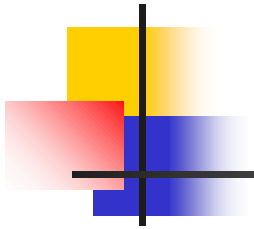
第十一章 简朴线性回归

Linear regression



- 回归是设法找出变量间在数量上的依存变化关系，用函数体现式体现出来，这个体现式称之为回归方程。

两变量间的关系

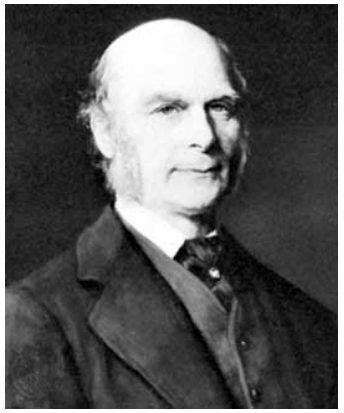
- 确定性关系：两变量间的函数关系
  - 圆的周长与半径的关系： C＝2 ⁄⁄R
  - 速度、时间与旅程的关系：L＝ST
  - X与Y的函数关系： Y＝a+bX
- 非确定性关系：两变量在宏观上存在关系，但并未精确到可以用函数关系来体现。
  - 青少年身高与年龄的关系；
  - 身高与体重的关系:原则体重(kg)=身高

一、线性回归的概念

- 当两个变量存在精确、严格的直线关系时，可以用Y=a+bX，表达两者的函数关系。
- 其中X 为自变量（independent variable）；Y是因变量（dependent variable ）。
- 但在实际生活当中，由于其他原因的干扰，许多双变量之间的关系并不是严格的函数关系，不能用函数方程来精确反应，为了区别于两变量间的函数方程，我们称这种关系为回归关系，用直线方程来表达这种关系称为回归直线或线性回归。

$$\hat{Y} = a + bx$$

小插曲：为何叫”回归“？



F. Galton

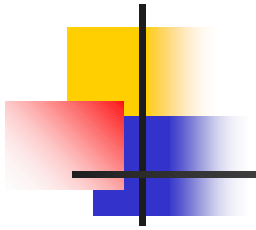
K.Pearson

二、回归参数的估计

$$\hat{Y} = a + bx$$

- 式中的 $\hat{Y}$ 是由自变量X推算应变量Y的估计值，a是回归直线在Y 轴上的截距；b为样本的回归系数，即回归直线的斜率，表达当X变动一种单位时，Y平均变动b个单位。
- 计算原理：最小二乘法，即保证各实测点到回归直线的纵向距离的平方和最小，并使计算出的回归方程最能代表实测数据所反应出的直线趋势。

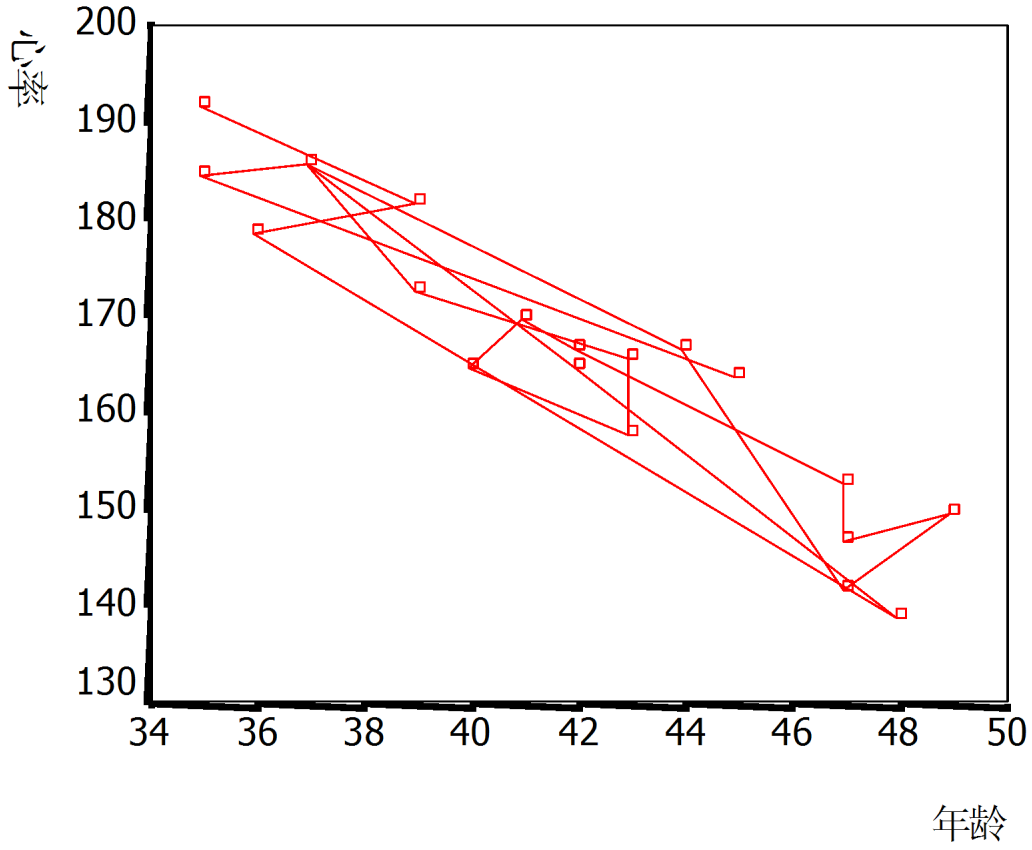
$$\sum\left(Y-\hat{Y}\right)^2=\sum\left[Y-\left(a+bX\right)\right]^2$$



$$b = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sum (X - \overline{X})^2} = \frac{l_{XY}}{l_{XX}}$$

$$a = \overline{Y} - b\overline{X}$$

例11-1　某医师为了研究正常成年男性的运动后最大心率与年龄的关系，测得20名正常成年男性的有关数据，散点图如下。



年龄与运动后最大心率的回归方程

$\bar{X}$=41.8　　$\bar{Y}$ = 166.8

l_{XX} = 381.2　　l_{YY} = 4477.2　　l_{XY} = - 1226.8

$$b = \frac{l_{XY}}{l_{XX}} = \frac{-\ 1226.8}{381.2} = -\ 3.218$$

a = 166.8-(-3.218)′ 41.8 = 301.3124

$$\hat{Y} = 301.3124 - 3.218X$$

回归系数和回归方程的意义及性质

$$\hat{Y} = a + bX$$

- b 的意义
- a 的意义
- $\hat{Y}$ 的意义
- $\hat{Y} - Y$ 的意义
- $\sum_{i=1}^{n} (\hat{Y}_i - Y_i)$ 的意义

b 的意义

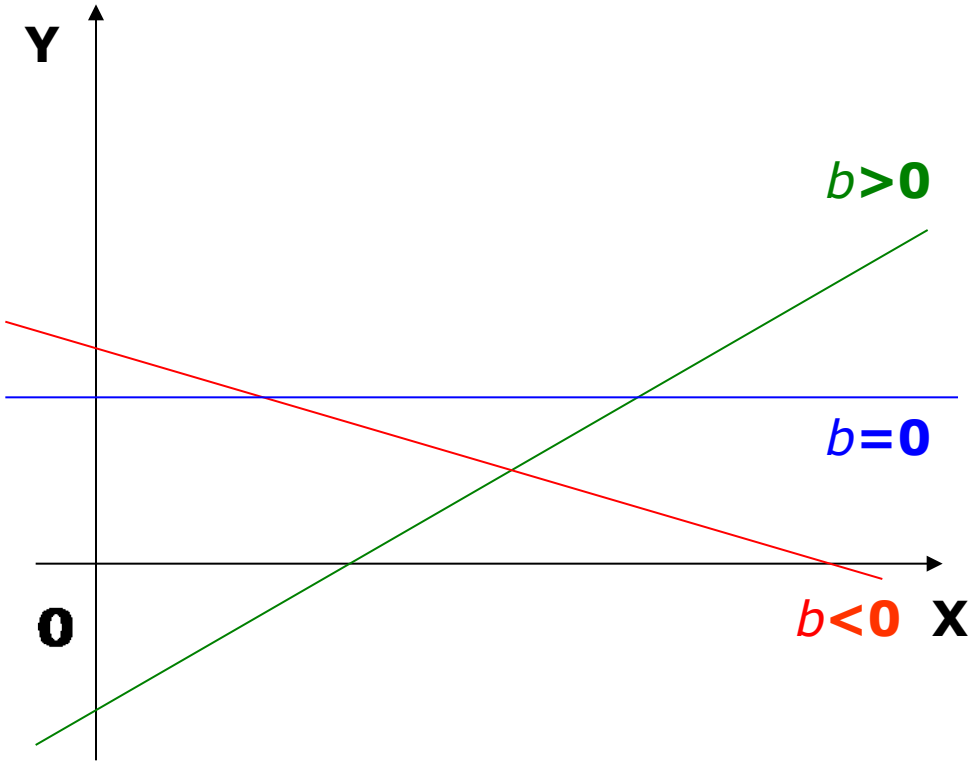
- 斜率(slope)
- $\hat{Y}$ =301.3124 -3.218 X
- 年龄每增长 1 岁，其运动后最大心率平均减少 3.218（次/分钟)
- b 的单位为 (Y的单位/X的单位)

b is the regression coefficient and the slope of the line 。

b>0，y increase with the increase of X

b<0，y decrease with the increase of X

b=0，no linear correlation between two variables.



statistical significance of *b*： when X changed a unit , the Y changed b units on average.

a 的意义

$$\hat{Y} = a + bX$$

- a 截距(intercept, constant)
- X=0 时，Y的估计值
- a的单位与Y值相似
- 当X也许取0时，a才有实际意义。

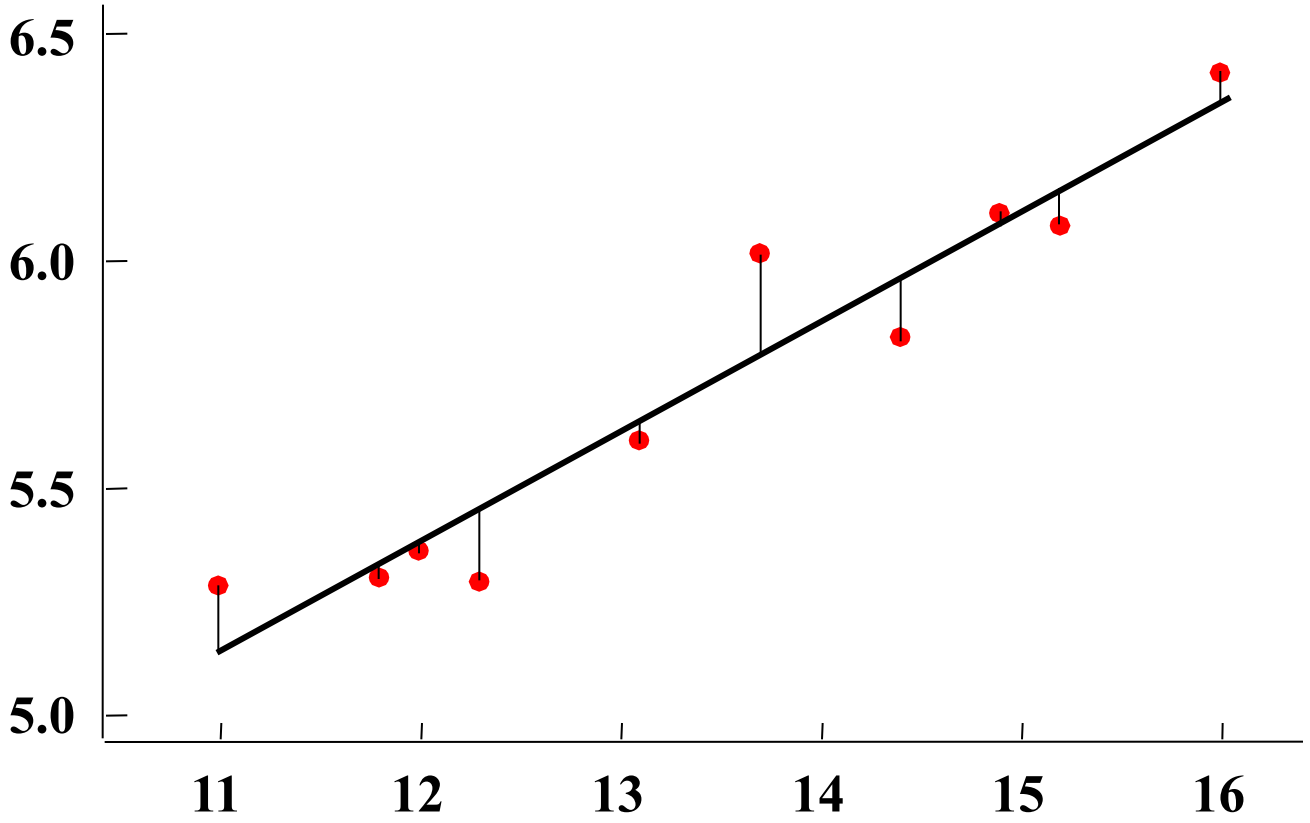
估计值 $\hat{Y}$ 的意义

- X=46时，$\hat{Y}$ =153.2844,

  即年龄为 46岁 的正常成年男性, 其平均运动后最大心率估计值为 153.2844 （次/分钟）;
- 给定X时，Y的估计值。
- 当 $X=\overline{X}$ 时，$\hat{Y}=\overline{Y}$

$\hat{Y}-Y$ 的意义

- $\hat{Y}-Y$ 为残差：实测点到回归直线的纵向距离。



$\sum(\hat{Y}-Y)^2$ 的意义

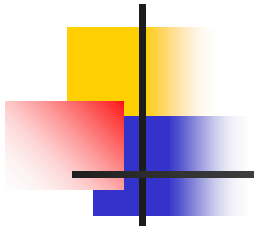
- 残差平方和 (residual sum of squares).
- 综合表达点距直线的纵向距离。
- 在所有的直线中，回归直线的残差平方和是最小的。(最小二乘)

三、总体回归系数的假设检查

- 与直线有关同样，直线回归方程也是从样本资料计算而得的，同样也存在着抽样误差问题。因此，需要对样本的回归系数**b**进行假设检查，以判断**b**与否从回归系数为零的总体中抽得。总体的回归系数用**β**表达。

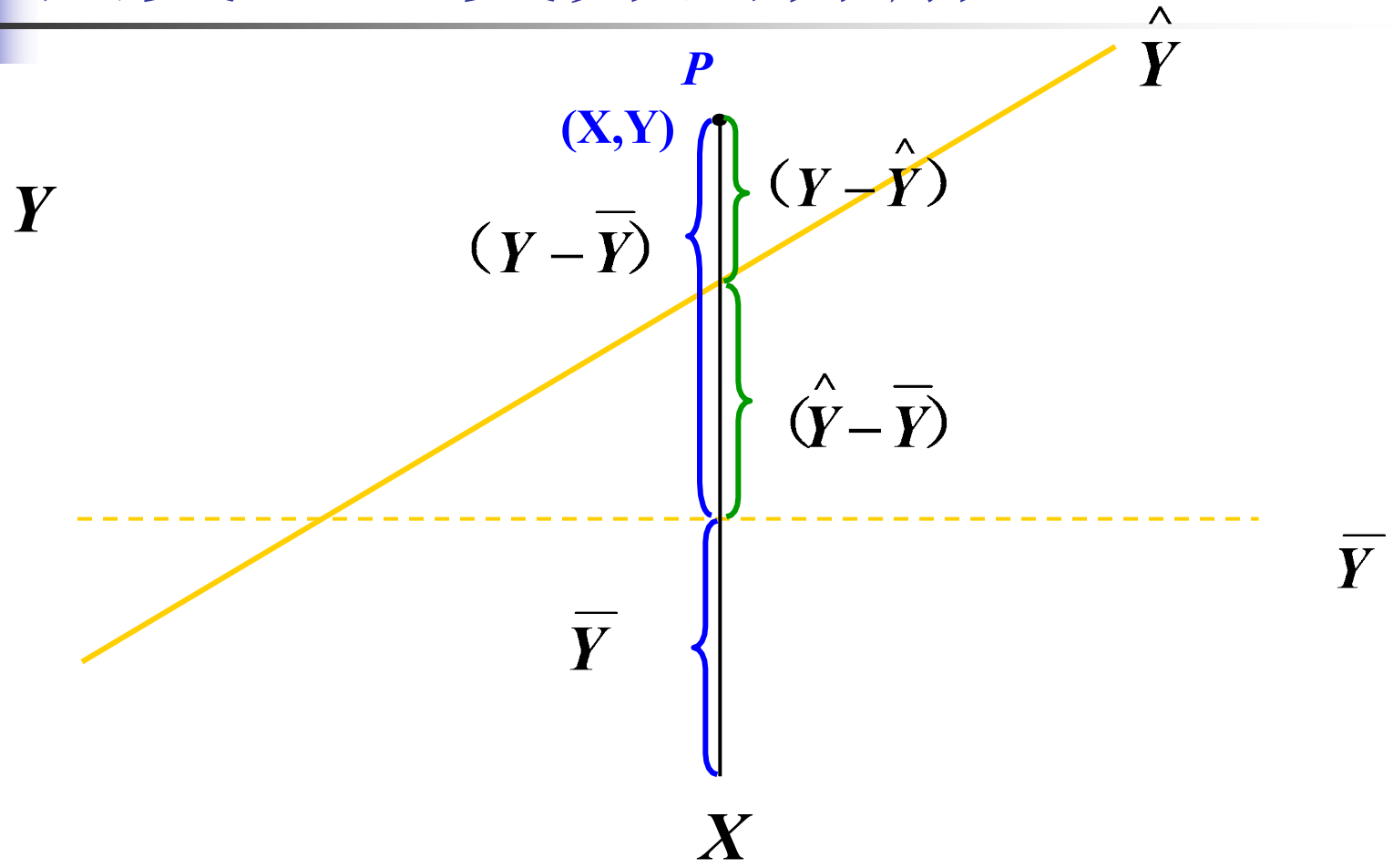
一般环节

1. **H0：β=0　回归方程无意义**
2. **H1：β≠0　回归方程故意义**
3. **α=0.05**
4. **选择合适的假设检查措施（方差分析或t检查），计算记录量**
5. **计算概率值P**
6. **做出推论：记录学结论和专业结论**



方差分析法

因变量总变异的分解


$\hat{Y}$
P
(X,Y)
Y
$(Y-\hat{Y})$
$(Y-\overline{Y})$
$(\hat{Y}-\overline{Y})$
$\overline{Y}$
$\overline{Y}$
X

Y的总变异分解

- 未引进回归时的总变异： $\sum (Y - \bar{Y})^2$
- (sum of squares about the mean of Y)

$$\sum (Y - \hat{Y})^2$$

- 引进回归后来的变异(剩余):
- (sum of squares about regression)—

$$\sum (\hat{Y} - \bar{Y})^2$$

- 回归的奉献，回归平方和：
- (sum of squares due to regression)

Y的总变异分解

$$\sum\left(Y-\bar{Y}\right)=\sum\left(\hat{Y}-\bar{Y}\right)^2+\sum\left(Y-\hat{Y}\right)^2$$

$$SS_{总}=SS_{回}+SS_{剩}$$

$$\nu_{总}=\nu_{回}+\nu_{剩}$$

剩余原则差

$$s_{Y\cdot X}=\sqrt{\frac{\sum\left(Y-\hat{Y}\right)}{n-2}}$$

(1) 扣除了X的影响后Y方面的变异;
(2) 引进 回归方程后，Y方面的变异。

回归系数检查的基本思想

- 假如X与Y无线性回归关系，在SS回归和SS剩余都是其他随机原因对Y的影响，由此，MS回归≈MS剩余，总体回归系数β=0，反之， β ≠ 0。因此用F检查对X与Y之间有无回归关系进行检查。

公式

$$SS_{总}=\sum\left(Y-\overline{Y}\right)^2=\sum Y^2-\frac{\left(\sum Y\right)^2}{n}$$

$$SS_{回归}=\sum\left(\hat{Y}-\overline{Y}\right)^2=bl_{xy}=\frac{l_{xy}^2}{l_{xx}}$$

$$SS_{剩余}=SS_{总}-SS_{回归}$$

$$v_{总}=n-1 \qquad v_{回归}=1 \qquad v_{剩余}=n-2$$

$$MS_{回归}=\frac{SS_{回归}}{v_{回归}} \qquad MS_{剩余}=\frac{SS_{剩余}}{v_{剩余}}$$

$$F=\frac{MS_{回归}}{MS_{剩余}}$$

H_0：β=0

H_1：β≠0

α=0.05

$$SS_{总} = \sum (Y - \overline{Y})^2 = 4477.2$$

$$SS_{回归} = \sum (\hat{Y} - \overline{Y})^2 = 39481591$$

$$SS_{剩余} = SS_{总} - SS_{回归} = 529.0409$$

$$F = \frac{MS_{回归}}{MS_{剩余}} = \frac{SS_{回归} / v_{回归}}{SS_{剩余} / v_{剩余}} = 134.3313$$

查F界值表，$F_{0.05（1，18）}$=4.41，F> $F_{0.05（1，18）}$，P<0.05，拒绝H_0