# 回归分析:最小二乘法

学业辅导中心

## 1 为什么要学习回归分析

研究收入如何受个体教育程度, 行业, 性别等因素影响, 要用到“回归分析”, 它与计量经济是十分类似的.

线性回归模型是现代统计学中应用最为广泛的模型之一，它也是其它统计模型研究或应用的基础。这主要有下列几个原因:

1. 在实际问题中，变量之间的关系常具有线性或近似线性的依赖关系。
2. 在现实世界中，虽然许多变量间的关系是非线性的，但经过适当的变换，将会成为线性关系。
3. 线性关系是变量之间最简单的关系，容易处理，理论和方法比较完善，这些为实际应用提供了有效算法。

## 2 单变量普通最小二乘

### 2.1 最简单的情况

现在我们有$n$个数据点$(x_i, y_i), i = 1, \cdots, n$. 我们的目标是: 拟合数据$(x_i, \hat{y}_i = \hat{\alpha} + \hat{\beta} x_i, i = 1, \cdots, n)$ Gauss给出来一个“最佳”拟合的办法:最小二乘法.

$$(\hat{\alpha}, \hat{\beta}) = \operatorname{argmin} \frac{1}{n} \sum_{i=1}^{n} \underbrace{(y_i - \alpha - \beta x_i)^2}_{\text{“}misfits\text{''}}$$

在求一阶导后,

$$\begin{cases} -\dfrac{2}{n} \displaystyle\sum_{i=1}^{n} (y_i - \hat{\alpha} - \hat{\beta} x_i) & = 0, \\ -\dfrac{2}{n} \displaystyle\sum_{i=1}^{n} x_i (y_i - \hat{\alpha} - \hat{\beta} x_i) & = 0. \end{cases} \quad \Rightarrow \quad \begin{array}{rcl} \hat{\beta} & = & \dfrac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2} \\ \hat{\alpha} & = & \bar{y} - \hat{\beta} \bar{x} \end{array}$$

**注记.** 这里不用$PPT$中的$SS_{xx}, SS_{xy}$等记号是为了方便记忆, 在计量经济中, 回归系数被记为

$$\beta_1 = \frac{\operatorname{cov}(Y_i, X_i)}{V(X_i)}.$$

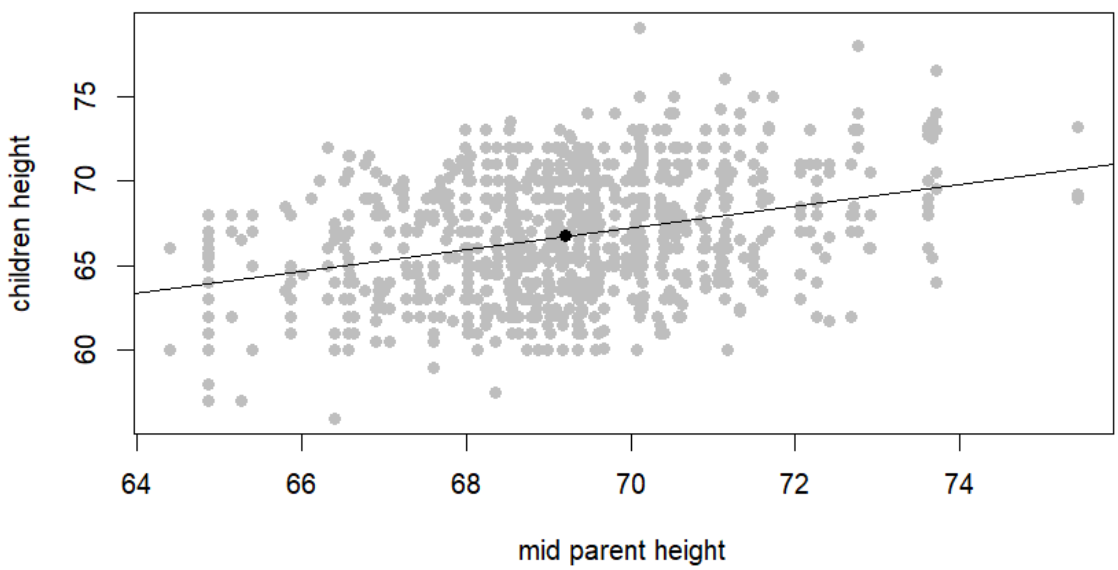
怎么方便怎么来.

## 2.2 过原点回归

$$
\begin{aligned}
& y = \hat{\alpha} + \hat{\beta} x = \bar{y} - \hat{\beta}\bar{x} + \hat{\beta} x \\
& \Rightarrow y - \bar{y} = \hat{\beta}(x - \bar{x}) \\
& \Rightarrow y - \bar{y} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2}(x - \bar{x}) = \frac{\hat{\rho}\hat{\sigma}_x\hat{\sigma}_y}{\hat{\sigma}_x^2}(x - \bar{x}) \\
& \Rightarrow \frac{y - \bar{y}}{\hat{\sigma}_y} = \hat{\rho}\frac{x - \bar{x}}{\hat{\sigma}_x},
\end{aligned}
$$

## 2.3 Galton's regression

This data set lists the individual observations for 934 children in 205 families on which Galton (1886) based his cross-tabulation.




拟合直线:$y = 22.64 + 0.64x$.

```r
install.packages("HistData")
library(HistData)
x <- GaltonFamilies$midparentHeight
y <- GaltonFamilies$childHeight

x.mean <- mean(x)
y.mean <- mean(y)
x.sd <- sd(x)
y.sd <- sd(y)
rho.xy <- cor(x,y)

beta.hat <- rho.xy * y.sd / x.sd
```

```
alpha.hat <- y.mean - x.mean * beta.hat

alpha.hat

beta.hat

fit <- lm(y ~ x)
summary(fit)


plot(x, y, xlab = "mid parent height", ylab = "children height", pch = 21, bg = "grey", col = "
    grey")
abline(fit)
points(x.mean, y.mean, pch = 16)
```

# 3 多变量普通最小二乘

## 3.1 求导解系数

$$Y=\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad X=\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}=\begin{pmatrix} x_1^{\mathrm{T}} \\ x_2^{\mathrm{T}} \\ \vdots \\ x_n^{\mathrm{T}} \end{pmatrix}=(X_1,\ldots,X_p)$$

目标: 找最佳线性拟合.

$$\hat{\beta}=\arg\min_b n^{-1}\sum_{i=1}^n (y_i-x_i^\top b)^2=\arg\min_b n^{-1}\|Y-Xb\|^2$$

这里$\hat{\beta}$是最小二乘系数, $\hat{y}_i$是拟合值, $\hat{\epsilon}_i=y_i-\hat{y}_i$是残差.

**注记** (常用向量求导法则)**.** 假设 $\boldsymbol{X}$ 为 $n\times m$ 矩阵, $y=f(\boldsymbol{X})$ 为 $\boldsymbol{X}$ 的一个实值函数, 矩阵

$$\frac{\partial y}{\partial \boldsymbol{X}}=\begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \frac{\partial y}{\partial x_{12}} & \cdots & \frac{\partial y}{\partial x_{1m}} \\ \frac{\partial y}{\partial x_{21}} & \frac{\partial y}{\partial x_{22}} & \cdots & \frac{\partial y}{\partial x_{2m}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial y}{\partial x_{n1}} & \frac{\partial y}{\partial x_{n2}} & \cdots & \frac{\partial y}{\partial x_{nm}} \end{bmatrix}$$

称为 $y$ 对 $\boldsymbol{X}$ 的微商.

特别的,

1. 设 $a,x$ 均为$n\times 1$ 向量, $y=a^\top x$, 则$\dfrac{\partial y}{\partial x}=a$.
2. 设 $A_{n\times n}$对称.$x_{n\times 1}$,$y=x^\top Ax$ ,则$\dfrac{\partial y}{\partial x}=2Ax$.

证明如下.

$$\frac{\partial a^{\mathrm{T}}x}{\partial x}=\begin{pmatrix}\frac{\partial a^{\mathrm{T}}x}{\partial x_1}\\ \vdots\\ \frac{\partial a^{\mathrm{T}}x}{\partial x_p}\end{pmatrix}=\begin{pmatrix}\frac{\partial\sum_{j=1}^{p}a_jx_j}{\partial x_1}\\ \vdots\\ \frac{\partial\sum_{j=1}^{p}a_jx_j}{\partial x_p}\end{pmatrix}=\begin{pmatrix}a_1\\ \vdots\\ a_p\end{pmatrix}=a$$

$$\frac{\partial x^{\mathrm{T}}Ax}{\partial x}=\begin{pmatrix}\frac{\partial x^{\mathrm{T}}Ax}{\partial x_1}\\ \vdots\\ \frac{\partial x^{\mathrm{T}}Ax}{\partial x_p}\end{pmatrix}=\begin{pmatrix}\frac{\partial\sum_{i=1}^{p}\sum_{j=1}^{p}a_{ij}x_ix_j}{\partial x_1}\\ \vdots\\ \frac{\partial\sum_{i=1}^{p}\sum_{j=1}^{p}a_{ij}x_ix_j}{\partial x_p}\end{pmatrix}=\begin{pmatrix}2a_{11}x_1+\cdots+2a_{1p}x_p\\ \vdots\\ 2a_{p1}x_1+\cdots+2a_{pp}x_p\end{pmatrix}=2Ax.$$

因此对$b$求一阶导, 对于多元场合, 一阶条件就是:

$$-\frac{2}{n}\sum_{i=1}^{n}x_i(y_i-x_i^{\mathrm{T}}\hat{\beta})=0.$$

于是

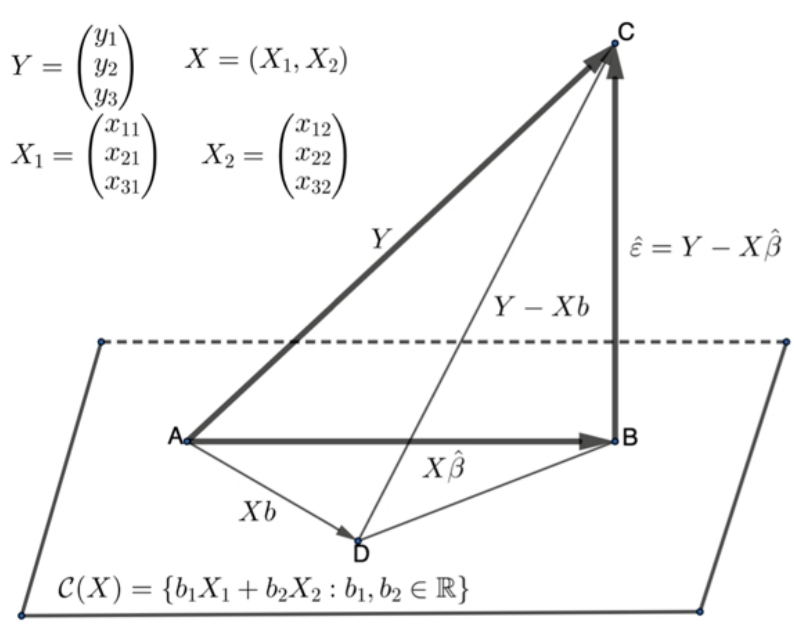
$$\sum_{i=1}^{n}x_i(y_i-x_i^{\mathrm{T}}\hat{\beta})=0\Longleftrightarrow X^{\mathrm{T}}(Y-X\hat{\beta})=0.$$

即

$$\hat{\beta}=\left(\sum_{i=1}^{n}x_ix_i^{\mathrm{T}}\right)^{-1}\left(\sum_{i=1}^{n}x_iy_i\right)=(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Y$$

这里我们有必要要求$X^{\top}X$是可逆的.

## 3.2 最小二乘的几何意义



从图中我们可以看到: 残差和协变量是正交的, 这是因为:

$$
\begin{aligned}
& X_1^{\mathrm{T}}\hat{\varepsilon}=0,\ldots,X_p^{\mathrm{T}}\hat{\varepsilon}=0,\\
\Longleftrightarrow\ & X^{\mathrm{T}}\hat{\varepsilon}=\begin{pmatrix}X_1^{\mathrm{T}}\hat{\varepsilon}\\ \vdots\\ X_p^{\mathrm{T}}\hat{\varepsilon}\end{pmatrix}=0,\\
\Longleftrightarrow\ & X^{\mathrm{T}}(Y-X\hat{\beta})=0,
\end{aligned}
$$

大多数情况下, $X$包含1向量, 因此

$$1_n^{\top}\hat{\varepsilon}=0\Rightarrow n^{-1}\sum_{i=1}^{n}\hat{\varepsilon}_i=0$$

## 3.3 矩阵的列空间

$$A=(a_{ij})=\begin{pmatrix}a_{11} & \cdots & a_{1m}\\ \vdots & & \vdots\\ a_{n1} & \cdots & a_{nm}\end{pmatrix}$$

$$A=(A_1,\ldots,A_m)$$

$$\mathcal{C}(A)=\{\alpha_1A_1+\cdots+\alpha_mA_m:\alpha_1,\ldots,\alpha_m\in\mathbb{R}\}$$

$$
\begin{aligned}
v\in\mathcal{C}(X) &\quad\Longleftrightarrow\quad v=Xb\\
w\perp\mathcal{C}(X) &\quad\Longleftrightarrow\quad X^Tw=0
\end{aligned}
$$

## 3.4 最小二乘法的另一种证明: 加一项减一项

后续在统计计算中还会推广到不可逆的情况.

$$
\begin{aligned}
\|Y-Xb\|^2 &= (Y-Xb)^{\mathrm{T}}(Y-Xb)\\
&= (Y-X\hat{\beta}+X\hat{\beta}-Xb)^{\mathrm{T}}(Y-X\hat{\beta}+X\hat{\beta}-Xb)\\
&= (Y-X\hat{\beta})^{\mathrm{T}}(Y-X\hat{\beta})+(X\hat{\beta}-Xb)^{\mathrm{T}}(X\hat{\beta}-Xb)\\
&\quad +(Y-X\hat{\beta})^{\mathrm{T}}(X\hat{\beta}-Xb)+(X\hat{\beta}-Xb)^{\mathrm{T}}(Y-X\hat{\beta})
\end{aligned}
$$

，

$$(X\hat{\beta}-Xb)^{\mathrm{T}}(Y-X\hat{\beta})=(\hat{\beta}-b)^{\mathrm{T}}X^{\mathrm{T}}(Y-X\hat{\beta})=0$$

$$\|Y-Xb\|^2=\|Y-X\hat{\beta}\|^2+\|X(\hat{\beta}-b)\|^2$$

$$\|Y-Xb\|^2\geq\|Y-X\hat{\beta}\|^2$$

Equality holds only if $b=\hat{\beta}$.

**定义 (加号逆)** 设 $A$ 为 $n\times m$ 矩阵，若 $m\times n$ 矩阵 $G$ 满足

$$
\begin{gathered}
\text{i)}AGA=A;\\
\text{ii)}GAG=G;\\
\text{iii)}(AG)^T=AG;\\
\text{iv)}(GA)^T=GA
\end{gathered}
$$

则称矩阵 $G$ 为矩阵 $A$ 的**加号逆**或 Moore-Penrose 广义逆，记作 $A^+$。如果 $G$ 满足定义中的第一个条件，则称 $G$ 为 $A$ 的**减号逆**，记作 $A^-$。显然，如果 $A$ 本身就是 $n$ 阶可逆方阵，则 $A^{-1}$ 满足上述四个条件。

广义逆可以用来分析回归分析和线性模型问题中最小二乘解的结构。设 $X$ 为 $n\times m$ 矩阵 $(n>m)$，则当 $X$ 列满秩时矩阵 $P=X(X^TX)^{-1}X^T$ 是对称幂等矩阵，可以把向量 $\boldsymbol{y}$ 正交投影到 $X$ 的各列张成的线性空间 $\mu(X)$ 中, 这时最小二乘问题

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^m}\|\boldsymbol{y}-X\boldsymbol{\beta}\|_2^2 \tag{5.99}$$

有唯一解 $\hat{\boldsymbol{\beta}}=(X^TX)^{-1}X^T\boldsymbol{y}$。对一般情况有如下结论。

**定理** 5.6.4. 设 $X$ 为 $n\times m$ 矩阵 $(n>m)$, 则最小二乘问题(5.99)的所有的最小二乘解可以写成

$$\hat{\boldsymbol{\beta}}=X^+\boldsymbol{y}+(I-X^+X)\boldsymbol{z},\ \forall\boldsymbol{z}\in\mathbb{R}^m. \tag{5.100}$$

在这些最小二乘解中 $\boldsymbol{\beta}_0=X^+\boldsymbol{y}$ 是唯一的长度最短的解。

## 3.5 帽子矩阵

$HY = \hat{Y} = \arg\min_{v \in C(X)} \|Y - v\|^2$, 其中$H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$.

它是对称幂等的:

$$H^2 = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} = H,$$
$$H^{\mathrm{T}} = \left\{X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\right\}^{\mathrm{T}} = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} = H.$$

根据高等代数知识我们知道, 它的特征值只能是0或1. 再由于

$$\begin{aligned} \mathrm{trace}(H) = &\quad \mathrm{trace}\left\{X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\right\} \\ = &\quad \mathrm{trace}\left\{(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X\right\} \\ = &\quad \mathrm{trace}(I_p) = p. \end{aligned}$$

因此, $H$一定可以正交对角化为$H = P\mathrm{diag}\{1, \ldots, 1, 0, \ldots, 0\}P^{\mathrm{T}}$. 且

$$Hv = v \Longleftrightarrow v \in C(X);$$

$$Hw = 0 \Longleftrightarrow w \perp C(X).$$

帽子矩阵告诉我们:

1. $\hat{y}_i$是所有观测$y_i$的线性组合:

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij}y_j = h_{ii}y_i + \sum_{j \neq i} h_{ij}y_j.$$

2. 当$X$有截距项时,

$$H1_n = 1_n \Longrightarrow \sum_{j=1}^{n} h_{ij} = 1 \quad (i = 1, \ldots, n).$$

**例 1.** *在处理-对照试验中, 有$m$个处理个体和$n$个对照个体, 则设计矩阵, 帽子矩阵分别为*

$$X = \begin{pmatrix} 1_m & 1_m \\ 1_n & 0_n \end{pmatrix} \quad H = \mathrm{diag}\{m^{-1}1_m1_m^{\mathrm{r}}, n^{-1}1_n1_n^{\mathrm{r}}\}.$$

## 3.6 回顾最小二乘的几何意义和正交分解

$$Y = \hat{Y} + \hat{\varepsilon},$$

$$\hat{Y} = X\hat{\beta} = HY$$

$$Y - \hat{Y} = (I_n - H)Y.$$

在PPT中指出, $H, I - H$都是投影矩阵, 且$H(I_n - H) = (I_n - H)H = 0$. 因此,

$$\hat{Y}^{\mathrm{T}}\hat{\varepsilon} = Y^{\mathrm{T}}H^{\mathrm{T}}(I_n - H)Y = Y^{\mathrm{T}}H(I_n - H)Y = 0.$$