第四章 回归预测技术

变量多少：一元回归，多元回归

回归方程的性质：线性、非线性

变量的属性：数量回归、非数量回归

第一节　一元线性回归预测

- 一元线性回归模型的基本假设
- 回归模型的参数估计
- 回归模型的假设检验
- 预测、控制和风险分析
- 举例

一、一元线性回归模型的基本假设

- 模型

$$y = a + bx + \varepsilon \qquad y_i = a + bx_i + \varepsilon_i$$

  - 其中，ε 是一种随机干扰或误差项，
  - x 是确定性的变量，自变量
  - y 是随机变量，因变量

目标：由一组统计数据 $(x_i,\ y_i)$，求出 a, b 的估计值 $\hat{a},\ \hat{b}$，建立预测方程 $\hat{y} = \hat{a} + \hat{b}x$ 。

其中 $\hat{a},\ \hat{b}$ 称为回归系数，再通过预测方程进行预测，并给出预测的置信区间。

- 随机误差 ε 的来源
  - 模型中被忽略掉的影响因素造成的误差

    消费支出=f(收入,人口数,消费习惯,存款利率,商品价格水平变化趋势)
  - 模型关系设定不准确造成的误差
  - 变量的测量误差
  - 随机误差

    销售量=f(收入,价格,消费心理)　　亩产量=f(施肥量,天气)

- 回归模型的基本假定
  - 关于变量和模型的假定

    1. x_i 是非随机的，或者 x_i 虽然是随机的，但与 ε_i 是独立的。
    2. x_i 无测量误差
    3. 不存在设定误差

- 关于随机误差项 ε_i 统计分布的假定

1. 零均值假定 $E(\varepsilon_i)=0$

$$\Rightarrow E(y_i)=E(a+bx_i+\varepsilon_i)=a+bx_i$$

2. 同方差假定 $Var(\varepsilon_i)=E(\varepsilon_i-E\varepsilon_i)^2=E\varepsilon_i^2=\sigma^2$

$$\Rightarrow Var(y_i)=E(y_i-Ey_i)^2=E(a+bx_i+\varepsilon_i-a-bx_i)^2=\sigma^2$$

3. 无自相关假定 $Cov(\varepsilon_i,\varepsilon_j)=E\varepsilon_i\varepsilon_j=0$

$$\Rightarrow Cov(y_i,y_j)=E\varepsilon_i\varepsilon_j=0$$

4. x_i与ε_i不相关假定$Cov(x_i,\varepsilon_i)=0$

5. 正态性假设 $\varepsilon_i\sim N(0,\sigma^2)\Rightarrow y_i\sim N(a+bx_i,\sigma^2)$

如果只利用最小二乘法进行参数估计,不需要假设5

如果进行假设检验和预测,则需要假设5,可知y的分布

二、回归模型的参数估计

- 普通最小二乘法（OLS）

$$\hat{y} = \hat{a} + \hat{b}x$$

残差： $e_i = y_i - \hat{y}_i = y_i - (\hat{a} + \hat{b}x_i)$

残差平方和： $Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}x_i)^2$

最小二乘法： 求解$\hat{a}$，$\hat{b}$，使得Q最小

由极值定理

$$\begin{cases} \dfrac{\partial Q}{\partial \hat{a}} = -2\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i) = 0 \\ \dfrac{\partial Q}{\partial \hat{b}} = -2\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)x_i = 0 \end{cases}$$

解得：

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i} \qquad \hat{a} = \bar{y} - \hat{b}\bar{x}$$

其中：

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

例1、已知某种商品的销售量同居民的可支配收入有关，并有统计数据见书63页表5.4。

(1)试建立回归方程，并求出相应参数的最小二乘估计；

年　　份	实际可支配收入 x 单位：10元	某种商品的销售量 y 单位：件
1983	522	6 700
1984	539	7 316
1985	577	7 658
1986	613	8 784
1987	644	8 408
1988	670	7 583
1989	695	8 600
1990	713	8 442
1991	741	7 158
1992	769	8 683
1993	801	9 317
1994	855	9 675
1995	842	7 542
1996	860	7 084
1997	890	8 612
1998	920	9 119

- 解

样本序号	实际可支配收入 x_i	x_i^2	某种商品销售量 y_i	y_i^2	x_iy_i
1	522	272 484	6 700	4 489 000	3 197 400
2	539	290 521	7 316	53 523 856	3 943 324
3	577	332 929	7 658	58 644 964	4 418 666
4	613	375 769	8 784	77 158 656	5 384 592
5	644	414 736	8 408	70 691 464	5 414 752
6	670	448 900	7 583	57 501 819	5 080 610
7	695	483 025	8 600	73 960 000	5 977 000
8	713	508 369	8 442	71 267 364	6 019 146
9	741	549 081	7 158	51 236 964	5 304 078
10	769	591 361	8 683	75 394 489	6 677 227
11	801	641 601	9 317	86 806 489	7 462 917
12	855	731 025	9 675	93 605 625	8 272 165
13	842	708 964	7 542	56 881 764	6 350 364
14	860	739 600	7 084	50 183 056	6 092 240
15	890	792 100	8 612	74 166 544	7 692 240
16	926	857 476	9 119	83 156 161	8 444 194
$\sum$	11 657	2 737 941	130 681		96 003 315

$$\bar{x} = \sum_{i=1}^{n} x_i / n = \frac{11\,657}{16} = 728.56$$

$$\bar{y} = \sum_{i=1}^{n} y_i / n = \frac{130\,681}{16} = 8\,167.6$$

由于 $\hat{b} = \dfrac{\sum_{i=1}^{n} x_i y_i - \bar{x}\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \bar{x}\sum_{i=1}^{n} x_i}$ $\qquad \hat{a} = \bar{y} - \hat{b}\bar{x}$

$$\hat{b} = \frac{96\,003\,315 - 728.56 \times 130\,681}{8\,737\,941 - 728.56 \times 11\,657}$$

$$= \frac{794\,365.64}{245\,117.08} = 3.24$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x} = 8\,167.6 - 3.24 \times 728.56 = 5\,807$$

$$\hat{y} = 5\,807 + 3.24x$$

■ 最小二乘估计量的性质

1.线性： $\hat{a},\hat{b}$分别为y_i和ε_i的线性函数或线性组合

证明：$$\hat{b}=\frac{\sum_{i=1}^{n}x_iy_i-\bar{x}\sum_{i=1}^{n}y_i}{\sum_{i=1}^{n}x_i^2-\bar{x}\sum_{i=1}^{n}x_i}=\frac{\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}\bar{x}y_i}{\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}\bar{x}x_i}=\sum_{i=1}^{n}\frac{(x_i-\bar{x})y_i}{\sum_{i=1}^{n}x_i^2-\sum_{i=1}^{n}\bar{x}x_i}$$

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

$$\hat{a}=\bar{y}-\hat{b}\bar{x}=\frac{1}{n}\sum y_i-\bar{x}\sum k_iy_i=\sum(\frac{1}{n}-\bar{x}k_i)y_i$$

其中：$$k_i=\frac{(x_i-\bar{x})}{\sum(x_i-\bar{x})^2}$$

■ 最小二乘估计量的性质

2.无偏性： $E(\hat{a}) = a, E(\hat{b}) = b$

证明：$\hat{b} = \sum k_i y_i = \sum k_i (a + bx_i + \varepsilon_i) = a\sum k_i + b\sum k_i x_i + \sum k_i \varepsilon_i$

可以证明 $\sum k_i = 0 \quad \sum k_i x_i = 1$

因此 $\hat{b} = b + \sum k_i \varepsilon_i \quad E(\hat{b}) = E(b + \sum k_i \varepsilon_i) = b$

$$\hat{a} = \sum (\frac{1}{n} - \bar{x}k_i) y_i = \sum (\frac{1}{n} - \bar{x}k_i)(a + bx_i + \varepsilon_i)$$

$$= a\sum (\frac{1}{n} - \bar{x}k_i) + b\sum (\frac{1}{n} - \bar{x}k_i) x_i + \sum (\frac{1}{n} - \bar{x}k_i) \varepsilon_i$$

$$\hat{a} = a + \sum (\frac{1}{n} - \bar{x}k_i) \varepsilon_i \qquad E(\hat{a}) = a$$

■ 最小二乘估计量的性质

3.有效性：在所有线性、无偏估计量中，最小二乘估计量$\hat{a},\hat{b}$的方差最小。

证明：$D(\hat{b})=D(\sum k_i y_i)=\sum k_i^2 D(y_i)=\sum k_i^2\sigma^2$

可以证明 $\sum k_i^2=\dfrac{1}{\sum(x_i-\bar{x})^2}$

因此 $D(\hat{b})=\dfrac{\sigma^2}{\sum(x_i-\bar{x})^2}$

$$D(\hat{a})=D\left(\sum(\frac{1}{n}-\bar{x}k_i)y_i\right)=\sum\left(\frac{1}{n}-\bar{x}k_i\right)^2\sigma^2=\frac{\sigma^2\sum x_i^2}{n\sum(x_i-\bar{x})^2}$$

假设 $\hat{b}^* = \sum c_i y_i$ 是 b 的另一个线性无偏估计量

$c_i \neq k_i \ (i = 1, 2, \cdots, n)$

$$E(\hat{b}^*) = b = E(\sum c_i y_i) = \sum c_i (a + b x_i) = a \sum c_i + b \sum c_i x_i$$

比较上式两边：$\sum c_i = 0 \quad \sum c_i x_i = 1$

$$D(\hat{b}^*) = \sigma^2 \sum c_i^2 = \sigma^2 \sum (c_i - k_i + k_i)^2 = \sigma^2 \sum \left((c_i - k_i)^2 + 2k_i (c_i - k_i) + k_i^2 \right)$$

又因为 $\sum k_i (c_i - k_i) = 0$

$$\text{上式} = \sigma^2 \sum k_i^2 + \sigma^2 \sum (c_i - k_i)^2 \geq \sigma^2 \sum k_i^2 = D(\hat{b})$$

- 回归参数的区间估计

1.$\hat{a},\hat{b}$的 分 布

$$\hat{a} \sim N(a, D(\hat{a}))$$

$$\hat{b} \sim N(b, D(\hat{b}))$$

2.随机误差项方差的估计

σ^2的无偏估计量 $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{n} e_i^2}{n-2}$ $E(\hat{\sigma}^2) = \sigma^2$

记估计标准误差 $S.E. = \hat{\sigma}$

$\hat{a}$的标准误差 $s(\hat{a}) = \sqrt{\dfrac{\hat{\sigma}^2 \sum x_i^2}{n\sum (x_i - \bar{x})^2}}$

$\hat{b}$的标准误差 $s(\hat{b}) = \sqrt{\dfrac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}}$

- 估计标准误差 $\hat{\sigma}$

  反映应变量的实际值 y_i 与估计值 $\hat{y}_i$ 的平均误差程度，$\hat{\sigma}$ 越大，则回归直线精度越低

- 参数估计量的标准误差 $s(\hat{a})$ $s(\hat{b})$

  反映参数估计值与真实值的误差程度

3.回归系数的区间估计

找出δ、α，使得 $P(\hat{b}-\delta \le b \le \hat{b}+\delta)=1-\alpha$

（1）σ^2已知，ε_i服从正态分布

$$Z=\frac{\hat{b}-b}{\sigma(\hat{b})}\sim N(0,1) \quad \text{其中} \quad \sigma(\hat{b})=\sqrt{\frac{\sigma^2}{\sum(x_i-\bar{x})^2}}$$

（2）σ^2未知，样本容量充分大

$$Z=\frac{\hat{b}-b}{s(\hat{b})}\sim N(0,1)$$

（3）σ^2未知，样本容量较小

$$t=\frac{\hat{b}-b}{s(\hat{b})}\sim t(n-2)$$

例：若σ^2已知，ε_i服从正态分布，α=0.05, 试求δ

解：$P(\hat{b}-\delta \le b \le \hat{b}+\delta)=0.95=P(-\frac{\delta}{\sigma(\hat{b})} \le \frac{b-\hat{b}}{\sigma(\hat{b})} \le \frac{\delta}{\sigma(\hat{b})})$

$$\Rightarrow 0.975=P(\frac{b-\hat{b}}{\sigma(\hat{b})} \le \frac{\delta}{\sigma(\hat{b})})$$

$$\therefore \frac{\delta}{\sigma(\hat{b})}=1.96 \qquad \delta=1.96\sigma(\hat{b})$$

表 I $_6$ 由 $\Phi(x)$ 求 x

$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x
0.50	0.0000	0.70	0.5244	0.900	1.2816	0.991	2.3656
0.51	0.0251	0.71	0.5534	0.905	1.3106	0.992	2.4089
0.52	0.0502	0.72	0.5828	0.910	1.3408	0.993	2.4573
0.53	0.0753	0.73	0.6128	0.915	1.3722	0.994	2.5121
0.54	0.1004	0.74	0.6434	0.920	1.4051	0.995	2.5758
0.55	0.1257	0.75	0.6745	0.925	1.4395	0.996	2.6521
0.56	0.1510	0.76	0.7063	0.930	1.4758	0.997	2.7478
0.57	0.1764	0.77	0.7389	0.935	1.5141	0.998	2.8782
0.58	0.2019	0.78	0.7722	0.940	1.5548	0.999	2.0902
0.59	0.2275	0.79	0.8064	0.945	1.5982		
0.60	0.2534	0.80	0.8416	0.950	1.6449	0.9991	3.1214
0.61	0.2793	0.81	0.8779	0.955	1.6954	0.9992	3.1559
0.62	0.3055	0.82	0.9154	0.960	1.7507	0.9993	3.1947
0.63	0.3319	0.83	0.9542	0.965	1.8119	0.9994	3.2389
0.64	0.3585	0.84	0.9945	0.970	1.8808	0.9995	3.2905
0.65	0.3853	0.85	1.0364	0.975	1.9600	0.9996	3.3528
0.66	0.4125	0.86	1.0803	0.980	2.0538	0.9997	3.4316
0.67	0.4399	0.87	1.1264	0.985	2.1701	0.9998	3.5401
0.68	0.4677	0.88	1.1750	0.990	2.3264	0.9999	3.7190
0.69	0.4959	0.89	1.2265				

三、回归模型的假设检验

先进行符号检验：由参数最小二乘估计值的符号及取值大小，
判断是否符合经济理论规定

1.总离差平方和的分解

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

$$S_{总}=U+Q$$

$S_{总}$：y_i与其平均值的总离差平方和

U：回归平方和，由x的变动引起的，表示$S_{总}$中可由回归直线解释

Q：残差平方和，由未能控制的因素引起的，不能由回归直线解释

2.可决系数：反映拟和优度，

即样本回归直线与样本观测值数据之间的拟合程度

$$r^2 = \frac{U}{S_{总}} = 1 - \frac{Q}{S_{总}}$$

r^2越大，拟合得越好

r^2越小，拟合得越差

3.相关系数：描述变量x、y之间的线性关系密切程度

$$\hat{r} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\overline{y}^2}}$$

性质：$\hat{r}^2 = r^2$

$\hat{r} = 0$　　零相关

$\hat{r} = \pm 1$　　完全正负相关

$|\hat{r}|$愈接近 1，x、y之间的线性关系越密切

$|\hat{r}|$愈接近 0，x、y之间的线性关系越不密切

4.显著性检验

(1) R检验：对于给定的显著性水平α，查相关系数表得$R_\alpha(n-2)$，若$|r| \geq R_\alpha(n-2)$，则两变量间线性相关关系显著，回归模型可以用来预测。否则不能。

(2) F检验：

$F = \dfrac{(n-2)U}{Q} = (n-2)\dfrac{\hat{r}^2}{1-\hat{r}^2}$ ，给定的显著性水平α，查F分布表得$F_\alpha(1, n-2)$， 若$F \geq F_\alpha(1, n-2)$，则线性相关关系显著，回归模型可以用来预测。否则不能。

在一元回归中，R检验、t检验、F检验是等价的

(3)Jarque-Bera(雅克—贝拉)检验：检验随机误差的正态性

偏度系数：$S=\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{n\sigma_x^3}$ ，峰度系数：$K=\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{n\sigma_x^4}$

$JB=\dfrac{n}{6}\left[S^2+\dfrac{(K-3)^2}{4}\right]$ ， 对于给定的显著性水平α，

查$\chi_\alpha^2(2)$， 若$JB\geq\chi_\alpha^2(2)$，则不满足正态性假设。否则满足。

其中： $\sigma_x=\sqrt{\dfrac{1}{n-1}\sum(x_i-\bar{x})^2}$

例1、已知某种商品的销售量同居民的可支配收入有关，并有统计数据见书59页表5.4。

(1)试建立回归方程，并求出相应参数的最小二乘估计；

(2)对回归方程进行显著性检验。

解：(1) $\hat{y}=5807+3.24x$

(2) 若用 R 检验：
$$\hat{r}=\frac{\sum_{i=1}^{n}x_i y_i-n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}\sqrt{\sum_{i=1}^{n}y_i^2-n\bar{y}^2}}$$
$$=\frac{96003315-16\times728.56\times8167.6}{\sqrt{8737941-16\times(728.56)^2}\sqrt{1038671215-16\times(8167.6)^2}}=0.468$$

α=0.05, $R_\alpha(14)$=0.497>0.468, 所以线性关系不显著。

若用 F 检验：$F=14\times\frac{0.219}{1-0.219}=4.08=(n-2)\frac{\hat{r}^2}{1-\hat{r}^2}$

α=0.05, F_α=4.60>4.08, 所以线性关系不显著

α=0.1, F_α=3.10<4.08, 所以线性关系显著

相关系数表（一）

表 V—1

自由度	$\alpha=5\%$			
	自变量和因变量总数			
	2	3	4	5
1	0.997	0.999	0.999	0.999
2	0.950	0.975	0.983	0.987
3	0.878	0.930	0.950	0.961
4	0.811	0.881	0.912	0.930
5	0.754	0.836	0.874	0.898
6	0.707	0.795	0.839	0.867
7	0.666	0.758	0.807	0.838
8	0.632	0.726	0.777	0.811
9	0.602	0.697	0.750	0.786
10	0.576	0.671	0.726	0.763
11	0.553	0.618	0.703	0.741
12	0.532	0.627	0.683	0.722
13	0.514	0.608	0.664	0.703
14	0.497	0.590	0.646	0.686
15	0.482	0.574	0.630	0.670

表 Ⅳ—2　　$\alpha = 0.05$

n \ m	1	2	3	4	5	6	7	8	9
1	161	200	216	225	230	234	237	239	241
2	18.5	19.0	19.2	19.2	1.93	19.3	19.4	19.4	1.94
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59

四、预测、控制和风险分析

1.预测： $\hat{y}_0 = \hat{a} + \hat{b}x_0$

2.风险分析： 给定α，求δ，使得$P(\hat{y}_0 - \delta < y_0 < \hat{y}_0 + \delta) = 1 - \alpha$

即：y_0落在$(\hat{y}_0 - \delta, \hat{y}_0 + \delta)$之外的风险为$\alpha$.

其中：
$$\delta = \sqrt{F_\alpha(1, n-2) \cdot \hat{\sigma}^2 \cdot \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$$

或者
$$\delta = t_\alpha(n-2) \cdot \hat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

其中：
$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2$$

证明：$E(\hat{y}_0 - y_0)=E(\hat{a}+\hat{b}x_0\text{-}a-bx_0\text{-}\varepsilon_0)=0$。　　$\hat{y}_0$与y_0不相关，$D(y_0)=\sigma^2$，

$$D(\hat{y}_0)=D(\hat{a}+\hat{b}x_0)=D\left[\sum(\frac{1}{n}-\bar{x}k_i)y_i+\sum k_i y_i x_0\right]=D\left[\sum(\frac{1}{n}-\bar{x}k_i+k_i x_0)y_i\right]=\sigma^2\sum\left[\frac{1}{n}+(x_0-\bar{x})k_i\right]^2$$

$$=\sigma^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]$$

$$D(y_0-\hat{y}_0)=D(y_0)+D(\hat{y}_0)=\sigma^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]\qquad \therefore y_0-\hat{y}_0\sim N\left(0,\sigma^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]\right)$$

由于σ^2未知，用$\hat{\sigma}^2=\frac{1}{n-2}\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)^2$代替，则　$t=\dfrac{y_0-\hat{y}_0}{\hat{\sigma}\cdot\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}}\sim t(n-2)$

对于给定的显著性水平α，由t分布查得$t_\alpha(n-2)$

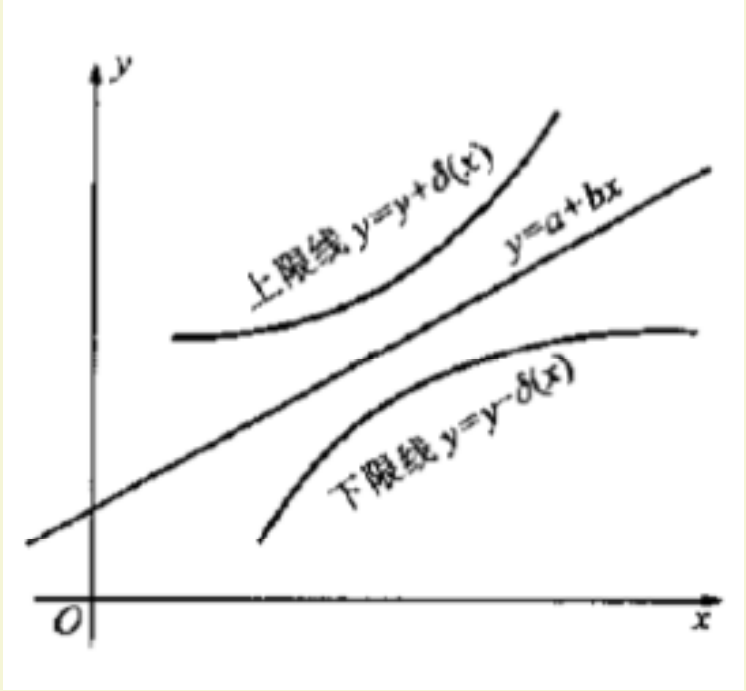
影响预测区间大小的因素：

(1) $\hat{\sigma}$：$\hat{\sigma}$越小，预测精度越高

(2) n：n越大，δ越小，预测精度越高

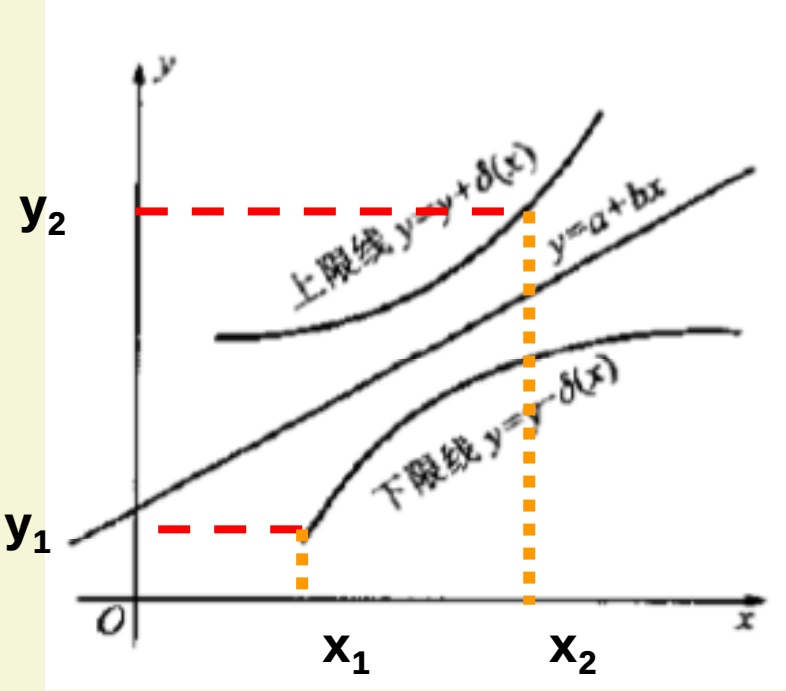
(3) $\sum_{i=1}^{n}(x_i-\bar{x})^2$:$\sum_{i=1}^{n}(x_i-\bar{x})^2$越大，$\delta$越小，预测精度越高

(4) $(x_0-\bar{x})^2$：x_0离$\bar{x}$越，预测精度越低



3.控制问题： 给定α，要使 y 以$1-\alpha$ 的概率落在$[y_1, y_2]$中，求 x 的控制范围$[x_1, x_2]$。

解法：$\begin{cases} \hat{y}-\delta(x_1)=y_1 & \text{解出} x_1 \\ \hat{y}+\delta(x_2)=y_2 & \text{解出} x_2 \end{cases}$



注：近似解法 $E(y_0-\hat{y}_0)=0$，$D(y_0-\hat{y}_0)=\hat{\sigma}^2\cdot\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right]\approx\hat{\sigma}^2$

因此： $y_0-\hat{y}_0\sim N(0,\hat{\sigma}^2)$

有性质：$P(\hat{y}_0-\hat{\sigma}\le y_0\le\hat{y}_0+\hat{\sigma})=68.3\%$

$P(\hat{y}_0-2\hat{\sigma}\le y_0\le\hat{y}_0+2\hat{\sigma})=95.4\%$

$P(\hat{y}_0-3\hat{\sigma}\le y_0\le\hat{y}_0+3\hat{\sigma})=99.7\%$

若要使y以68.3%的概率落在$[y_1,y_2]$中，那么由下列方程

$$\begin{cases}\hat{y}-\hat{\sigma}=\hat{a}+\hat{b}x_1-\hat{\sigma}=y_1 & \text{解出}x_1\\ \hat{y}+\hat{\sigma}=\hat{a}+\hat{b}x_2+\hat{\sigma}=y_2 & \text{解出}x_2\end{cases}$$

五、举例

例1、已知市场上某种商品的销售量 y 与价格 x 的关系为

$$\hat{y} = 20 - 0.6\hat{x} \quad \hat{\sigma}^2 = 2.25$$

若要使销售量 y 以95%的可能性落在(10, 15)内，试问价格应控制在什么范围？

解： $\hat{\sigma} = 1.5$

$$\begin{cases} 10 = 20 - 0.6x_1 - 3 \\ 15 = 20 - 0.6x_2 + 3 \end{cases}$$

解得 $x_1 = 11.6$， $x_2 = 13.3$

所以价格应控制在(11.6 , 13.3)内。

例 5．假设模型为 $Y_t = \alpha + \beta X_t + \mu_t$。给定 n 个观察值 (X_1, Y_1)，(X_2, Y_2)，…，(X_n, Y_n)，按如下步骤建立 β 的一个估计量：在散点图上把第 1 个点和第 2 个点连接起来并计算该直线的斜率；同理继续，最终将第 1 个点和最后一个点连接起来并计算该条线的斜率；最后对这些斜率取平均值，称之为 $\hat{\beta}$，即 β 的估计值。

（1）画出散点图，给出 $\hat{\beta}$ 的几何表示并推出代数表达式。

（2）计算 $\hat{\beta}$ 的期望值并对所做假设进行陈述。这个估计值是有偏的还是无偏的？解释理由。

（3）证明为什么该估计值不如我们以前用 OLS 方法所获得的估计值，并做具体解释。

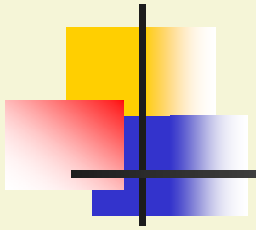
首先计算每条直线的斜率并求平均斜率。连接(X_1,Y_1)和(X_t,Y_t)的直线斜率为$(Y_t-Y_1)/(X_t-X_1)$。由于共有$n-1$条这样的直线，因此

$$\hat{\beta}=\frac{1}{n-1}\sum_{t=2}^{t=n}[\frac{Y_t-Y_1}{X_t-X_1}]$$

（2）因为 X 非随机且$E(\mu_t)=0$，因此

$$E[\frac{Y_t-Y_1}{X_t-X_1}]=E[\frac{(\alpha+\beta X_t+\mu_t)-(\alpha+\beta X_1+\mu_1)}{X_t-X_1}]=\beta+E[\frac{\mu_t-\mu_1}{X_t-X_1}]=\beta$$

这意味着求和中的每一项都有期望值β，所以平均值也会有同样的期望值，则表明是无偏的。



（3）根据高斯－马尔可夫定理，只有 β 的 OLS 估计量是最付佳线性无偏估计量，因此，这里得到的 $\hat{\beta}$ 的有效性不如 β 的 OLS 估计量，所以较差。