线性回归

hongqianglv@xjtu.edu.cn
西安交通大学

回归分析概述

一、变量间的关系及回归分析的基本概念

1、变量间的关系

变量之间的关系，大体可分为两类：

（1）确定性关系或函数关系：研究的是确定现象非随机变量间的关系。

（2）统计依赖或相关关系：研究的是非确定现象随机变量间的关系。

函数关系：

$$\text{圆面积} = f(\pi, \text{半径}) = \pi \cdot \text{半径}^2$$

统计依赖关系/统计相关关系：

$$\text{农作物产量} = f(\text{气温}, \text{降雨量}, \text{阳光}, \text{施肥量})$$

对变量间统计依赖关系的考察主要是通过回归分析(regression analysis)或相关分析(correlation analysis)来完成的：

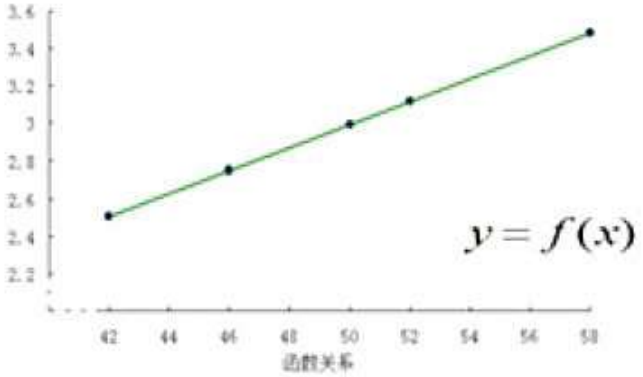
回归分析：自变量是可测和可控的非随机变量

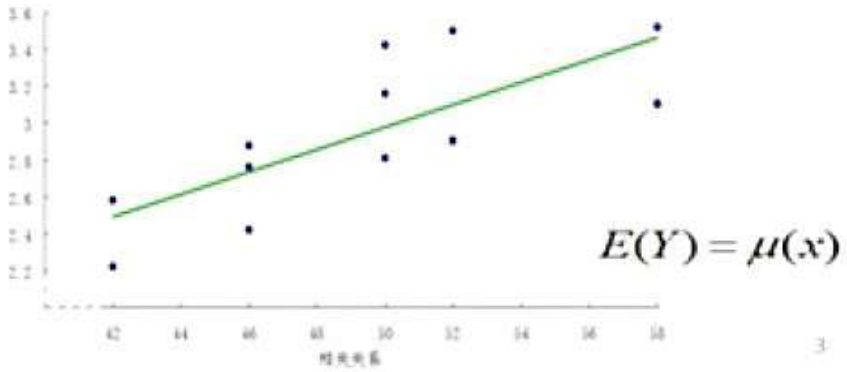
相关分析：自变量也是随机变量或者不可控变量

一元线性回归

相关性关系：

变量之间的关系并不确定，而是表现为具有随机性的一种“趋势”。即对自变量x的同一值，在不同的观测中，因变量Y可以取不同的值，而且取值是随机的，但对应x在一定范围的不同值，对Y进行观测时，可以观察到Y随x的变化而呈现有一定趋势的变化。





一元线性回归

- 如：身高与体重，不存在这样的函数可以由身高计算出体重，但从统计意义上来说，身高者，体也重。
- 如：父亲的身高与儿子的身高之间也有一定联系，通常父亲高，儿子也高。

一元线性回归

我们以一个例子来建立回归模型

- 例1：根据2013年《中国统计年鉴》的数据，2012年中国各地区城镇居民人均年消费支出和可支配收入数据见下表.

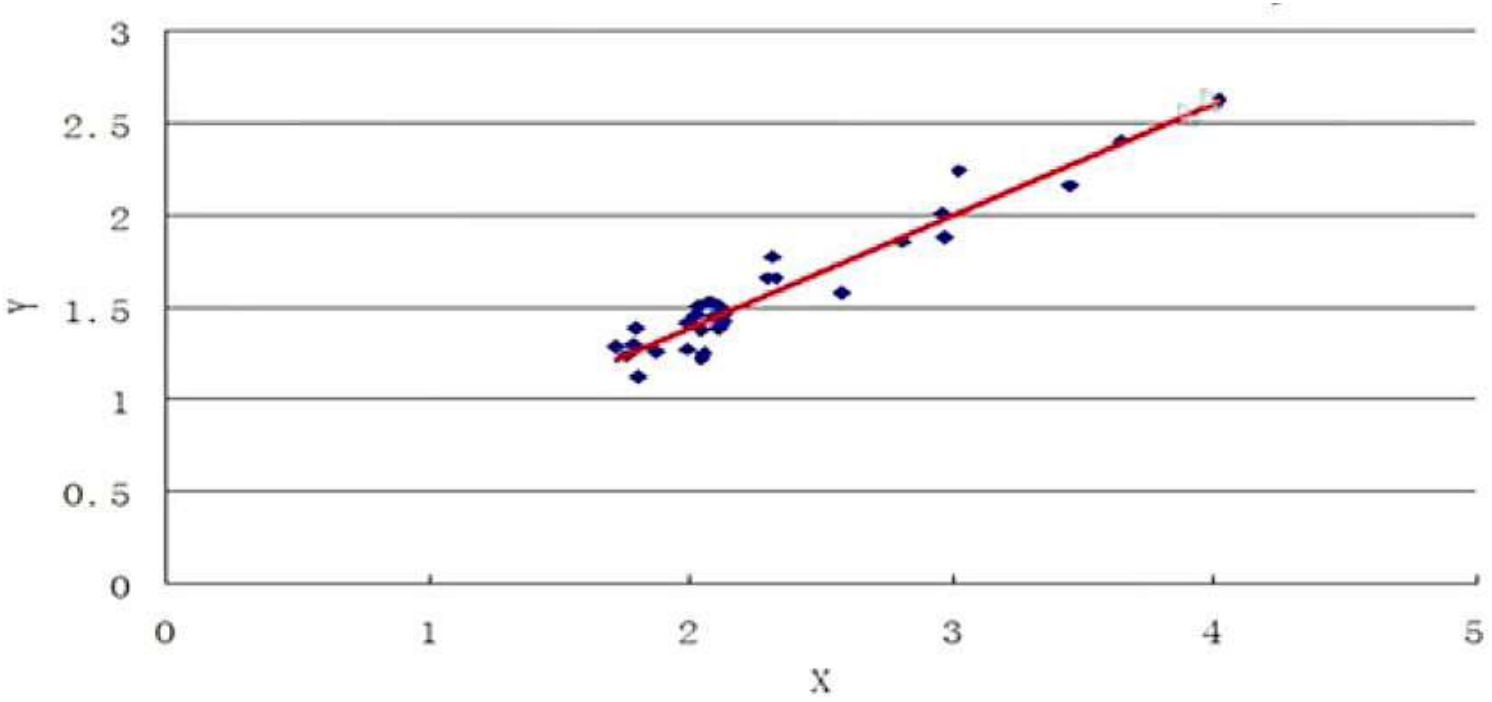
一元线性回归

地区	可支配收入x(万元)	消费支出y(万元)	地区	可支配收入x(万元)	消费支出y(万元)
北京	3.647	2.405	上海	4.019	2.625
天津	2.963	2.002	江苏	2.968	1.883
河北	2.054	1.253	浙江	3.455	2.155
山西	2.041	1.221	安徽	2.102	1.501
内蒙古	2.315	1.772	福建	2.806	1.859
辽宁	2.322	1.659	江西	1.986	1.278
吉林	2.021	1.461	山东	2.576	1.578
黑龙江	1.776	1.298	河南	2.044	1.373

一元线性回归

地区	可支配收入x（万元）	消费支出y（万元）	地区	可支配收入x（万元）	消费支出y（万元）
湖北	2.084	1.450	云南	2.107	1.388
湖南	2.132	1.461	西藏	1.803	1.118
广东	3.023	2.240	陕西	2.073	1.533
广西	2.124	1.424	甘肃	1.716	1.285
海南	2.092	1.446	青海	1.757	1.235
重庆	2.297	1.657	宁夏	1.983	1.407
四川	2.031	1.505	新疆	1.792	1.389
贵州	1.870	1.259			

一元线性回归



散点图　X：可支配收入，
Y：消费支出

一元线性回归

可支配收入x的变化是引起消费支出Y的变化的主要因素，其他因素的影响是次要的.

从散点图看出，引起消费支出Y的变化的主要部分可以表示为

$\mu(x) = a + bx$，其中a，b是未知参数.

另一部分是由其他随机因素引起的，记为ε

即$Y = a + bx + \varepsilon$.

一元线性回归

- 一般概率模型：Y=确定性成分+随机误差，Y为因变量。
- 确定性成分= E(Y)，随机误差的均值等于 0, 则Y= E(Y)+ ε
- 若E(Y)与自变量 x 之间为线性关系时， E(Y)= $a+bx$
- 一元线性回归模型：$Y = a + bx + \varepsilon$
  - 模型中，Y是 x 的线性函数(部分)加上误差项
  - 线性部分反映了由于 x 的变化而引起的Y的变化
  - 误差项 ε 是随机变量
    - 反映了除 x 和 Y 之间的线性关系之外的随机因素对 Y 的影响
    - 是不能由 x 和 Y之间的线性关系所解释的变异性
  - a和 b 称为模型的参数

一元线性回归

1. **线性性**：Y与X存在线性相关关系，即

   $E(Y)=a+bx$

2. **正态性**：误差项ε是一个服从数学期望为0的正态分布的随机变量，即$\varepsilon \sim N(0,\sigma^2)$，$E(\varepsilon)=0$

3. **均等性**：对于所有的 x 值，ε的方差σ^2 都相同

4. **独立性**：对于每一个特定的 x 值，它所对应的ε与其他 x 值所对应的ε不相关，对于一个特定的 x 值，它所对应的 Y 值与其他 x 所对应的 Y值也不相关

一元线性回归

对从总体(x, Y)中抽取的一个样本

$(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$

一元线性回归模型：

$$\begin{cases} Y_i = a + bx_i + \varepsilon_i, i = 1, 2, \ldots, n, \\ \varepsilon_i \sim N(0, \sigma^2), \text{且相互独立}, \\ a, b \text{(回归系数)}, \sigma^2 \text{未知}. \end{cases}$$

一元线性回归

根据样本

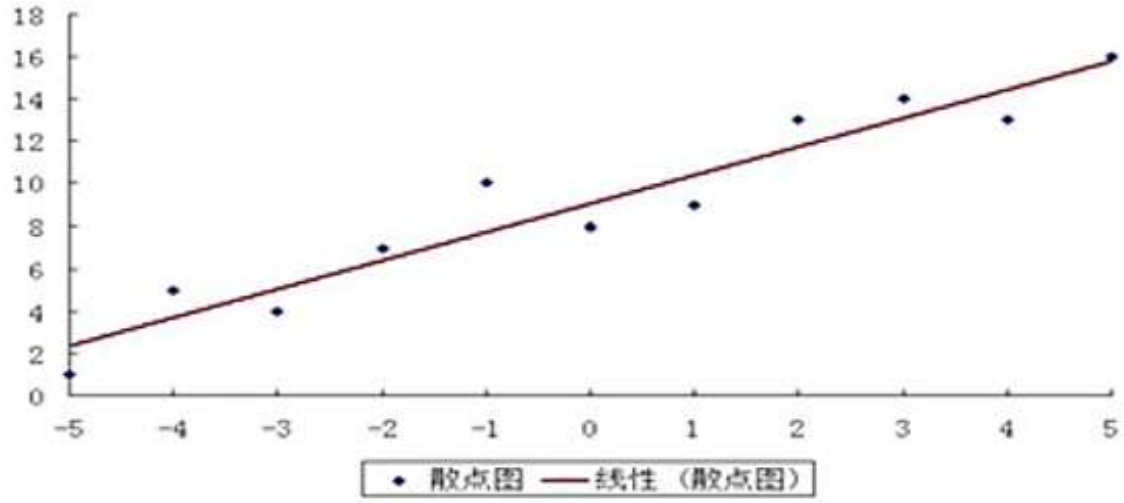
估计a，b，

记为$\hat{a}$，$\hat{b}$，

称为y关于x

一元线性回归

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



一元线性回归

一元线性回归要解决的问题：

参数估计：
- (1) a, b 的估计；
- (2) σ^2 的估计；

参数检验及模型应用：
- (3) 线性假设的显著性检验；
- (4) 回归系数 b 的置信区间；
- (5) Y 的点预测.

一元线性回归

a和b 的最小二乘估计

使因变量的观察值与估计值之间的离差平方和达到最小来求得 $\hat{a}$ 和 $\hat{b}$ 的方法。

$$Q(a,b)=\sum_{i=1}^{n}(y_i-a-bx_i)^2 \text{取最小值.}$$

根据

$$\left.\begin{aligned}\frac{\partial Q}{\partial a}&=-2\sum_{i=1}^{n}(y_i-a-bx_i)=0\\ \frac{\partial Q}{\partial b}&=-2\sum_{i=1}^{n}(y_i-a-bx_i)x_i=0\end{aligned}\right\}$$

一元线性回归

得方程组

$$\begin{cases} na+(\sum_{i=1}^{n}x_i)b=\sum_{i=1}^{n}y_i \\ (\sum_{i=1}^{n}x_i)a+(\sum_{i=1}^{n}x_i^2)b=\sum_{i=1}^{n}x_iy_i \end{cases}$$

由于x_i不全相同， 方程组的系数行列式

$$\begin{vmatrix} n & \sum_{i=1}^{n}x_i \\ \sum_{i=1}^{n}x_i & \sum_{i=1}^{n}x_i^2 \end{vmatrix} = n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2=n\sum_{i=1}^{n}(x_i-\bar{x})^2\neq 0$$

一元线性回归

$$\hat{b}=\frac{n\sum_{i=1}^{n}x_iy_i-(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)}{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

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

其中 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\quad \bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i.$

在得到a,b的估计$\hat{a},\hat{b}$后，对于给定的x，

一元线性回归

方程

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

称为Y关于x的经验回归方程，简称回归方程.

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

对于样本值$(x_1, y_2), (x_2, y_2), \cdots (x_n, y_n)$，回归直线通过散点图的几何中心$(\bar{x}, \bar{y})$.

一元线性回归

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

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

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

$$=\sum_{i=1}^{n}x_iy_i-\frac{1}{n}(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i),$$

$$\hat{b}=\frac{S_{xy}}{S_{xx}},\qquad \hat{a}=\frac{1}{n}\sum_{i=1}^{n}y_i-(\frac{1}{n}\sum_{i=1}^{n}x_i)\hat{b}.$$

一元线性回归

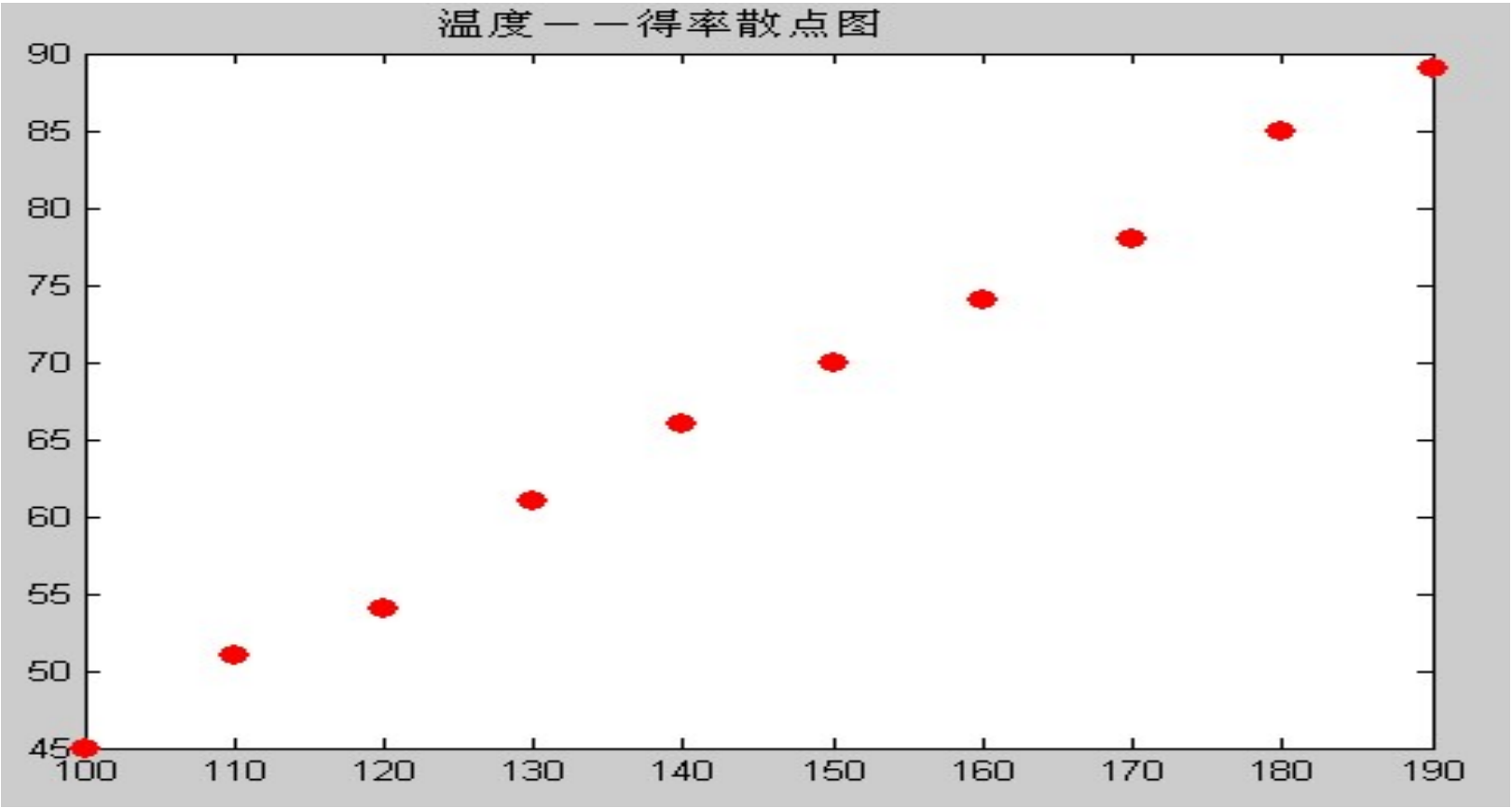
例 为研究某一化学反应过程中，温度$x(^0C)$对产品得率$Y(\%)$的影响，测得数据如下 .

温度$x(^oC)$	100	110	120	130	140	150	160	170	180	190
得率$Y(\%)$	45	51	54	61	66	70	74	78	85	89

这里自变量x是普通变量， Y是随机变量.

画出散点图如下，

一元线性回归



观察散点图，$\mu(x)$具有线性函数$a+bx$的形式.

一元线性回归

随机变量 Y 符合一元线性回归模型所述的条件，求 Y 关于 x 的线性回归方程 .

	x	y	x^2	y^2	xy
	100	45	10000	2025	4500
	110	51	12100	2601	5610
	120	54	14400	2916	6480
	130	61	16900	3721	7930
	140	66	19600	4356	9240
	150	70	22500	4900	10500
	160	74	25600	5476	11840
	170	78	28900	6084	13260
	180	85	32400	7225	15300
	190	89	36100	7921	16910
Σ	1450	673	218500	47225	101570

一元线性回归

解： $S_{xx}=218500-\frac{1}{10}\times1450^2=8250$

$$S_{xy}=101570-\frac{1}{10}\times1450\times673=3985$$

$$\hat{b}=S_{xy}/S_{xx}=0.48303$$

$$\hat{a}=\frac{1}{10}\times673-\frac{1}{10}\times1450\times0.48303=-2.73935$$

回归直线方程 $\hat{y}=-2.73935+0.48303x$

一元线性回归

σ^2的估计

σ^2 越小,用回归函数 $\mu(x)=a+bx$ 作为 Y 的近似导致的均方误差就越小.

$$E\{[Y-(a+bx)]^2\}=E(\varepsilon^2)=D(\varepsilon)+[E(\varepsilon)]^2=\sigma^2.$$

利用回归函数 $\mu(x)=a+bx$

去研究随机变量 Y 与 x 的关系就愈有效

为了估计 σ^2,引入残差平方和

$$\hat{y}_i=\hat{y}\big|_{x=x_i}=\hat{a}+\hat{b}x_i\ ,$$

$y_i-\hat{y}_i$ 为 x_i 处的残差 .

残差平方和

$$Q_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2$$

它是经验回归函数在x_i处的函数值 $\widehat{\mu(x)} = \hat{a} + \hat{b}x$

与x_i处的观察值y_i的偏差的平方和.

$$Q_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}[y_i - \bar{y} - \hat{b}(x_i - \bar{x})]^2$$

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

一元线性回归

$$= S_{yy} - 2\hat{b}S_{xy} + (\hat{b})^2 S_{xx}$$

由 $\hat{b} = S_{xy} / S_{xx}$

$$Q_e = S_{yy} - (\hat{b})^2 S_{xx}.$$

σ^2 的无偏估计量为

$$\sigma^2 = \frac{Q_e}{n-2} = \frac{1}{n-2}\left[S_{yy} - (\hat{b})^2 S_{xx}\right].$$

一元线性回归

例　求上例中方差的无偏估计.

解　$$S_{yy}=\sum_{i=1}^{n}y_i^2-\frac{1}{n}(\sum_{i=1}^{n}y_i)^2$$

$$=47225-\frac{1}{10}\times 673^2$$

$$=1932.1$$

又知$S_{xx}=8250$,$\hat{b}=0.48303$

$$Q_e=S_{yy}-(\hat{b})^2S_{xx}=7.23$$

$$\hat{\sigma}^2=Q_e/(n-2)=7.23/8=0.90$$

可化为一元线性回归

方法——通过适当的变量变换,化成一元线性回归问题进行分析处理.

几种常见的可转化为一元线性回归的模型

1. $Y=\alpha e^{\beta x}\cdot\varepsilon,\qquad \ln\varepsilon\sim N(0,\sigma^2).$

其中α,β,σ^2是与x无关的未知参数.

将$Y=\alpha e^{\beta x}\cdot\varepsilon$两边取对数,

得 $$\ln Y=\ln\alpha+\beta x+\ln\varepsilon.$$

令$\ln Y=Y',\ln\alpha=a,\ \beta=b,\quad x=x',\quad \ln\varepsilon=\varepsilon'$

转化为一元线性回归模型：

$$Y'=a+bx'+\varepsilon', \quad \varepsilon'\sim N(0,\sigma^2).$$

2. $Y=\alpha+\beta h(x)+\varepsilon, \qquad \varepsilon\sim N(0,\sigma^2).$

其中α,β,σ^2是与x无关的未知参数.

$h(x)$是x的已知函数，

令$\alpha=a, \quad \beta=b, \quad h(x)=x'$,

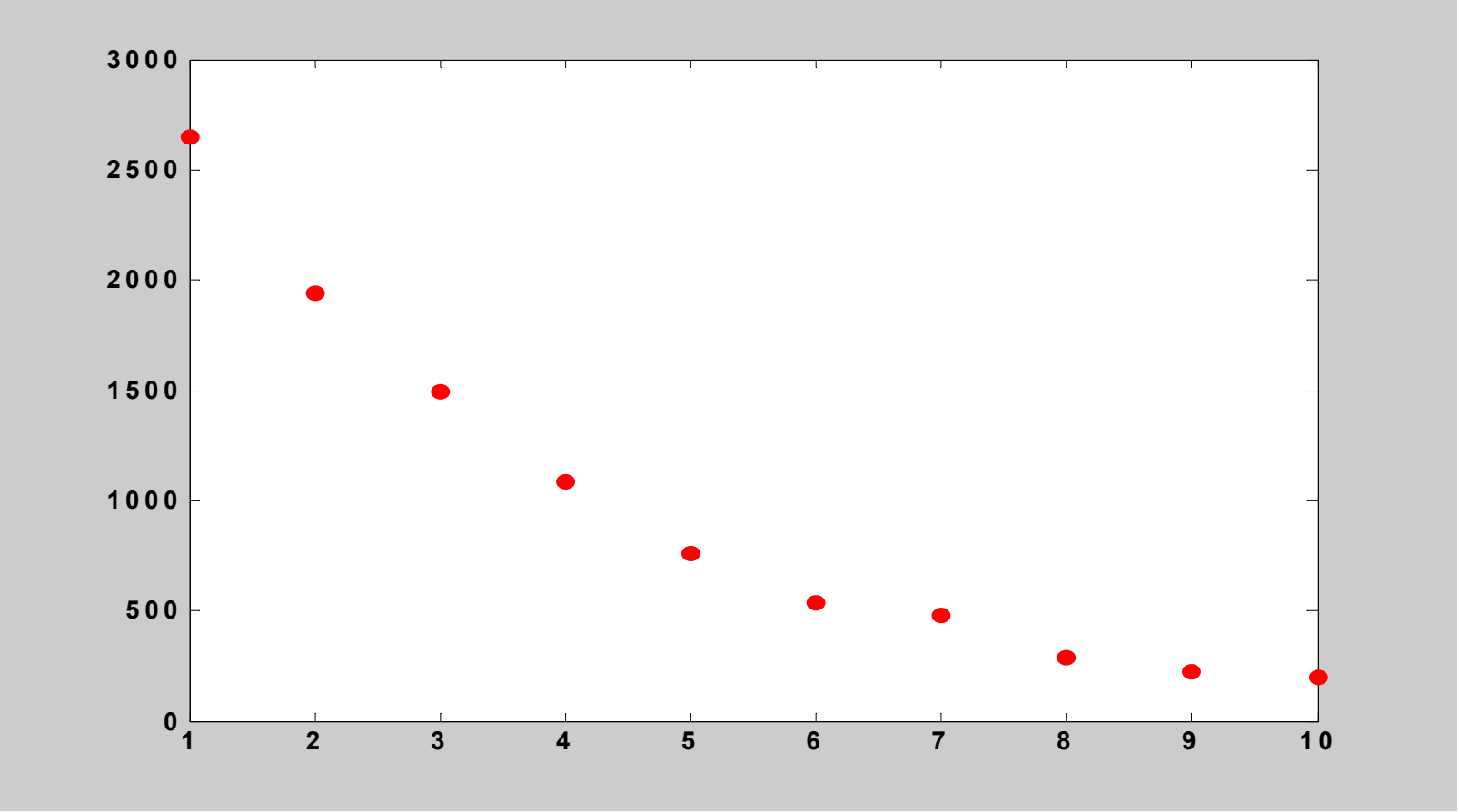
转化为一元线性回归模型：

$$Y'=a+bx'+\varepsilon, \quad \varepsilon\sim N(0,\sigma^2).$$

例 下表是 **1957** 年美国旧轿车价格的调查资料, 今以 ***x*** 表示轿车的使用年数, ***Y*** 表示相应的平均价格(以美元计), 求 ***Y*** 关于 ***x*** 的回归方程 .

年数x	1	2	3	4	5	6	7	8	9	10
价格Y	2651	1943	1494	1087	765	538	484	290	226	204

解　　做散点图，Y与x呈指数关系，



选择模型 $Y = \alpha e^{\beta x} \cdot \varepsilon, \quad \ln \varepsilon \sim N(0, \sigma^2).$

变量变换 $Y' = a + bx' + \varepsilon', \quad \varepsilon' \sim N(0, \sigma^2).$

其中 $\ln Y = Y', \ a = \ln \alpha, \ b = \beta, \ x' = x, \ \varepsilon' = \ln \varepsilon$

数据变换后得

$x' = x$	1	2	3	4	5
$y' = \ln y$	7.8827	7.5720	7.3092	6.9912	6.6399
$x' = x$	6	7	8	9	10
$y' = \ln y$	6.2879	6.1821	5.6699	5.4205	5.3181

经计算 $\hat{b} = -0.2977, \quad \hat{a} = 8.1646.$

$$\hat{y}' = -0.2977x' + 8.1646.$$

代回原变量,得曲线回归方程

$$\hat{y} = \exp(\hat{y}') = \exp(-0.2977x + 8.1646)$$

$$= 3514.3e^{-0.2977x}.$$