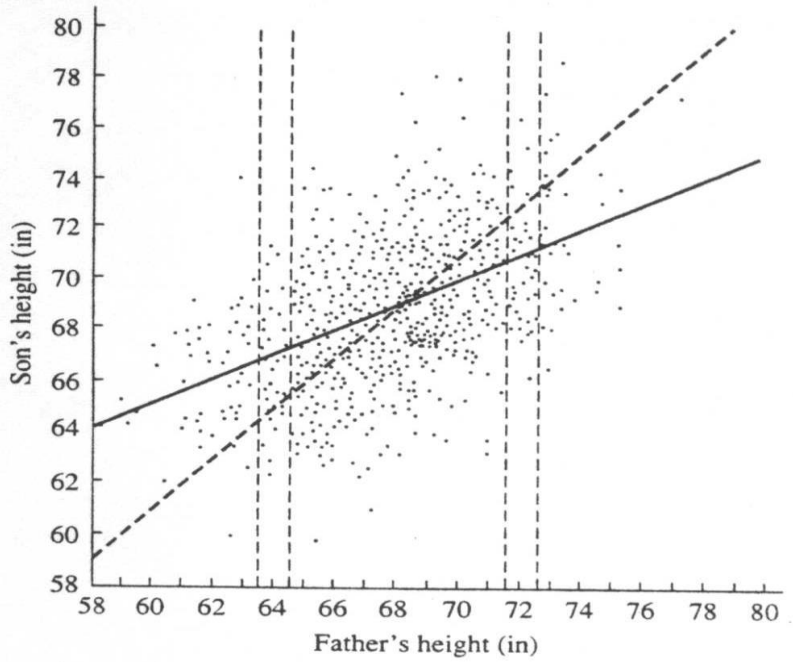
应用统计

第12讲 一元回归模型

高尔顿的身高回归直线

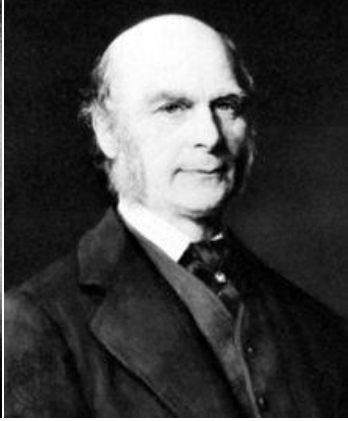
高（矮）个子的父母的子女的身高比其父母矮（高），子女身高有向中心回归的趋势。这也是“回归”一词的来源。



回归直线 $y=0.516\,x+33.73$

Pearson: 1078个父亲和儿子身高的散点图

Galton（高尔顿）



高尔顿（右）是达尔文（左）的表弟，他的父亲是银行家。他从小就聪颖过人，出生12个月后，他便能认识所有的大写字母，18个月后则能辨别大写和小写两种字母。到了两岁半左右，高尔顿已能阅读《蛛网捕蝇》之类的儿童读物。3岁时他学会签名，4岁时他能写诗，5岁时已能背诵并理解苏格兰叙事诗《马米翁》，6岁时，他已精熟荷马史诗，7岁能欣赏莎士比亚名著，对博物学产生兴趣，并按自己的方法对昆虫、矿物标本进行分类。

弗朗西斯·高尔顿 [Francis Galton 1822 – 1911]，英国探险家、优生学家、统计学家、心理学家，也是心理测量学上生理计量法的创始人。1846 – 1850 年参加了皇家地理学会去非洲的旅行，从事气象学和地理学的研究。1856 年当选为皇家学会会员。1859 年表兄达尔文出版了《物种起源》引起了他对人类遗传的兴趣。1884 年，他创设了人类测量实验室，1901 年与其弟子皮尔逊创办了《生物统计学杂志》，1904 年捐赠基金在伦敦创办优生学实验室。

（线性）回归分析

- 回归分析是研究变量间关系的统计学课题
- 在数据的定量分析中，往往需要处理存在着一定联系的变量，需要刻画变量之间有怎样的相互关系，以及如何发生相互影响
- 一元线性回归分析、多元线性回归分析、非线性回归分析、曲线估计、时间序列分析，以及逻辑斯蒂克回归等

（线性）相关系数

不相关不仅不对应不独立，而且还可能具有很强的相互关联。

所以不相关指的是不存在线性相关的关系，相关系数不能表达非线性的相关关系。

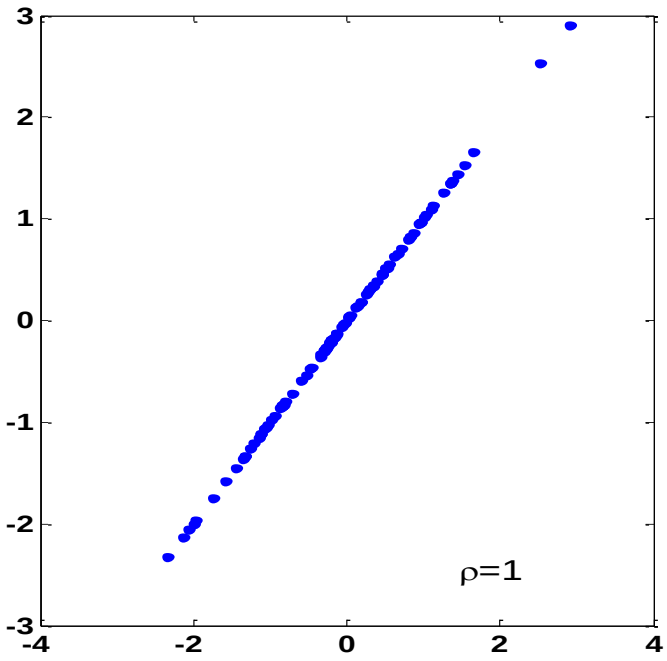
例如：$X \sim N(0,1)$，$Y = X^2$，显然 X, Y 不独立 （思考：对不独立直观上的理解），

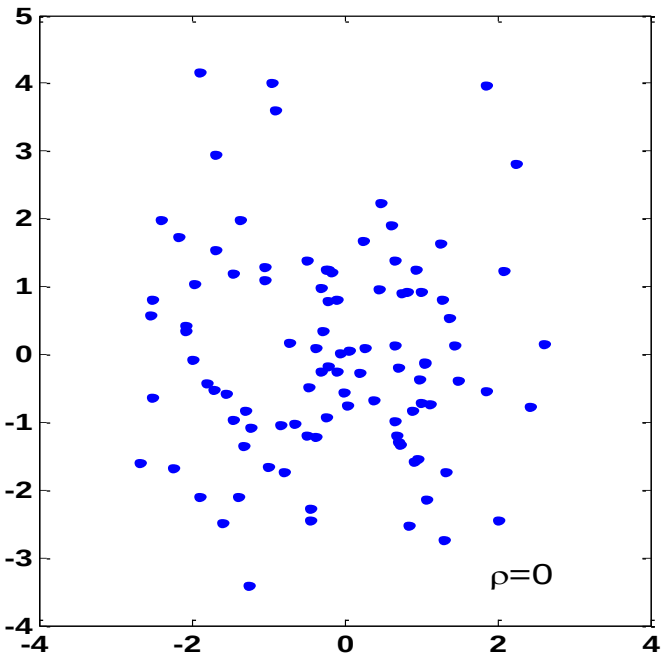
$Cov(X,Y) = Cov(X,X^2) = E(X^3) - E(X)E(X^2) = 0$，$X, Y$（线性）不相关。

$Corr(X,Y) = \pm 1$ 的充分必要件是 X, Y 之间几乎处处有线性关系，

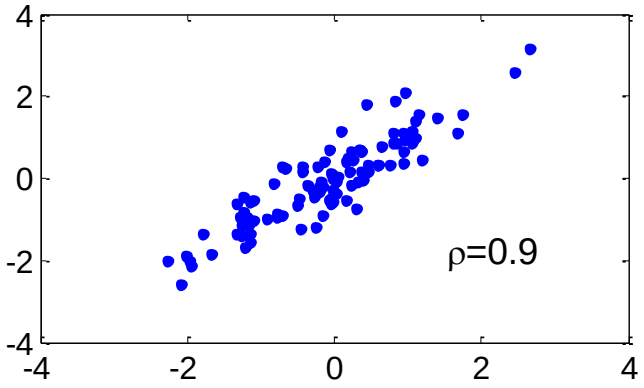
即存在常数 a、b, 使得 $P(Y = aX + b) = 1$

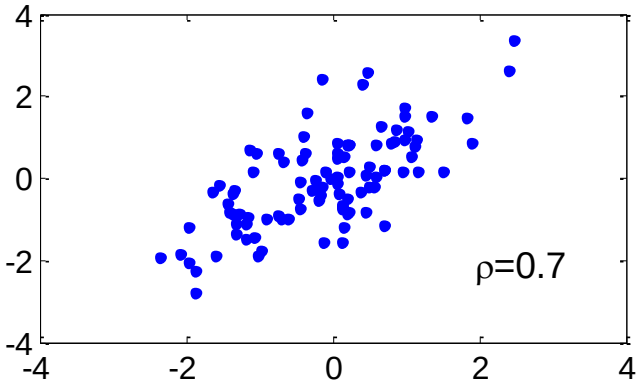
两个随机变量的相关性

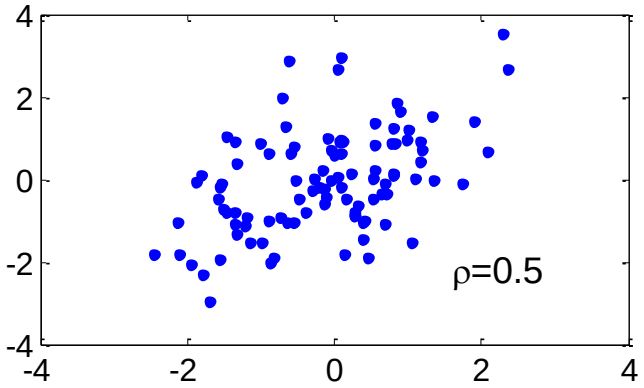


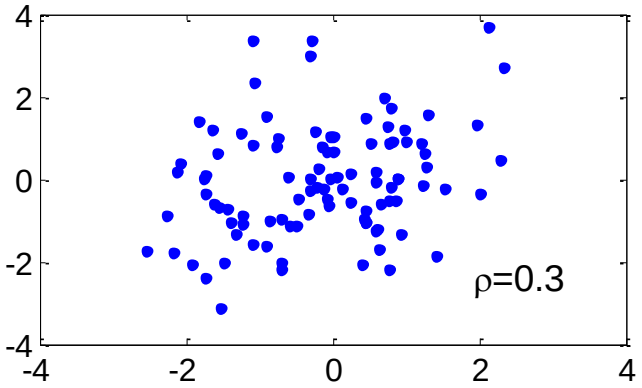


相关系数的大小	一般解释
0.8 ~ 1.0	非常强的相关
0.6 ~ 0.8	强相关
0.4 ~ 0.6	中度相关
0.2 ~ 0.4	弱相关
0.0 ~ 0.2	弱相关或无关









早期线性回归的使用实例

例 1.1　Forbes 数据

在十九世纪四、五十年代，苏格兰物理学家 James D. Forbes，试图通过水的沸点来估计海拔高度。他知道通过气压计测得的大气压可用于得到海拔高度，高度越高，气压越低。在这里讨论的实验中，他研究了气压和沸点之间的关系。由于在 40 年代运输精密的气压计相当困难，这引起了他研究此问题的兴趣。测量沸点将给旅行者提供一个快速估计高度的方法。

Forbes 在阿尔卑斯山及苏格兰收集数据。选定地点后，他装起仪器，测量气压及沸点。气压单位采用水银柱高度，并根据测量时周围气温与标准气温之间的差异校准气压。沸点用华氏温度表示。我们从他 1857 年的论文中选取了 $n=17$ 个地方的数据，见表 1.1（Forbes，1857）。在研究这些数据时，有若干可能引起兴趣的问题，气压及沸点是如何联系的？这种关系是强是弱？我们能否根据温度预测气压？如果能，有效性如何？

表 1.1　在阿尔卑斯山及苏格兰的 17 个地方沸点(°F)及大气压（英寸汞柱）的 Forbes 数据

案例号	沸　点（°F）	气　压（英寸汞柱）	log（气压）	100×log（气压）
1	194.5	20.79	1.3179	131.79
2	194.3	20.79	1.3179	131.79
3	197.9	22.40	1.3502	135.02
4	198.4	22.67	1.3555	135.55
5	199.4	23.15	1.3646	136.46
6	199.9	23.35	1.3683	136.83
7	200.9	23.89	1.3782	137.82
8	201.1	23.99	1.3800	138.00
9	201.4	24.02	1.3806	138.06
10	201.3	24.01	1.3805	138.05
11	203.6	25.14	1.4004	140.04
12	204.6	26.57	1.4244	142.44
13	209.5	28.49	1.4547	145.47
14	208.6	27.76	1.4434	144.34
15	210.7	29.04	1.4630	146.30
16	211.9	29.88	1.4754	147.54
17	212.2	30.06	1.4780	147.80

Forbes 的理论认为，在观测值范围内，沸点和气压值的对数成一直线。由此，我们取 10 作为对数的底数。事实上统计分析和对数的底数是没有关系的。由于气压的对数值变化不大，最小的为 1.318，而最大的为 1.478，我们将所有气压的对数值乘以 100，如表 1.1 中第 5 列所示。这将在不改变分析的主要性质的同时，避免研究非常小的数字。

着手进行回归分析的一个有效途径是，画一个变量对另一个变量的图。这图称为散点图，它既能用于提示某种关系，也能用于说明这种关系可能是不适当的。散点图可手工在一般作图纸上绘制。X 轴即水平轴，通常留作用于自变量。在 Forbes 的数据中为沸点。Y 轴即垂直轴，通常被用于表示响应变量。在本例中，Y 轴的值为 $100\times\log$（气压）。对 n 对 (x, y) 数据中的每一对，在图上作一个点。大多数回归分析的计算机程序可以作这个图。

Forbes 数据的散点图的总的印象是，这些点基本上，但并不精确地，落在一条直线上。图 1.1 所画的直线将在后面讨论。它指出两个变量之间的关系至少可以初步近似地用一条直线的方程来描述。

Forbes. J. D. (1857). "Further experiments and remarks on the measurement of heights by the boiling point of water." *Trans. R. Soc. Edinburgh*, **21**, 135—143.

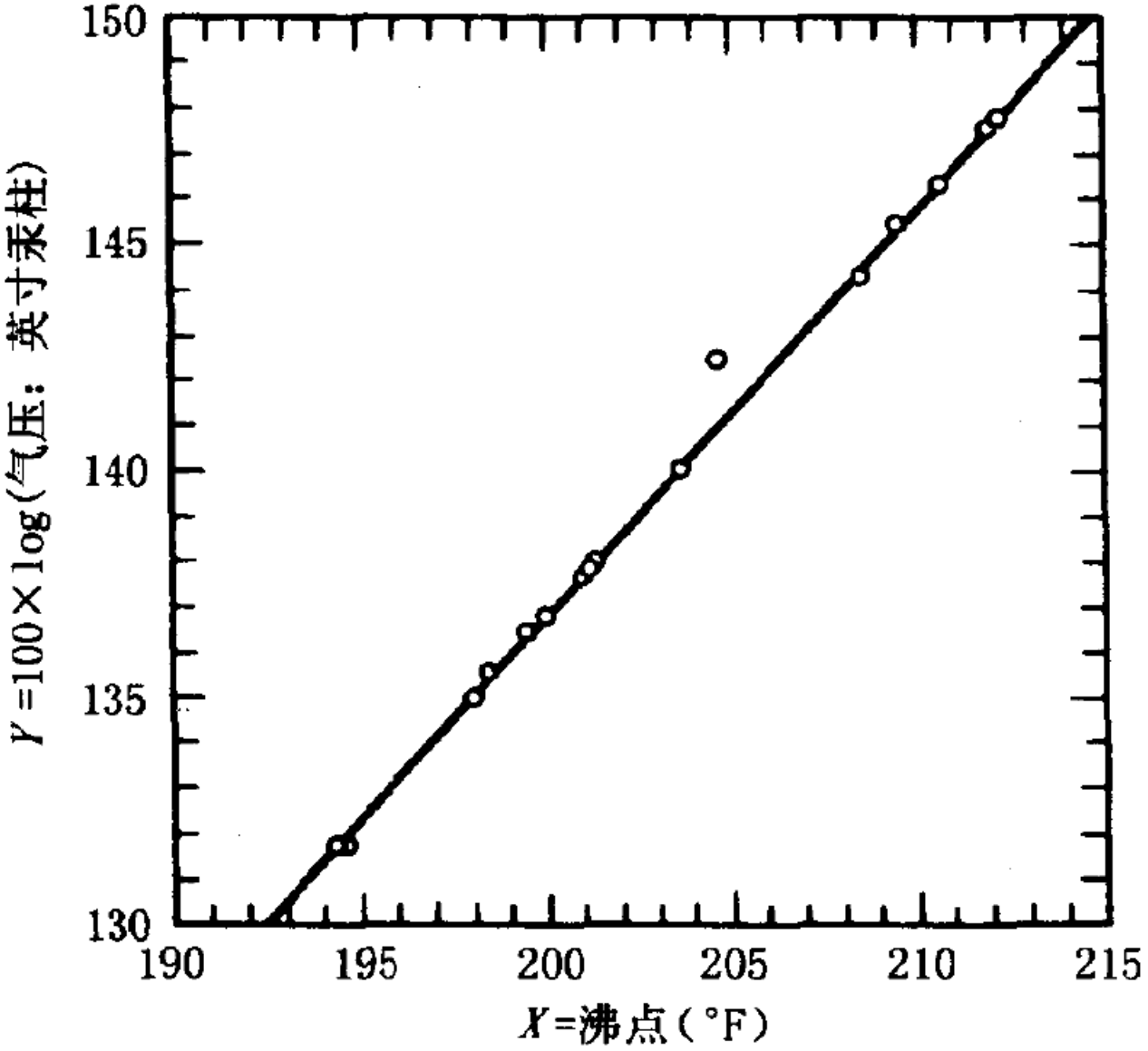


图 1.1　Forbes 数据的散点图

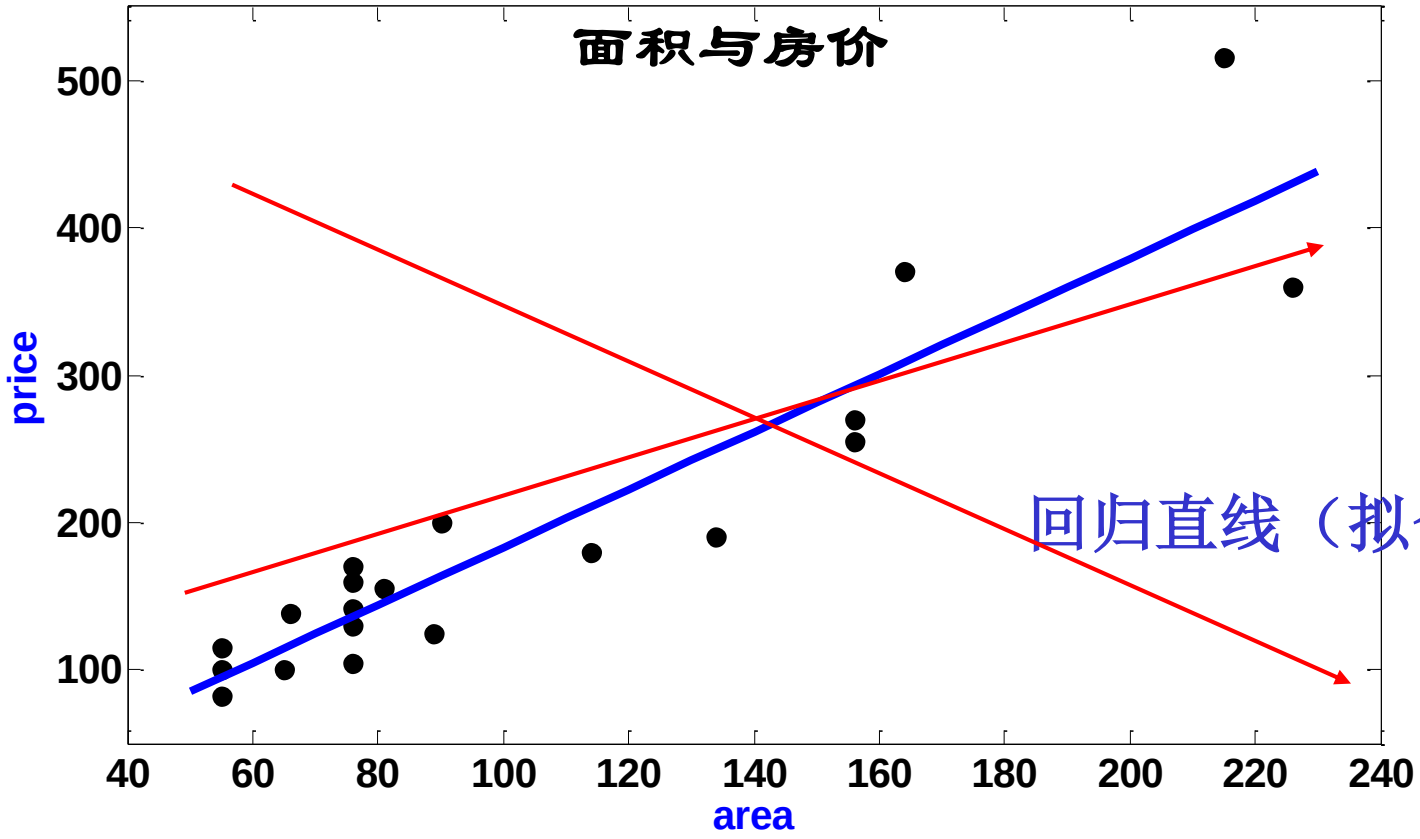
回归分析实例--北京房价分析

编号	面积（平方米）	价格(万元)
1	55	100
2	76	130
3	65	100
4	156	255
5	55	82
6	76	105
7	89	125
8	226	360
9	134	190
10	156	270

编号	面积（平方米）	价格(万元)
11	114	180
12	76	142
13	164	370
14	55	115
15	90	200
16	81	155
17	215	516
18	76	160
19	66	138
20	76	170

自变量（预报因子），　因变量（预报对象）

北京房价的回归分析



学会如何得到回归直线，并掌握其统计含义

SPSS运行输出结果

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	area[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: price

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.926[a]	.857	.849	43.12927

a. Predictors: (Constant), area

SPSS运行输出结果

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	200792.145	1	200792.145	107.945	.000[a]
	Residual	33482.405	18	1860.134		
	Total	234274.550	19			

a. Predictors: (Constant), area

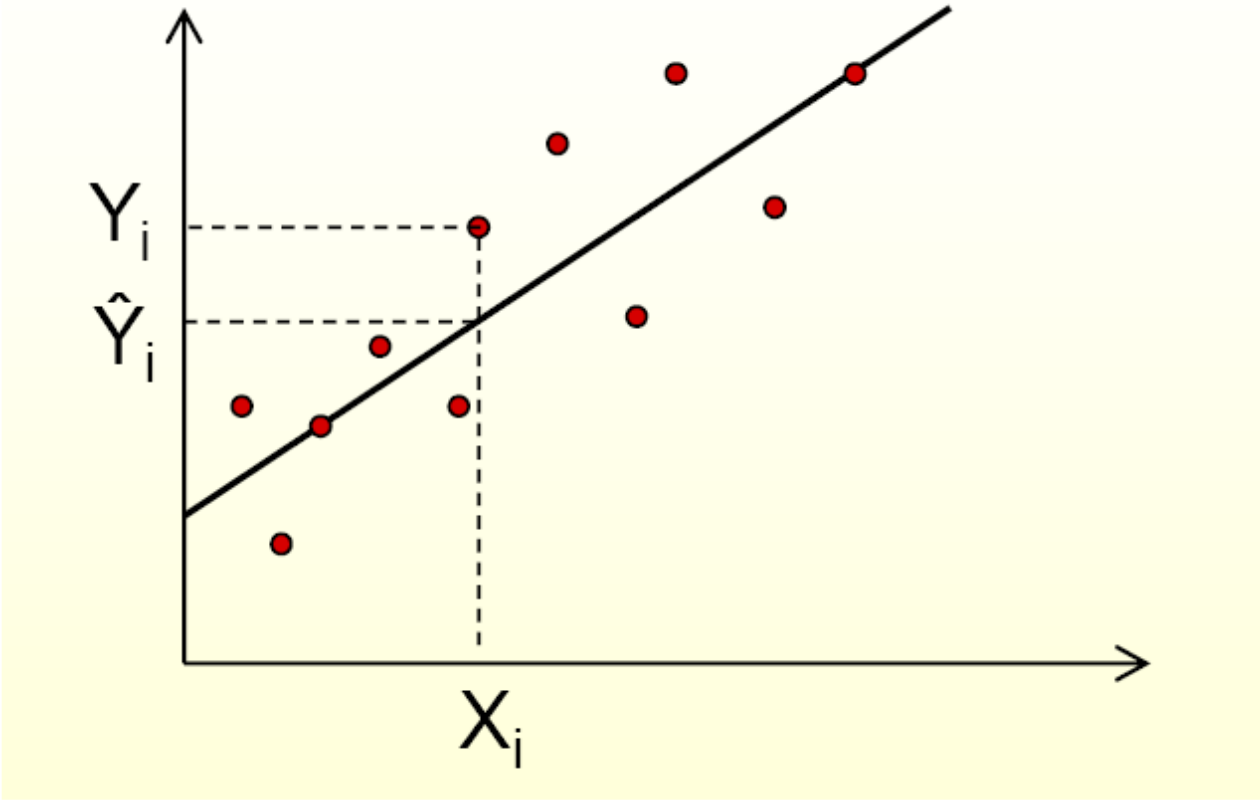
b. Dependent Variable: price

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-12.825	22.046		-.582	.568
	area	1.961	.189	.926	10.390	.000

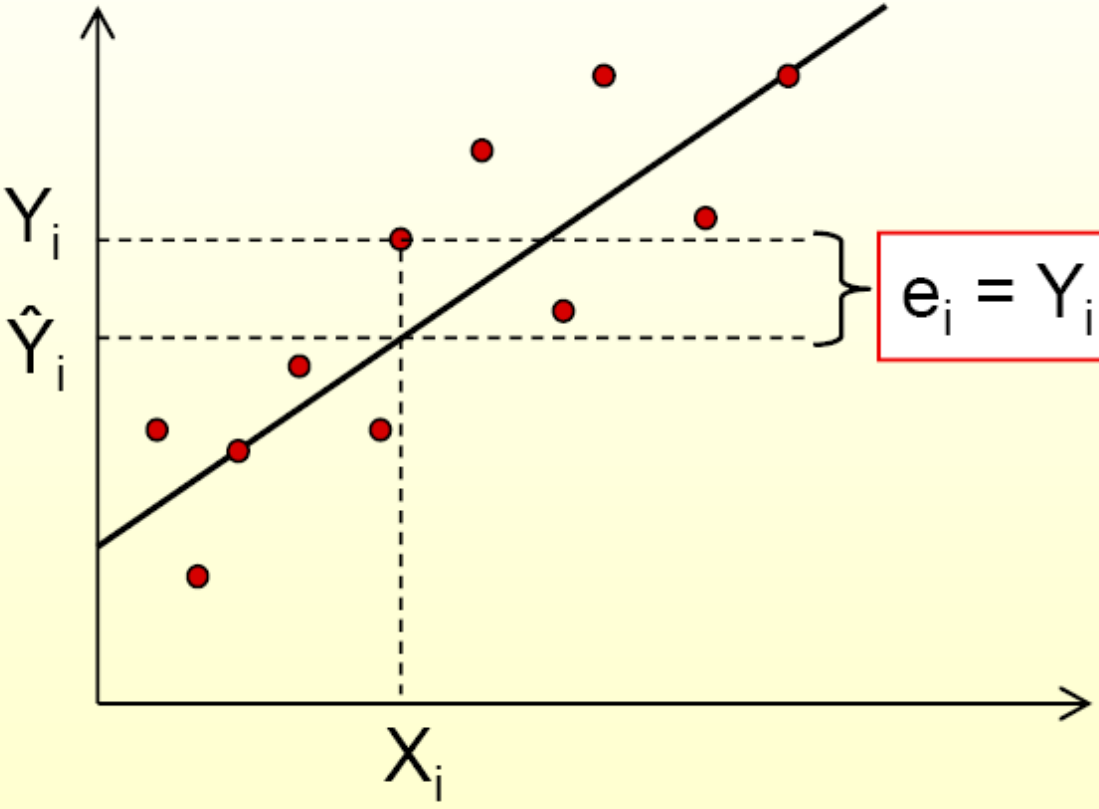
a. Dependent Variable: price

线性拟合与预报（拟合值）



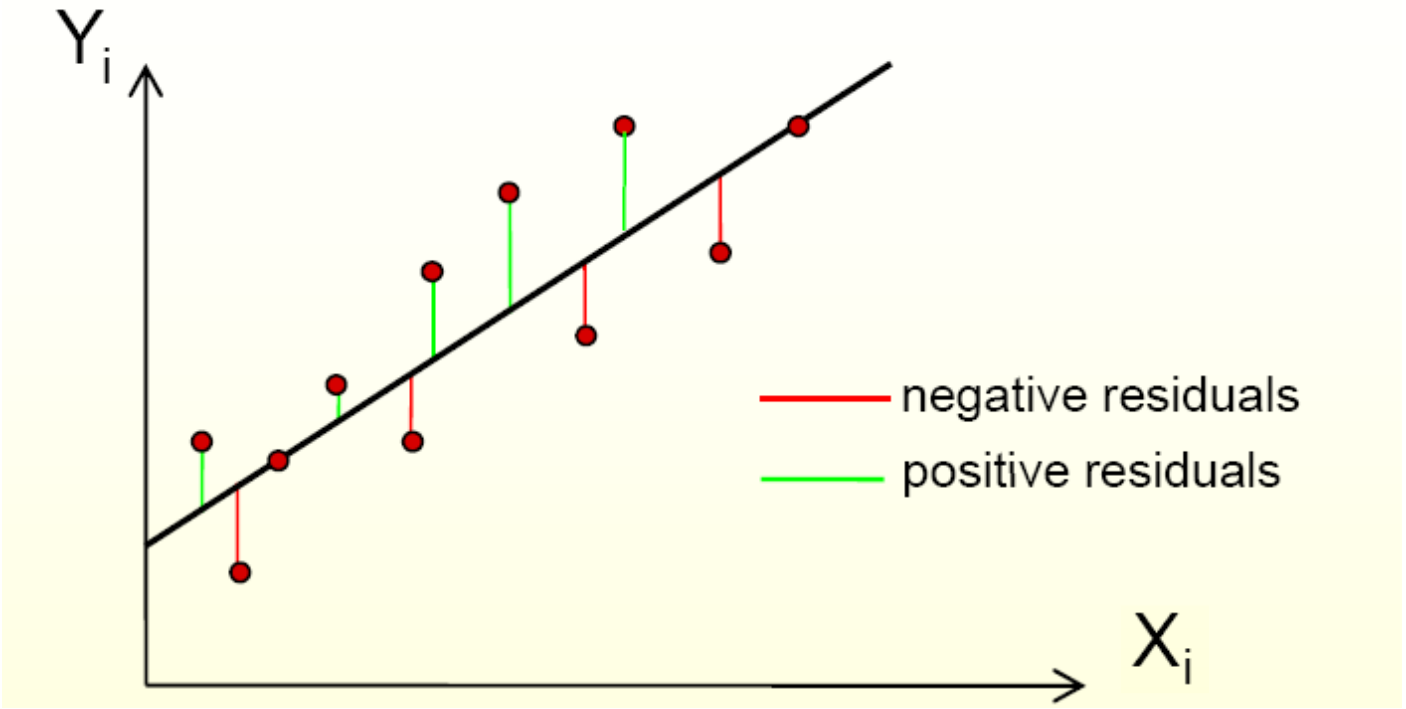
拟合方程：$\hat{Y}_i = b_0 + b_1 X_i$

第i个观测值的残差



$$e_i = Y_i - \hat{Y}_i = \text{Residual}$$

拟合值与残差



$$Y_i = \hat{Y}_i + (Y_i - \hat{Y}_i) = \hat{Y}_i + e_i$$

最小二乘拟合

- 最小二乘：
- 选择 b_0 和 b_1 使得残差的平方和最小

Least Squares:

choose b_0 and b_1 to minimize $\sum_{i=1}^{N} e_i^2$

$$\min \sum_{i=1}^{n} \left(y_i - \left(b_0 + b_1 x_i\right)\right)^2$$

最小二乘拟合公式

$$S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$L_{XX} = \sum_{i=1}^{n}(x_i-\bar{x})^2$$

$$L_{YY} = \sum_{i=1}^{n}(y_i-\bar{y})^2$$

$$L_{XY} = \sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$$b_1 = \frac{\sum_{i=1}^{N}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{N}(X_i-\bar{X})^2} = \frac{S_{XY}}{S_X^2}$$

S_{XY} = sample covariance (X, Y) 样本协方差

S_X^2 = sample variance of (X) 样本方差

$$b_0 = \bar{Y} - b_1\bar{X}$$

$$b_1 = \frac{S_{xy}}{S_x^2} = r_{xy} \times \frac{S_y}{S_x}$$

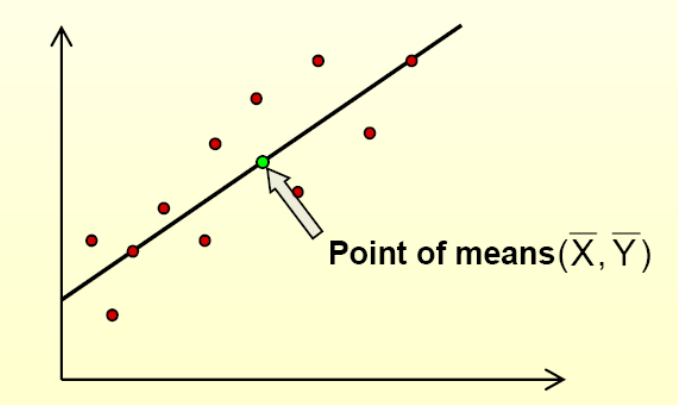
b_0（截距）与b_1（斜率）的直观解释

- b_0（截距）的取值保证了回归直线必然经过 $(\overline{X}, \overline{Y})$

$$\hat{y}(x) = (\bar{y} - b_1 \bar{x}) + b_1 x = \bar{y} + b_1 (x - \bar{x})$$

$$\hat{Y} - \overline{Y} = b_1 (X - \overline{X})$$

利用回归直线过样本均值



- b_1（斜率）的取值使得

残差与预报变量不相关 **Corr (e,X) =0**

$$\sum (X_i - \bar{X})(Y_i - b_0 - b_1 X_i) = \sum (X_i - \bar{X})(Y_i - [\bar{Y} - b_1 \bar{X}] - b_1 X_i)$$

$$= \sum (X_i - \bar{X})(Y_i - \bar{Y} - b_1 (X_i - \bar{X})) = \sum (X_i - \bar{X})(Y_i - \bar{Y}) - b_1 \sum (X_i - \bar{X})^2 = 0$$

方差的分解，ANOVA Table

- **ANOVA （analysis of variance）**

$$\underbrace{\sum_{i=1}^{N}(Y_i-\overline{Y})^2}_{\text{Total Sum of Squares SST}} = \underbrace{\sum_{i=1}^{N}(\hat{Y}_i-\overline{Y})^2}_{\text{Regression SS SSR}} + \underbrace{\sum_{i=1}^{N}e_i^{\,2}}_{\text{Error SS SSE}}$$

总偏差平方和：**SST**； 回归平方和：**SSR**； 剩余平方和：**SSE**

拟合优度的度量：R^2

- 好的拟合：

  回归平方和 SSR大；剩余平方和 SSE小

  如果SST=SSR，则得到完美的拟合

- 决定系数（coefficient of determination）

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

但是不易进行建立一个量化标准

Y（响应变量）可以在多大的程度（百分比）被X（预测变量）解释

$$Y = b_0 + b_1 X + e$$

简单线性回归的理论模型

$$Y = \underbrace{\beta_0 + \beta_1 X}_{\text{Part of Y related to X}} + \underbrace{\varepsilon}_{\text{Part of Y independent of X}}$$

$$\varepsilon \sim N(0, \sigma^2) \quad \text{Error Term}$$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

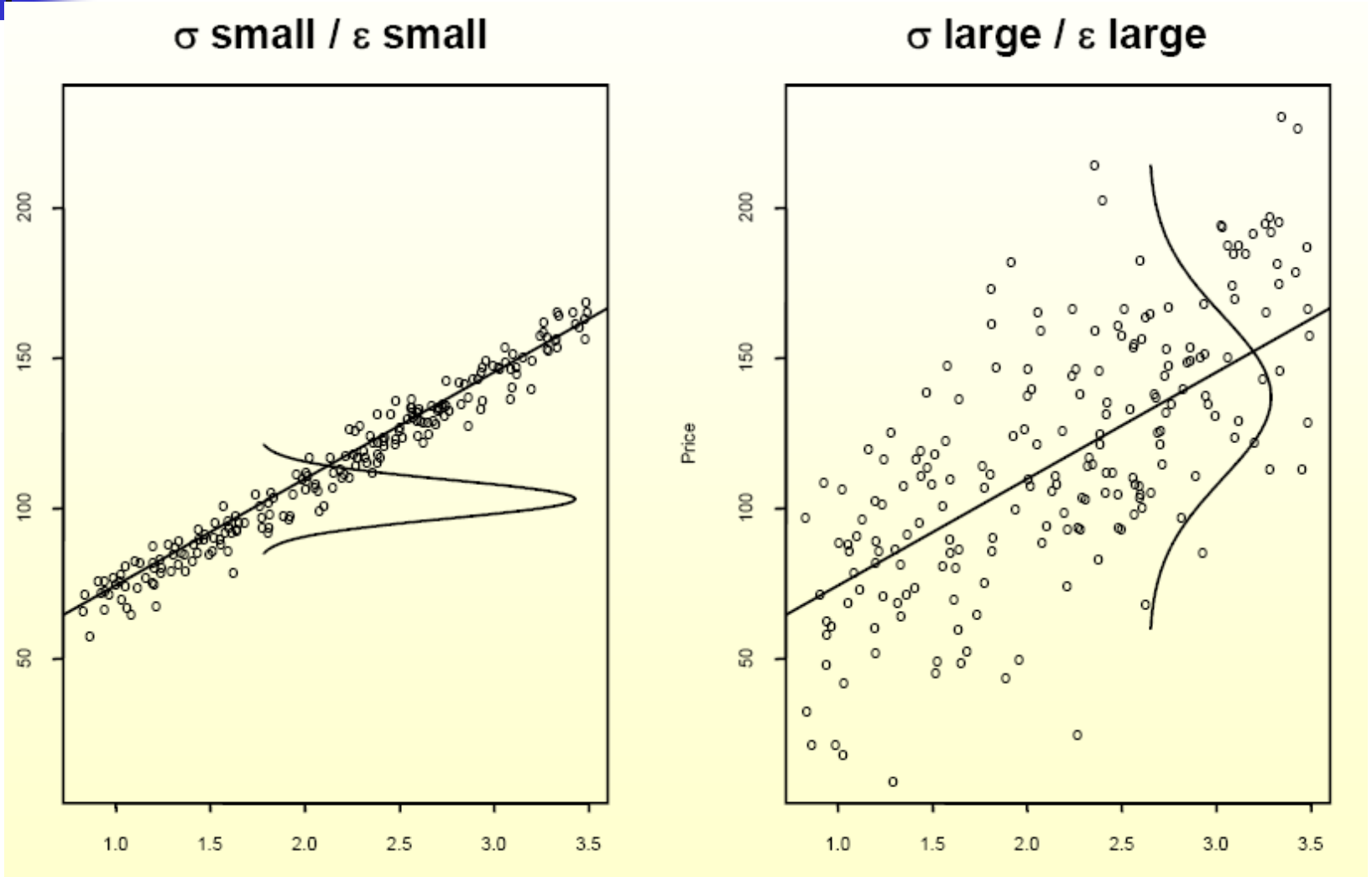
$$\varepsilon_i \sim \text{iid } N(0, \sigma^2)$$

MODEL

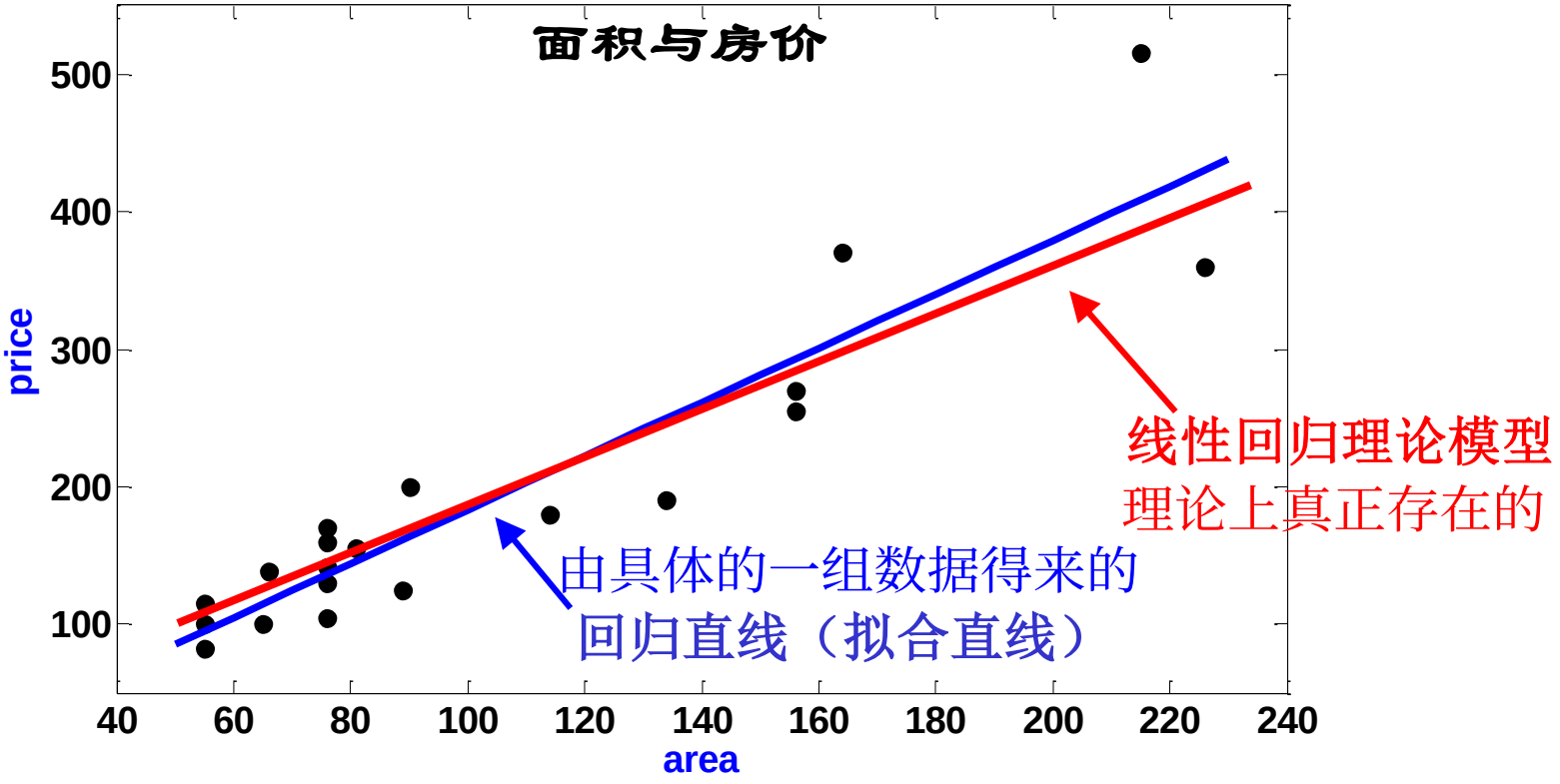
Independent and identically distributed

将Y分解为：与X相关的部分以及独立于X的残差这两部分之和

简单线性回归


σ small / ε small
σ large / ε large
Price
50
100
150
200
1.0
1.5
2.0
2.5
3.0
3.5

“真正的”回归直线


面积与房价
price
area
500
400
300
200
100
40
60
80
100
120
140
160
180
200
220
240
线性回归理论模型
理论上真正存在的
由具体的一组数据得来的
回归直线（拟合直线）

线性回归的任务

数据：

$$\begin{array}{cccc} x_1 & x_2 & \cdots & x_n \\ y_1 & y_2 & \cdots & y_n \end{array}$$

利用观测数据得到线性回归理论模型的近似

$$\hat{Y} = b_0 + b_1 X \qquad \approx \qquad Y = \beta_0 + \beta_1 X + \varepsilon$$

最小二乘近似　　理论模型

理论模型的参数β0,β1是具体存在的，ε是均值为0的随机变量，而具体的一个观测值可以找出一组b0、b1是随机变量，以此用作对β0,β1的估计

观测数据与记号（模型中的重要统计量）

$$\begin{matrix} x_1 & x_2 & \cdots & x_n \\ y_1 & y_2 & \cdots & y_n \end{matrix}$$

$$\hat{Y} = b_0 + b_1 X$$

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

$$\bar{x} = \frac{x_1 + x_2 + \cdots x_n}{n}$$

$$\bar{y} = \frac{y_1 + y_2 + \cdots y_n}{n}$$

$$S_X{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$S_Y{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$L_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$L_{YY} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$L_{XY} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{L_{XY}}{L_{XX}} = \frac{S_{XY}}{S_X{}^2}$$

$$b_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{L_{XX}}\right)\right), \quad b_1 \sim N\left(\beta_1, \frac{\sigma^2}{L_{XX}}\right)$$

最小二乘估计的统计性质

(1) $b_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{(n-1)s_x{}^2}\right)\right)$

(2) $b_1 \sim N\left(\beta_1, \frac{\sigma^2}{(n-1)s_x{}^2}\right)$

(3) $\frac{\mathrm{SSE}}{\sigma^2} \sim \chi^2(n-2)$

(4) $\bar{y}$，b_1，SSE 相互独立。

线性与非线性最小二乘

最小二乘： $\min E(X-c)^2 \Rightarrow c = E(X)$,

$$\min E(Y-(aX+b))^2 \Rightarrow a = \frac{Cov(X,Y)}{\sigma_X^{\,2}}, \quad b = E(Y) - \frac{Cov(X,Y)}{\sigma_X^{\,2}} E(X)$$

定理：二元随机变量(X,Y)的联合密度函数为$p_{XY}(x,y)$，$E(Y^2)$存在，令

$$\varphi(x) = \begin{cases} E(Y \mid X = x), & p_X(x) > 0 \\ 0, & ohterwise \end{cases}$$，则 $E(Y-\varphi(X))^2 = \min_{\psi} E(Y-\psi(X))^2$。

定理：二元随机变量(X,Y)的联合密度函数为$p_{XY}(x,y)$，$E(Y^2)$存在，令

$$\varphi(x)=\begin{cases} E(Y\mid X=x), & p_X(x)>0 \\ 0, & ohterwise \end{cases},$$ 则 $E(Y-\varphi(X))^2=\min_{\psi} E(Y-\psi(X))^2$。

证明：考虑连续随机变量的情形

$$E[Y\cdot(\varphi(X)-\psi(X))]=\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} y\cdot(\varphi(x)-\psi(x))p(x,y)dxdy=\int_{-\infty}^{+\infty}(\varphi(x)-\psi(x))dx\int_{-\infty}^{+\infty} y\cdot p(x,y)dy$$

$$=\int_{-\infty}^{+\infty}(\varphi(x)-\psi(x))\int_{-\infty}^{+\infty} y\cdot p(y\mid x)p_X(x)dxdy=\int_{-\infty}^{+\infty}(\varphi(x)-\psi(x))p_X(x)dx\int_{-\infty}^{+\infty} y\cdot p(y\mid x)dy$$

$$=\int_{-\infty}^{+\infty}(\varphi(x)-\psi(x))p_X(x)\varphi(x)dx=E[\varphi(X)\cdot(\varphi(X)-\psi(X))].$$

$$E(Y-\psi(X))^2=E(Y-\varphi(X)+\varphi(X)-\psi(X))^2$$

$$=E(Y-\varphi(X))^2+E(\varphi(X)-\psi(X))^2+2\cdot E[(Y-\varphi(X))\cdot(\varphi(X)-\psi(X))]$$

$$=E(Y-\varphi(X))^2+E(\varphi(X)-\psi(X))^2.$$

条件期望与最佳预报

注记: 条件期望 $E(Y \mid X)$ 是一个依赖随机变量 X 的随机变量，而它恰好是可用 X 表示的所有随机变量中最小二乘意义下与 Y 最为接近的随机变量。

例. 设 $p(x,y)=\begin{cases}1, & 0<|y|<x<1\\ 0, & ohterwise\end{cases}$，计算 $E(Y \mid X)$，$E(X \mid Y)$，并考虑这两个条件期望的直观意义。

解： $p_X(x)=2x\cdot I_{0<x<1}$，$p_Y(y)=(1-|y|)\cdot I_{|y|<1}$，

$$E(X \mid Y)=\frac{1+|Y|}{2}，\quad E(Y \mid X)=0。$$

一元线性回归模型

数据：

x_1	x_2	$\cdots$	x_n
y_1	y_2	$\cdots$	y_n

利用观测数据得到线性回归理论模型的近似

$$\hat{Y} = b_0 + b_1 X \quad \approx \quad Y = \beta_0 + \beta_1 X + \varepsilon$$

最小二乘近似　　理论模型

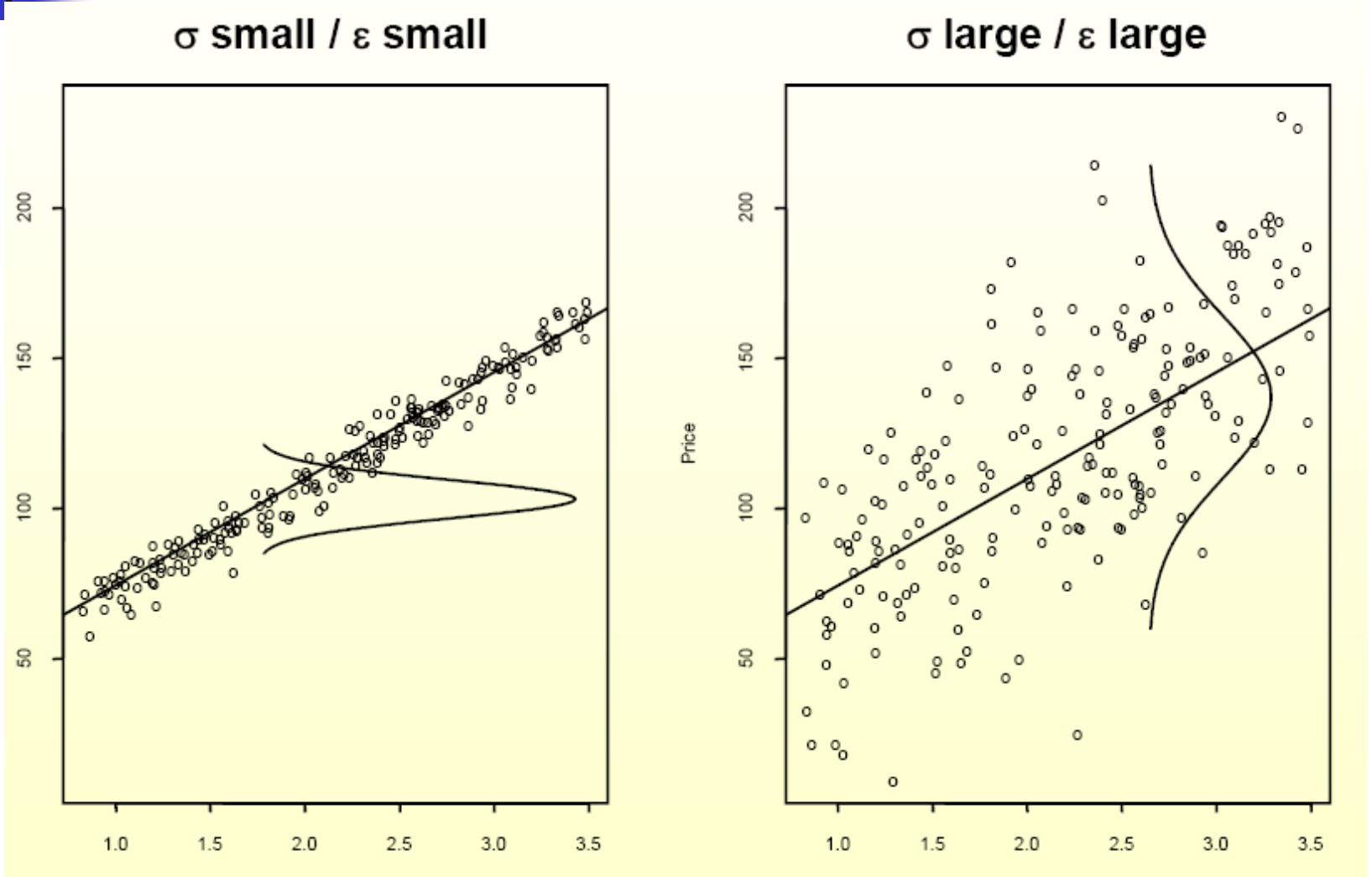
目的是估计理论模型！！详细可见统推3-2的67页

$$\min \sum_{i=1}^{n} \left(y_i - \left(b_0 + b_1 x_i \right) \right)^2$$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\varepsilon_i \overset{i.i.d}{\sim} N\left(0, \sigma^2\right)$$

简单线性回归


σ small / ε small
σ large / ε large
Price
50
100
150
200
1.0
1.5
2.0
2.5
3.0
3.5

观测数据与记号（模型中的重要统计量）

$$\begin{matrix} x_1 & x_2 & \cdots & x_n \\ y_1 & y_2 & \cdots & y_n \end{matrix}$$

$$\hat{Y} = b_0 + b_1 X$$

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

$$\bar{x} = \frac{x_1 + x_2 + \cdots x_n}{n}$$

$$\bar{y} = \frac{y_1 + y_2 + \cdots y_n}{n}$$

$$S_X{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$S_Y{}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$S_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$L_{XX} = \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$L_{YY} = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

$$L_{XY} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{L_{XY}}{L_{XX}} = \frac{S_{XY}}{S_X{}^2}$$

$$b_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{L_{XX}}\right)\right), \quad b_1 \sim N\left(\beta_1, \frac{\sigma^2}{L_{XX}}\right)$$

方差的分解
ANOVA （analysis of variance）

$$SST = L_{YY} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = b_1^2 L_{XX}, \quad SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$SST = SSR + SSE \qquad \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

总偏差平方和：**SST**； 回归平方和：**SSR**； 剩余平方和：**SSE**

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$$

$$\hat{\sigma}_e^2 = \frac{SSE}{n-2}$$ 是σ^2的无偏估计

σ^2的区间估计 $\left[\frac{SSE}{\chi^2_{1-\alpha/2}(n-2)}, \frac{SSE}{\chi^2_{\alpha/2}(n-2)}\right]$

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$R_a^2 = 1 - \frac{SSE/(n-2)}{SST/(n-1)}$$

最小二乘估计的统计性质

(1) $b_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{(n-1)s_x{}^2}\right)\right)$

(2) $b_1 \sim N\left(\beta_1, \frac{\sigma^2}{(n-1)s_x{}^2}\right)$

(3) $\frac{\mathrm{SSE}}{\sigma^2} \sim \chi^2(n-2)$

(4) $\bar{y}$，b_1，SSE 相互独立。

回归方程的显著性检验

- 实际工作中，我们并不能事先断定 X 与 Y 之间确有线性关系，$Y = \beta_0 + \beta_1 x + \varepsilon$ 只是一种假设。当然，这个假设是否有根据，我们可以通过专业知识和散点图做粗略的判断。更可靠地做出定量判断的方法基于假设检验的理论，定义如下假设检验问题：$H_0 : \beta_1 = 0 \quad VS \quad H_1 : \beta_1 \neq 0$

- 若拒绝 H_0，则认为 X 与 Y 之间确有线性关系，所求线性回归方程有意义；若接受 H_0，则意味着 X 与 Y 之间没有明显的线性关系，所得回归方程缺乏实际意义。

回归方程的显著性检验，F检验

$$b_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{L_{XX}}\right)\right), \quad b_1 \sim N\left(\beta_1, \frac{\sigma^2}{L_{XX}}\right)$$

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$$

$$H_0: \beta_1 = 0 \quad VS \quad H_1: \beta_1 \neq 0$$

当$H_0: \beta_1 = 0$成立时，$b_1 \sim N\left(0, \frac{\sigma^2}{L_{XX}}\right) \Rightarrow \frac{b_1\sqrt{L_{XX}}}{\sigma} \sim N(0,1)$

回归平方和 $SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = b_1^2 L_{XX} \Rightarrow \frac{SSR}{\sigma^2} \sim \chi^2(1)$

且SSE与SSR独立。 $F = \frac{SSR}{SSE/(n-2)} \sim F(1, n-2)$

F 越小说明相关程度越小

回归方程系数的 t 检验和区间估计

$$b_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{L_{XX}}\right)\right), \quad b_1 \sim N\left(\beta_1, \frac{\sigma^2}{L_{XX}}\right)$$

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$$

$$S_{\beta_0}{}^2 = \left(\frac{1}{n}+\frac{\bar{x}^2}{L_{XX}}\right)\cdot\frac{SSE}{n-2}, \qquad S_{\beta_1}{}^2 = \frac{SSE}{(n-2)L_{XX}}$$

当$H_0: \beta_0 = 0$成立时，$\dfrac{b_0}{S_{\beta_0}} \sim t(n-2)$；　当$H_0: \beta_1 = 0$成立时，$\dfrac{b_1}{S_{\beta_1}} \sim t(n-2)$

β_0的区间估计 $\left[b_0 - t_{\alpha/2}(n-2)\cdot S_{\beta_0},\ b_0 + t_{\alpha/2}(n-2)\cdot S_{\beta_0}\right]$

β_1的区间估计 $\left[b_1 - t_{\alpha/2}(n-2)\cdot S_{\beta_1},\ b_1 + t_{\alpha/2}(n-2)\cdot S_{\beta_1}\right]$

回归方程的显著性检验

F统计量：当$H_0:\beta_1=0$成立时，回归平方和 $SSR=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$

满足$\frac{SSR}{\sigma^2}\sim\chi^2(1)$，且SSE与SSR独立。

$$F=\frac{SSR}{SSE/(n-2)}\sim F(1,n-2)$$

t统计量：$\frac{\sqrt{n-1}s_X(b_1-\beta_1)}{\sqrt{SSE/(n-2)}}\sim t(1,n-2)$

$$\frac{(b_0-\beta_0)/\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{(n-1)s_X{}^2}}}{\sqrt{SSE/(n-2)}}=\frac{b_0-\beta_0}{\sqrt{\frac{1}{n}+\frac{\bar{x}^2}{(n-1)s_X{}^2}}\cdot\sqrt{SSE/(n-2)}}\sim t(1,n-2)$$

当$H_0:\beta_1=0$成立时，$T=\frac{b_1}{\sqrt{SSE/(n-2)}/\sqrt{n-1}s_X}=\frac{b_1}{s_{\beta_1}}\sim t(1,n-2)$

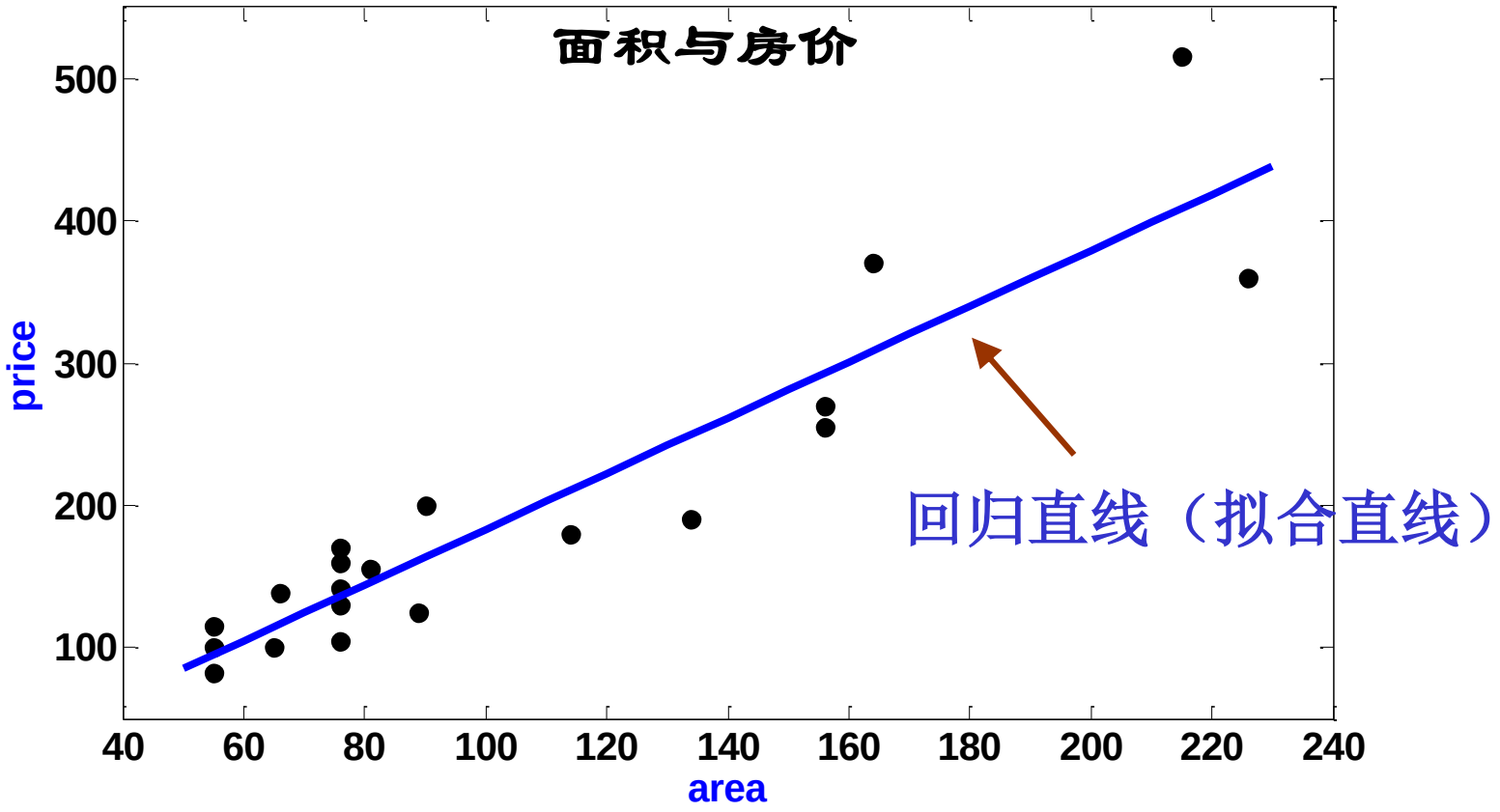
回归分析实例--北京房价分析

编号	面积（平方米）	价格(万元)
1	55	100
2	76	130
3	65	100
4	156	255
5	55	82
6	76	105
7	89	125
8	226	360
9	134	190
10	156	270

编号	面积（平方米）	价格(万元)
11	114	180
12	76	142
13	164	370
14	55	115
15	90	200
16	81	155
17	215	516
18	76	160
19	66	138
20	76	170

因变量（预报变量）， 因变量（响应变量）

北京房价的回归分析


面积与房价
price
area
500
400
300
200
100
40
60
80
100
120
140
160
180
200
220
240
回归直线（拟合直线）

回归计算

$$\bar{x} = \frac{x_1 + x_2 + \cdots x_n}{n} = 108.8$$

$$\bar{y} = \frac{y_1 + y_2 + \cdots y_n}{n} = 171.7$$

编号	面积（平方米）	价格(万元)
1	55	100
2	76	130
3	65	100
4	156	255
5	55	82
6	76	105
7	89	125
8	226	360
9	134	190
10	156	270

$$L_{XX} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = 29077.6$$

$$L_{XY} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = 47405.4$$

$$b_1 = \frac{L_{XY}}{L_{XX}} = 1.6303$$

$$b_0 = \bar{y} - b_1 \bar{x} = -5.677$$

$$SST = L_{YY} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 79090.1$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 77285.3$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 1804.8$$

$$R^2 = 1 - \frac{SSE}{SST} = 0.977, \quad R_a^{\,2} = 1 - \frac{SSE/8}{SST/9} = 0.974,$$

$$\hat{\sigma}_e^{\,2} = \frac{SSE}{n-2} = 225.6, \quad \hat{\sigma}_e = 15.02$$

回归计算

编号	面积（平方米）	价格(万元)
1	55	100
2	76	130
3	65	100
4	156	255
5	55	82
6	76	105
7	89	125
8	226	360
9	134	190
10	156	270

$$\bar{x} = \frac{x_1 + x_2 + \cdots x_n}{n} = 108.8$$

$$L_{XX} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = 29077.6$$

$$SST = L_{YY} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 79090.1$$

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 77285.3$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 1804.8$$

$$F = \frac{SSR}{SSE/(n-2)} = \frac{77285.3}{225.6} = 342.6, \quad F_{0.01}(1,8) = 11.3$$

$$\frac{b_0}{S_{\beta_0}} = \frac{-5.677}{10.696} = -0.531, \quad \frac{b_1}{S_{\beta_1}} = \frac{1.630}{0.088} = 18.509$$

$$t_{0.005}(8) = 3.3554, \quad t_{0.025}(8) = 2.306$$

$$S_{\beta_0} = \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{L_{XX}}\right) \cdot \frac{SSE}{n-2}} = \sqrt{114.4} = 10.696$$

$$S_{\beta_1} = \sqrt{\frac{SSE}{(n-2)L_{XX}}} = \sqrt{0.00776} = 0.088$$

SPSS运行输出结果

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	area[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: price

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.926[a]	.857	.849	43.12927

a. Predictors: (Constant), area

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

$\hat{\sigma}_e$

SPSS运行输出结果

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	200792.145	1	200792.145	107.945	.000[a]
	Residual	33482.405	18	1860.134		
	Total	234274.550	19			

a. Predictors: (Constant), area

b. Dependent Variable: price

S_{β_i}

$\beta = b_1 \frac{s_X}{s_Y} = b_1 \sqrt{\frac{L_{XX}}{L_{YY}}}$

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-12.825	22.046		-.582	.568
	area	1.961	.189	.926	10.390	.000

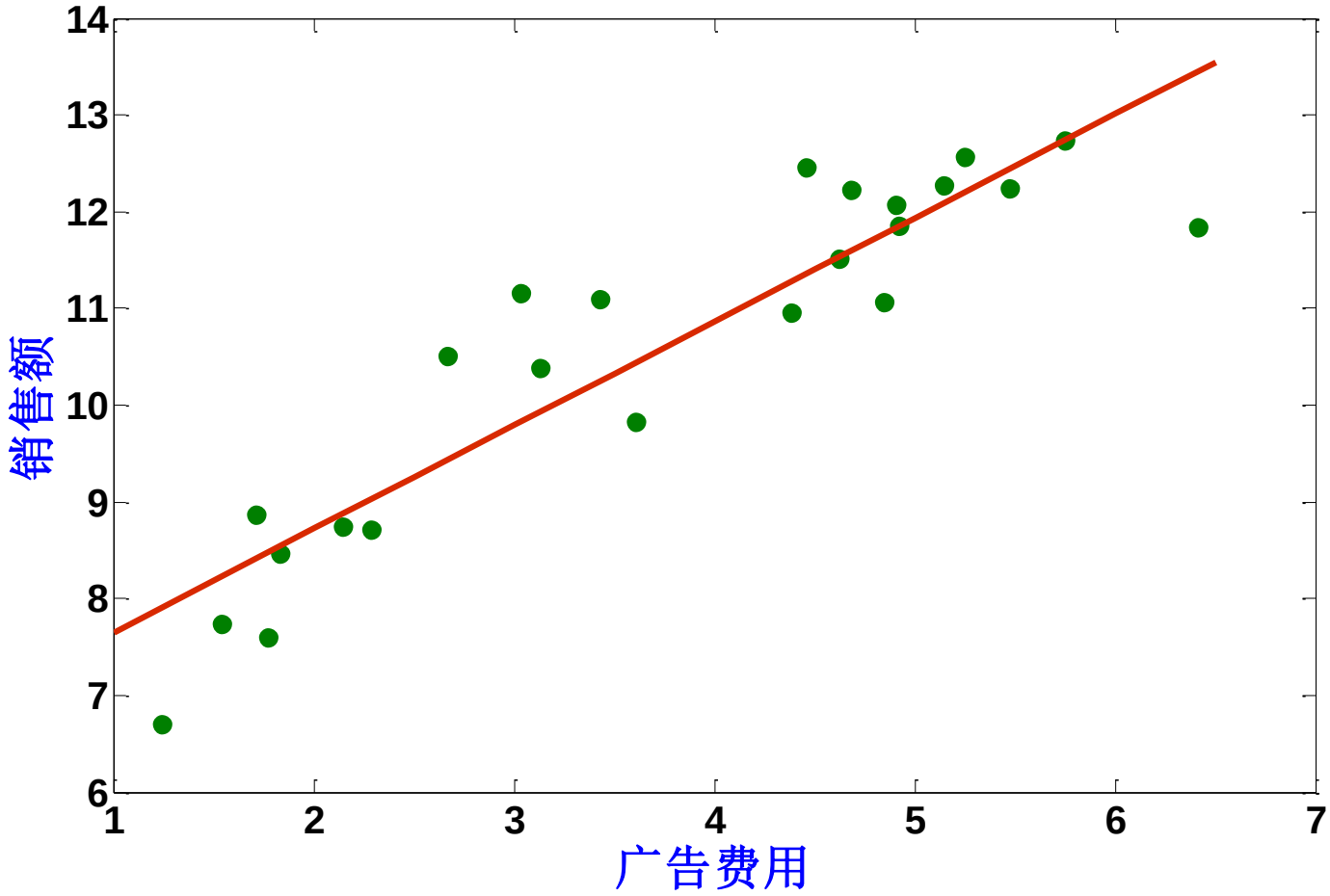
a. Dependent Variable: price

广告费用与销售额的回归分析

编号	广告费用（万元）	销售额（万元）
1	4.6864	12.2284
2	6.4108	11.8436
3	5.4734	12.2475
4	3.4323	11.099
5	4.3878	10.9666
6	2.1459	8.7517
7	1.539	7.7473
8	2.6655	10.5034
9	1.2445	6.7085
10	1.7747	7.6044
11	4.4602	12.461
12	1.8321	8.4699

编号	广告费用（万元）	销售额（万元）
13	5.1469	12.2733
14	5.2505	12.5655
15	1.716	8.8674
16	3.0363	11.1537
17	4.9228	11.8595
18	4.8472	11.0672
19	3.1284	10.3842
20	2.2864	8.7083
21	4.9048	12.0686
22	5.7479	12.7386
23	3.6116	9.8222
24	4.6223	11.5108

广告费用与销售额的回归分析



SPSS运行输出结果

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	广告费用[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: 销售量

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.916[a]	.839	.832	.73875

a. Predictors: (Constant), 广告费用

SPSS运行输出结果

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	62.514	1	62.514	114.548	.000[a]
	Residual	12.006	22	.546		
	Total	74.520	23			

a. Predictors: (Constant), 广告费用

b. Dependent Variable: 销售量

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	6.584	.402		16.391	.000
	广告费用	1.071	.100	.916	10.703	.000

a. Dependent Variable: 销售量

预测

设 Y 与 X 满足线性模型：$Y=\beta_0+\beta_1 X+\varepsilon$，其中 $\varepsilon\sim N(0,\sigma^2)$

根据历史样本 $(x_1,y_1),\cdots,(x_n,y_n)$ 求得回归方程 $\hat{Y}=b_0+b_1X$.

令 x_0 表示 X 的某个固定值，且 $y_0=\beta_0+\beta_1 x_0+\varepsilon_0$，$\varepsilon_0\sim N(0,\sigma^2)$.

设 y_0 与 $y_1,\cdots,y_n$ 相互独立，求 y_0 的预测值和预测区间。

以回归值 $\hat{y}_0=b_0+b_1x_0$ 作为 y_0 的预测值，是无偏估计

$$E(\hat{y}_0)=E(b_0+b_1x_0)=\beta_0+\beta_1x_0=E(y_0)$$

考虑 $y_0-\hat{y}_0$ 的分布：

$\hat{y}_0=b_0+b_1x_0=\bar{y}-b_1\bar{x}+b_1x_0=\bar{y}+b_1(x_0-\bar{x})$，$\bar{y}$ 与 b_1 独立

$$\Rightarrow Var(\hat{y}_0)=Var(\bar{y})+(x_0-\bar{x})^2Var(b_1)=\sigma^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{XX}}\right]$$

预测区间

$$\hat{y}_0 \sim N\left(\beta_0+\beta_1 x_0, \sigma^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{XX}}\right]\right)$$

$$y_0 \sim N\left(\beta_0+\beta_1 x_0, \sigma^2\right) \Rightarrow Var(y_0-\hat{y}_0)=\sigma^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{XX}}\right]$$

$$y_0-\hat{y}_0 \sim N\left(0, \sigma^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{XX}}\right]\right)$$

$$\frac{SSE}{\sigma^2} \sim \chi^2(n-2)$$

$$T=\frac{\dfrac{y_0-\hat{y}_0}{\sigma\sqrt{1+\dfrac{1}{n}+\dfrac{(x_0-\bar{x})^2}{L_{XX}}}}}{\sqrt{\dfrac{SSE}{\sigma^2(n-2)}}}=\frac{y_0-\hat{y}_0}{\sqrt{\dfrac{SSE}{(n-2)}}\sqrt{1+\dfrac{1}{n}+\dfrac{(x_0-\bar{x})^2}{L_{XX}}}}=\frac{y_0-\hat{y}_0}{\hat{\sigma}_e\cdot\sqrt{1+\dfrac{1}{n}+\dfrac{(x_0-\bar{x})^2}{L_{XX}}}} =\sim t(n-2)$$

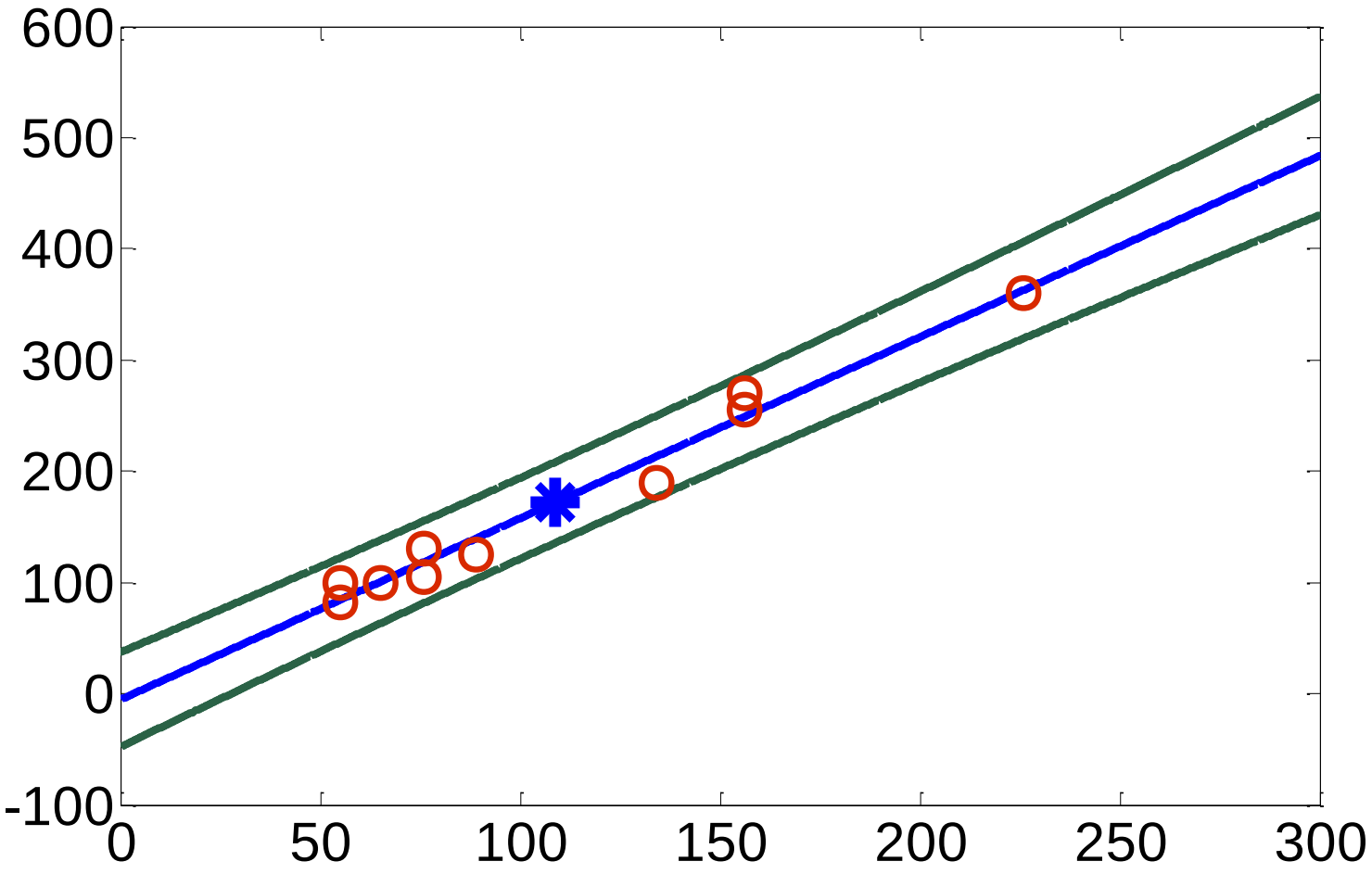
预测区间

$$T=\frac{y_0-\hat{y}_0}{\hat{\sigma}_e\cdot\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{XX}}}}\sim t(n-2)$$

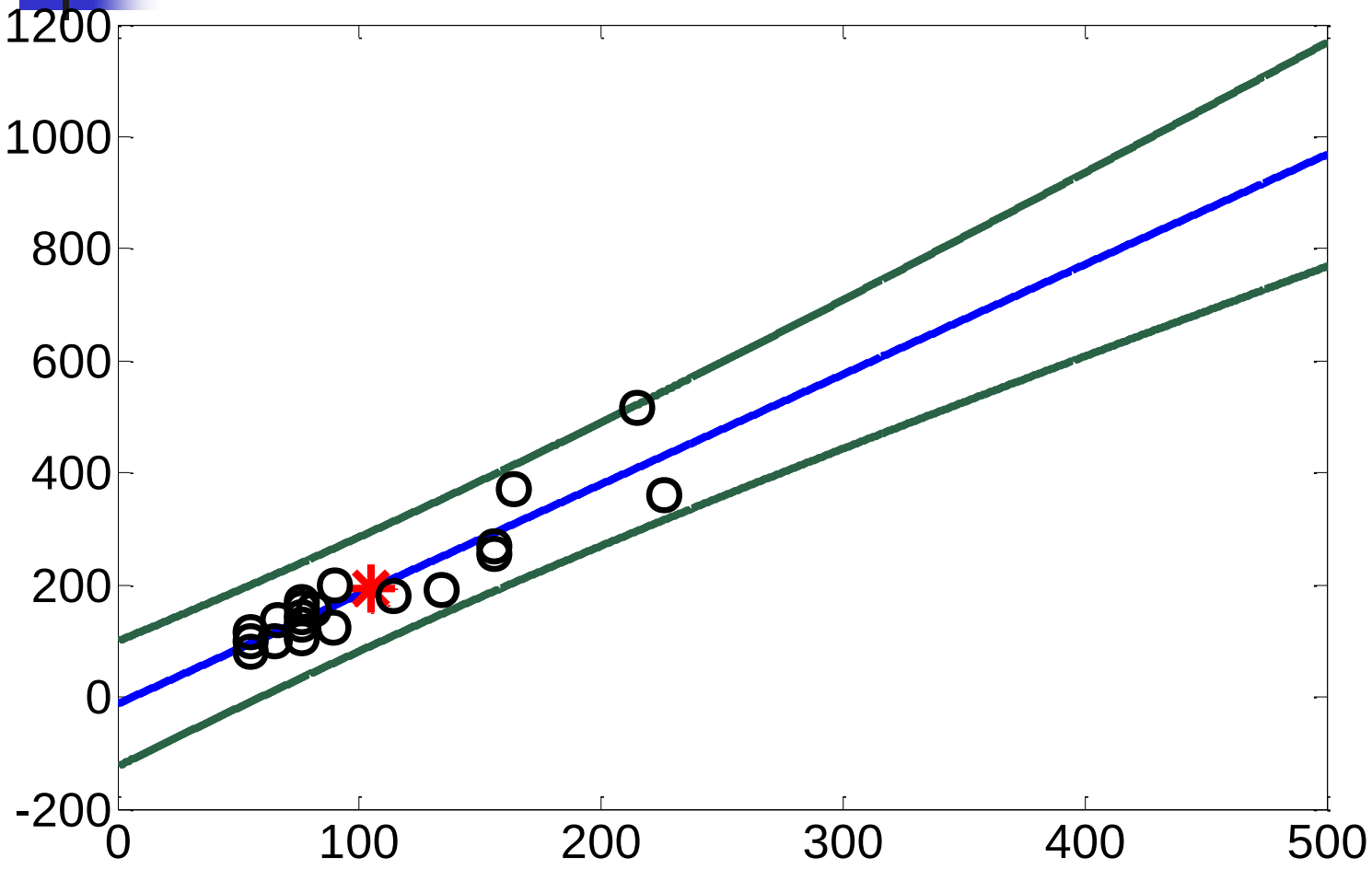
$$\delta(x_0)=t_{\alpha/2}(n-2)\cdot\hat{\sigma}_e\cdot\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{L_{XX}}}$$

y_0 的置信水平 $1-\alpha$ 的区间估计：$[\hat{y}_0-\delta(x_0),\hat{y}_0+\delta(x_0)]$

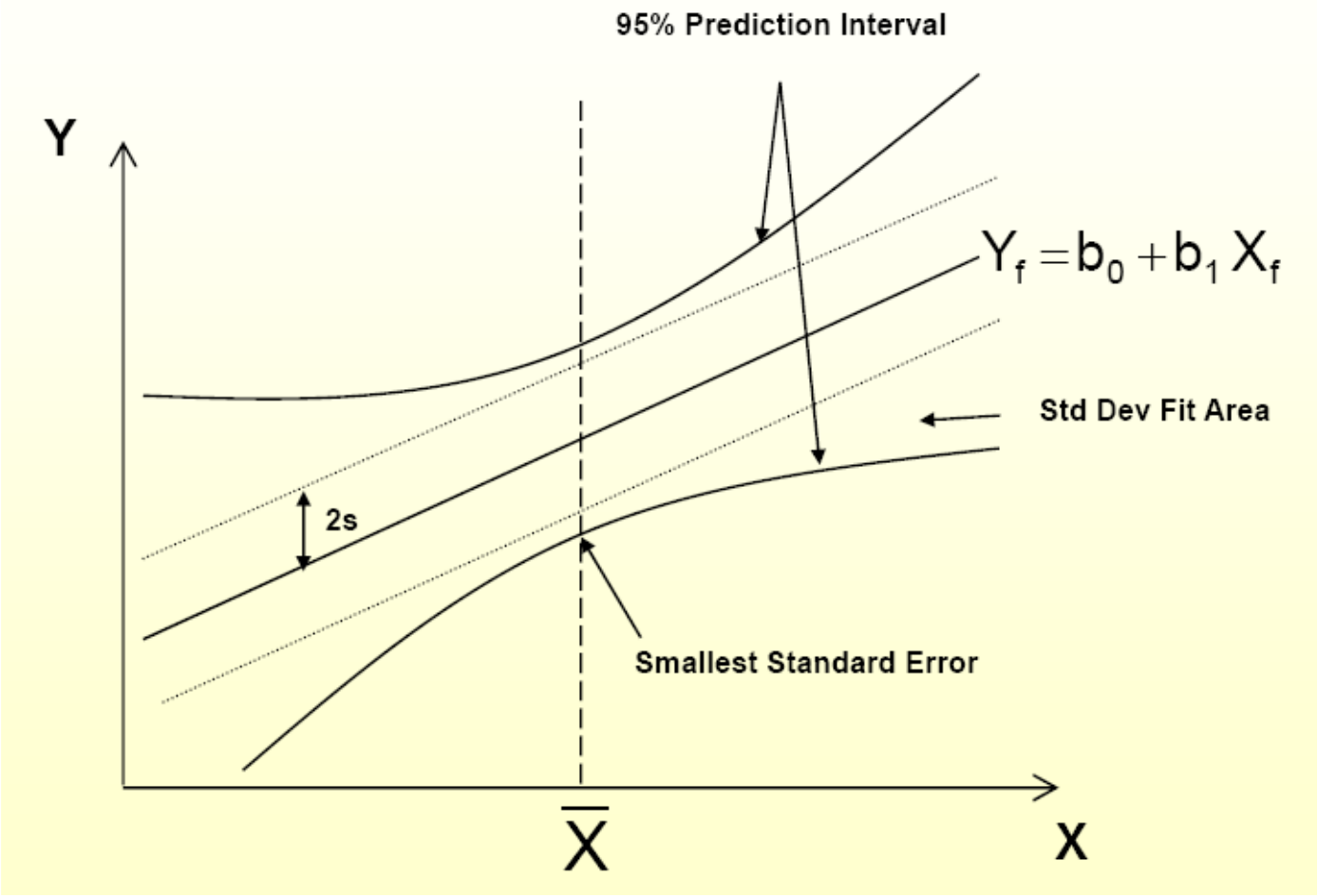
预测区间：北京房价，10个观测数据



预测区间：北京房价，20个观测数据



预测区间


95% Prediction Interval
Y
$Y_f = b_0 + b_1 X_f$
Std Dev Fit Area
2s
Smallest Standard Error
$\overline{X}$
X

作业

- 236页，习题五　8

补充题　设 $\hat{Y}=b_0+b_1X$ 是由 n 对观测数据 $(x_1,y_1),\cdots,(x_n,y_n)$ 得到的理想回归模型的最小二乘近似，其中 $b_1=\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$，

（1）一元回归理想模型 $Y=\beta_0+\beta_1X+\varepsilon$ 中 ε 应满足什么条件？

（2）证明 $b_1\sim N\left(\beta_1,\dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$；

（3）又已知 $\dfrac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sigma^2}\sim\chi^2(n-2)$，求参数 β_1 的 $1-\alpha$ 置信区间。