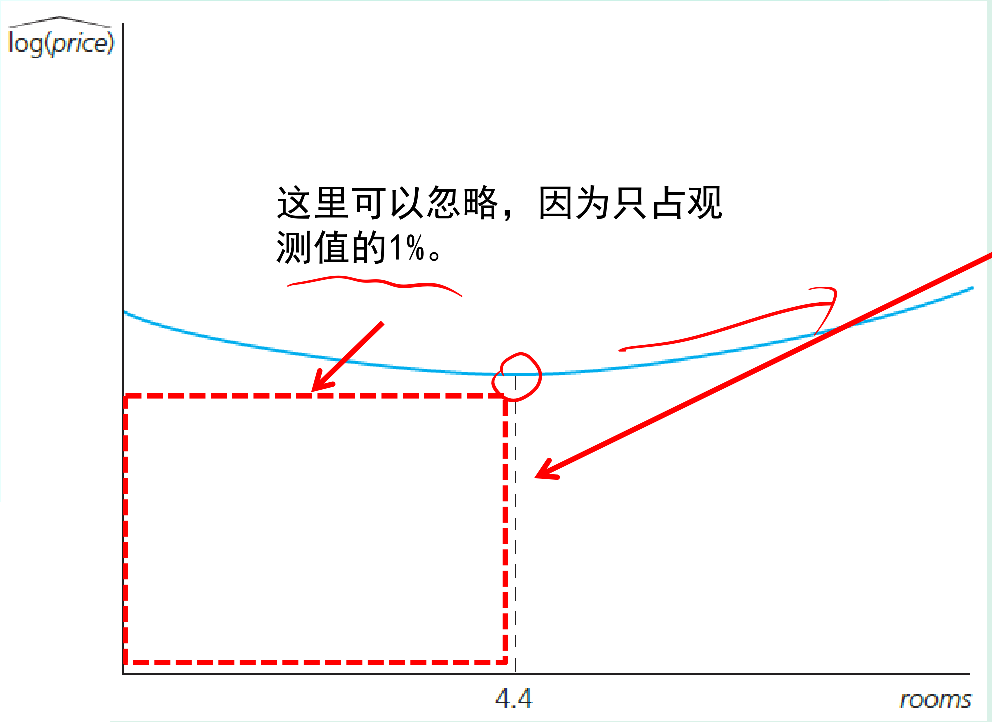
$$n = 506, R^2 = .603$$

这是否意味着，在房间数量较少的情况下，更多
的房间意味着更低的价格？（少于4间的只有1%）

$$\Rightarrow \frac{\Delta \log(price)}{\Delta rooms} = \frac{\%\Delta price}{\Delta rooms} = -.545 + .124 rooms$$

对函数形式的进一步讨论

- **计算转折点**



转折点：

$$x^* = \left|\frac{-.545}{2(.062)}\right| \approx 4.4$$

房间从5增加到6：

$$-.545 + .124(5) = +7.5\%\ price$$

房间从6增加到7：

$$-.545 + .124(6) = +19.9\%\ price$$

对函数形式的进一步讨论

- **其他可能性**

$$\log(price) = \beta_0 + \beta_1 \log(nox) + \beta_2 [\log(nox)]^2$$

$$+\beta_3 crime + \beta_4 rooms + \beta_5 rooms^2 + \beta_6 stratio + u$$

$$\Rightarrow \quad \frac{\Delta \log(price)}{\Delta \log(nox)} = \frac{\%\Delta price}{\%\Delta nox} = \beta_1 + 2\beta_2[\log(nox)]$$

- **更高阶的多项式：总成本函数**

$$cost = \beta_0 + \beta_1 quantity + \beta_2 quantity^2 + \beta_3 quantity^3 + u$$

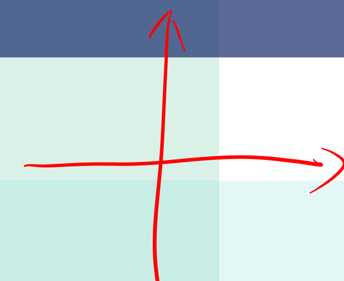
对函数形式的进一步讨论

- **含有交互项的模型**

面板中.

DID，双重差分，

准实验



$$price = \beta_0 + \beta_1 sqrft + \boxed{\beta_2} bdrms$$

交互项（interaction）

post：时间上已经开始后取1.

$$+ \beta_3 \boxed{sqrft \cdot bdrms} + \beta_4 bthrms + u$$

一组：对照

一组：实验.

treat：实验组取1

$$\Rightarrow \frac{\Delta price}{\Delta bdrms} = \beta_2 + \beta_3 sqrft$$

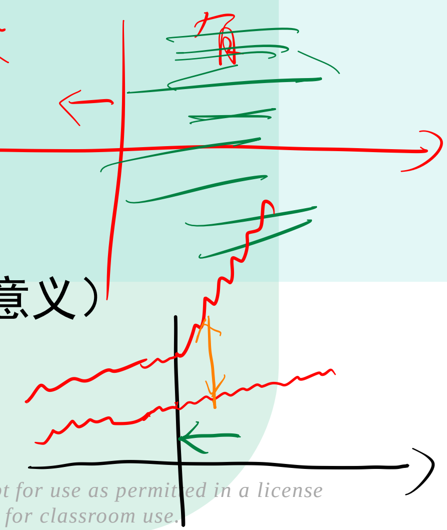
卧室数量的影响取决于建筑面积的大小（有意义）

对照组取0.

$y_{it} = \dots + \beta_1 \cdot treat + \beta_2 \cdot post + \beta_3 \cdot treat \cdot post$.

没

有



- **相互作用效应使参数的解释复杂化**

β_2 = bdrms对一套面积为零的住房的价格的影响。（没意义）

treat：1，0　1

post：0，1　1

平行趋势

对函数形式的进一步讨论

- **例子：大学前GPA, ACT成绩和出勤率对期末考试分数的影响**

出勤率的影响

$$\widehat{stndfnl} = \underset{(1.36)}{2.05} - \underset{(.0102)}{.0067}\, atndrte - \underset{(.48)}{1.63}\, priGPA - \underset{(.098)}{.128}\, ACT$$
$$+ \underset{(.101)}{.296}\, priGPA^2 + \underset{(.0022)}{.0045}\, ACT^2 + \underset{(.0043)}{.0056}\, priGPA \cdot atndrte$$
$$n = 680,\ R^2 = .229,\ \bar{R}^2 = .222.$$

- **相互作用效应使参数的解释复杂化**

-0.0067＝priGPA为零时atndrte对期末分数的影响。（没意义）

对函数形式的进一步讨论

- **交互效应的再参数化**

断点回归

总体均值，由样本均值代替

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + u$$

$$y = \alpha_0 + \delta_1 x_1 + \delta_2 x_2 + \beta_3 (x_1 - \boxed{\mu_1})(x_2 - \boxed{\mu_2}) + u$$

如果所有变量取其均值，x_2的影响

- **再参数化的优势**
  - 参数的含义简单化
  - 对于在均值处的偏效应的标准误可以计算

对函数形式的进一步讨论

- 在具有二次函数、交互作用和其他非线性函数形式的模型中，偏效应取决于一个或多个解释变量的值
- 例子：上个例子中atndrte对期末平均分数的偏效应：

$$-0.0067 + 0.0056priGPA$$

- **平均偏效应（average partial effects，APE）**是描述因变量和每个解释变量之间关系的一种测度值
- 计算偏效应并带入估计值后，对样本中每个元素的偏效应求平均值
- 例子：上个例子中atndrte的平均偏效应：

$$APE_{atndrte} = -0.0067 + 0.0056\overline{priGPA}$$

思考：priGPA的APE是多少？ $\overline{atndrte}$

拟合优度和回归元选择的进一步探讨

- **拟合优度和回归元选择的进一步探讨**
- **对R2的一般性评论**
  - 高R2不意味着因果关系
  - 低R2并不妨碍精确估计偏效应
  - 普通R2测量了什么?

R^2, MSE

$$R^2 = 1 - \frac{SSR}{SST} = 1 - \frac{(SSR/n)}{(SST/n)}$$ 是一个对于 $\underbrace{\boxed{1 - \frac{\sigma_u^2}{\sigma_y^2}}}_{\text{总体R2}}$ 的估计

有截距

  - 引入新的解释变量必然会增加R2, 即使该变量对被解释变量没有因果影响 不好

拟合优度和回归元选择的进一步探讨

- **调整R2**

分子和分母的自由度

  - 一个更好的考虑自由度的估计量是
  - $$\bar{R}^2 = 1 - \frac{(SSR/\boxed{(n-k-1)})}{(SST/\boxed{(n-1)})} = adjusted\ R^2$$

R^2

adj. R^2

  - 调整R2对于增加新的自变量进行惩罚（只有加入的新变量可以“明显地”降低SSR时，调整R2才会增加）
  - 数学上可以证明：当且仅当新变量的t统计量的绝对值大于1，调整R2才会增加

- **R2与调整R2的关系**

$n-k-1$ 小

调整R2可能为负

$$\bar{R}^2 = 1 - (1 - R^2)(n-1)/(n-k-1)$$

有可能负

拟合优度和回归元选择的进一步探讨

- **利用调整R2在两个非嵌套模型中进行选择**
  - 如果任一模型都不是其他模型的特例，则模型之间是非嵌套的

$$rdintens = \beta_0 + \beta_1 \log(sales) + u$$

$$\boxed{R^2 = .061, \bar{R}^2 = .030}$$

$$rdintens = \beta_0 + \beta_1 sales + \beta_2 sales^2 + u$$

$$\boxed{R^2 = .148, \bar{R}^2 = .090}$$

  - 比较两个模型的R2对第一个模型不公平，因为第一个模型包含的参数较少
  - 在给定的例子中，即使调整了自由度的差异，二次模型仍然是首选

拟合优度和回归元选择的进一步探讨

- **具有不同因变量模型的比较**
  - R2或调整R2不能用于比较具有不同因变量的模型

- **例子：CEO薪酬与企业业绩**

log(salary)比
salary的波动性
要小得多

$$\boxed{\widehat{salary}} = 830.63 + .0163\, sales + 19.03\, roe$$

$$(223.90) \quad (.0089) \qquad\qquad (11.08)$$

$$n = 209,\ R^2 = .029,\ \bar{R}^2 = .020,\ SST = 391{,}732{,}982$$

$$\boxed{\widehat{lsalary}} = \underset{(0.29)}{4.36} + \underset{(.033)}{.275}\, lsales + \underset{(.0040)}{.0179}\, roe$$

$$n = 209,\ R^2 = .282,\ \bar{R}^2 = .275,\ SST = 66.72$$

拟合优度和回归元选择的进一步探讨

- **为了避免遗漏变量偏误，有时回归分析中控制因素过多**
- **在某些情况下，一些变量不应被控制**
  - 在啤酒税（和其他因素）导致的交通事故死亡率回归中，不应直接控制啤酒消费：

$$fatalities = \beta_0 + \boxed{\beta_1} tax + \beta_2 beercons + \cdots$$

  - 在农民家庭健康对农药使用的回归中，不应控制看病次数
- **不同的回归可能有不同的目的**
  - 在房价对房屋特征回归中，如果想研究价格评估的有效性，就应该包括房价评估；
  - 但如果只是探讨房屋特征对房价的影响，会发现不应包含房价评估（控制房价评估价值不变，讨论增加一间卧室对房屋价值的影响没有意义）

拟合优度和回归元选择的进一步探讨

- **增加自变量以减少误差方差**
  - 增加自变量可能会增加多重共线性
  - 另一方面，增加自变量可以减少误差方差（残差的方差）
  - 应添加与其他自变量不相关的变量，因为它们在不增加多重共线性的情况下减少了误差方差
  - 然而，很难找

$\hat{u}$

R^2 残差平方和

- **例子：个人啤酒消费与啤酒价格**
  - 将个体特征纳入啤酒消费对啤酒价格的回归中，可以更精确地估计价格弹性

y

	(1)	(2)	(3)
x_1	0.51 (0.07)	0.54 (0.015)	0.53 (0.002)
x_2	—	0.8 (0.5)	0.75 (0.6)
x_3		—	0.06 (0.004)
x_4	—	—	0.08 (0.006)

u

预测

- 在第三章中，我们定义了OLS预测值或拟合值： Reduced, 解释.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + +\hat{\beta}_k x_k$$ Structrue

具体的，给定$x_1 = c_1, \ldots, x_k = c_k$，我们得到预测值：

$$\hat{\theta} = \hat{\beta}_0 + \hat{\beta}_1 c_1 + \cdots + +\hat{\beta}_k c_k$$

$\hat{\theta}$作为OLS估计值的线性组合，存在抽样波动的问题，如何得到置信区间？

$\hat{\theta}$可以视为预期$\theta_0 = \beta_0 + \beta_1 c_1 + \cdots + +\beta_k c_k$的估计，那么我们可以变换模型得到

$$y = \theta_0 + \beta_1(x_1 - c_1) + \cdots + \beta_k(x_k - c_k) + u$$

误差

预测

- 例子：大学GPA预测值的置信区间

$$
\begin{aligned}
\widehat{colgpa} = &\ 1.493 + .00149\, sat - .01386\, hsperc \\
&\ (0.075)\ (.00007)\qquad (.00056) \\
&- .06088\, hsize + .00546\, hsize^2 \\
&\ (.01650)\qquad\quad (.00227) \\
n = 4{,}137,\ R^2 &= .278,\ \bar{R}^2 = .277,\ \hat{\sigma} = .560,
\end{aligned}
$$

- 当$sat = 1200, hsperc = 30, hsize = 5$时，$colgpa$预测值为2.70

预测

- 例子：大学GPA预测值的置信区间
- 模型变换：

$$sat0 = sat - 1200, hsperc0 = hsperc - 30, hsize0 = hsize - 5, hsizesq0 = hsize^2 - 25$$

- 估计结果：

$$\begin{aligned}\widehat{colgpa} = \boxed{2.700} &+ .00149\ sat0 - .01386\ hsperc0 \\ (0.020)\ &\quad (.00007) \qquad\quad (.00056) \\ &- .06088\ hsize0 + .00546\ hsizesq0 \\ &\quad (.01650) \qquad\quad (.00227) \\ n = 4{,}137,\ R^2 &= .278,\ \bar{R}^2 = .277,\ \hat{\sigma} = .560.\end{aligned}$$

- 预期gpa的95%置信区间：$2.70 \pm 0.02 \times 1.96 = [2.66, 2.74]$

预测

- 当因变量为log(y)时对y的预测

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

- 给定OLS估计量，首先可以预测$\log y$

$$\widehat{\log y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + +\hat{\beta}_k x_k$$

- 之后，我们可以得到 y 的预测值：

$$\hat{y} = exp\left(\widehat{\log y}\right)$$

- 然而，这将系统性地低估y 的预测值

预测

- **当因变量为log(y)时对y的预测**

$u \sim N(\cancel{\mu}\,0, \sigma^2)$

$E[\exp(u)] = \exp(\cancel{\mu +} \boxed{\tfrac{1}{2}\sigma^2})$

Moment Generating function

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

$$\Rightarrow \quad y = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)\boxed{\exp(u)} \quad e^{\frac{1}{2}\sigma^2}$$

当假设 u 独立于 $x_1, \ldots, x_k$ 时:

$E[\exp(u)] \neq 1$.

$$\Rightarrow \quad E(y|\mathrm{x}) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k) E(\exp(u))$$

$$\Rightarrow \quad \hat{y} = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k)\left(\frac{1}{n}\sum_{i=1}^{n}\exp(\hat{u}_i)\right)$$

对y的预测

预测

M 95% 提出来

train, validation, test

训练 验证集 测试、

K-fold validation.

K

估计 X_{Train} X_{test}

训练 ~~测~~试集 验证.

Model 1 Model 1, $\hat{y}_1$ 85%

Model 2 2, $\hat{y}_2$ 87%

Model M $\hat{y}_K$ 97%

y真实已知. 98%

Selection bias.

M

小节

- 本章讨论了多元回归分析的一些重要专题
- 改变自变量或者因变量的单位，对统计量没有影响
- 对数、二次项、交互项的使用会对OLS估计值的解释产生影响
- 调整R2有些时候可以用来选择回归元
- 我们提出了预测值置信区间的构造方法并讨论了log(y)的预测问题