T7.3

已知数据点 $(1,3),(3,1),(5,7),(4,6),(7,3)$，分别求**总体最小二乘**和**一般最小二乘**的拟合直线，并分析它们的距离平方和。

普通最小二乘（OLS）

我们考虑**让拟合直线通过数据点中心**，则直线方程写为：

$$m(x-\bar{x})+(y-\bar{y})=0$$

或者

$$(x-\bar{x})+m(y-\bar{y})=0$$

这里

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

代价函数

将n个数据点带入方程，直线方程不会严格满足等号。因此设置代价函数为**拟合误差的平方和**：

$$D_{\mathrm{LS}}^{(1)}(m,\bar{x},\bar{y})=\sum_{i=1}^{n}[(x_i-\bar{x})+m(y_i-\bar{y})]^2$$
$$D_{\mathrm{LS}}^{(2)}(m,\bar{x},\bar{y})=\sum_{i=1}^{n}[m(x_i-\bar{x})+(y_i-\bar{y})]^2$$

求解

$$\bar{x}=4,\bar{y}=4.2$$

对于这两种代价函数，直接对m求偏导。

$$D_{LS}^{(1)}=(-3-1.2m)^2+(-1-3.2m)^2+(1+2.8m)^2+(1.8m)^2+(3-0.2m)^2$$

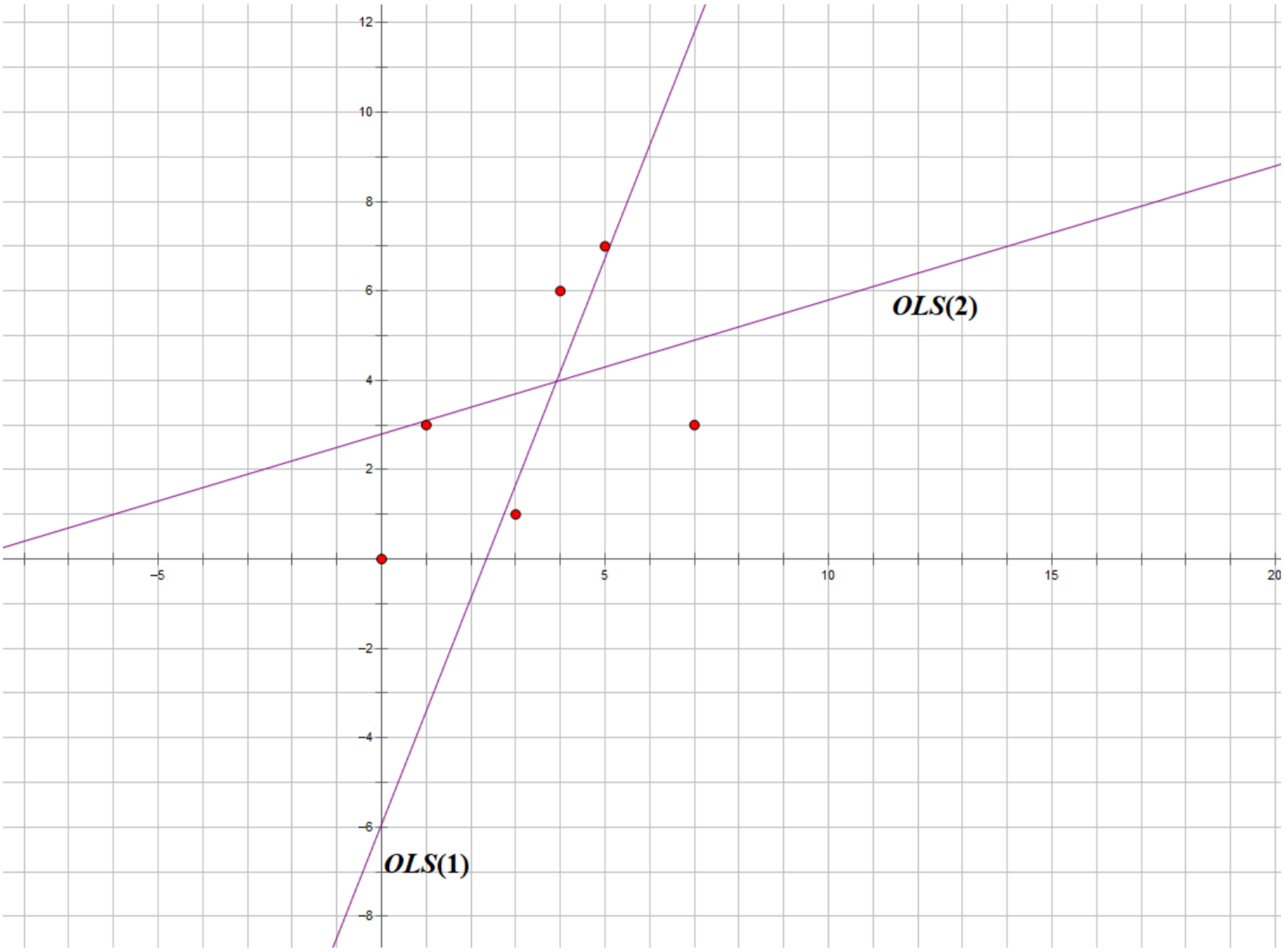
$$\frac{\partial D_{\mathrm{LS}}^{(1)}}{\partial m}=45.6m+18=0\Rightarrow m=-0.3947$$

$$(x-4)-0.3947(y-4.2)=0$$

$$D_{\mathrm{LS}}^{(2)}=(-3m-1.2)^2+(-m-3.2)^2+(m+2.8)^2+1.8^2+(3m-0.2)^2$$

$$\frac{\partial D_{\mathrm{LS}}^{(1)}}{\partial m}=40m+18=0\Rightarrow m=-0.45$$

$$-0.45(x-4)+(y-4.2)=0$$



总体最小二乘（TLS）

重新设计代价函数。总体最小二乘法的代价函数考虑：已知数据点到直线方程 $a(x-\overline{x})+b(y-\overline{y})$的**距离平方和最小化。**

$$\bar{\boldsymbol{x}}=\begin{bmatrix}\bar{x}\\ \bar{y}\end{bmatrix}$$

$$=\frac{1}{5}\sum_{i=1}^{5}\boldsymbol{x}_i=\frac{1}{5}\left\{\begin{bmatrix}1\\3\end{bmatrix}+\begin{bmatrix}3\\1\end{bmatrix}+\begin{bmatrix}5\\7\end{bmatrix}+\begin{bmatrix}4\\6\end{bmatrix}+\begin{bmatrix}7\\4\end{bmatrix}\right\}=\begin{bmatrix}4\\4.2\end{bmatrix}$$

- 为什么要通过均值点?

> **引理 6.3.1** [370] 对过直线的数据点 (x_0,y_0) 和数据点集合 $(x_1,y_1),\cdots,(x_n,y_n)$，恒有不等式
>
> $$D(a,b,\bar{x},\bar{y})\leqslant D(a,b,x_0,y_0)\tag{6.3.45}$$
>
> 等号成立，当且仅当 $x_0=\bar{x}$ 和 $y_0=\bar{y}$。
>
> 引理 6.3.1 表明，总体最小二乘拟合的直线必须通过 n 个数据点的中心 $(\bar{x},\bar{y})$，才能使偏差 D 最小。

代价函数$D(a,b)$

点(p,q)到直线$ax+by-c=0$的距离：

$$d^2=\frac{(ap+bq-c)^2}{a^2+b^2}$$

如果直线经过点(x_0,y_0)，则$c=ax_0+by_0$，距离化简为：

$$d^2=\frac{[a(p-x_0)+b(q-y_0)]^2}{a^2+b^2}$$

已知的n个数据点到直线$a(x-\bar{x})+b(y-\bar{y})=0$的距离平方和为：

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

写成矩阵形式：

$$D(a,b)=\|\boldsymbol{Mt}\|_2^2=\left\|\begin{bmatrix}x_1-\bar{x} & y_1-\bar{y}\\ x_2-\bar{x} & y_2-\bar{y}\\ \vdots & \vdots\\ x_n-\bar{x} & y_n-\bar{y}\end{bmatrix}\frac{1}{\sqrt{a^2+b^2}}\begin{bmatrix}a\\ b\end{bmatrix}\right\|_2^2$$

上式中

$$M=\begin{bmatrix}x_1-\bar{x} & y_1-\bar{y}\\ x_2-\bar{x} & y_2-\bar{y}\\ \vdots & \vdots\\ x_n-\bar{x} & y_n-\bar{y}\end{bmatrix}\quad \boldsymbol{t}=\frac{1}{\sqrt{a^2+b^2}}\begin{bmatrix}a\\ b\end{bmatrix}$$

求解最小值

把代价函数写成Rayleigh商的形式：

$$\|\boldsymbol{Mt}\|_2^2=\frac{t^{\mathrm{T}}M^{\mathrm{T}}Mt}{t^{\mathrm{T}}t}$$

答案就是Rayleigh商取最小值的解：M^TM**最小特征值对应的特征向量。**

对M^TM对特征值分解：

```
M =

   -3.0000   -1.2000
   -1.0000   -3.2000
    1.0000    2.8000
         0    1.8000
    3.0000   -0.2000
```

```
>> A=transpose(M)*M

A =

   20.0000    9.0000
    9.0000   22.8000
```

```
>> [V,D] = eig(A)

V =

   -0.7595    0.6505
    0.6505    0.7595

D =

   12.2918         0
         0   30.5082
```

所以

$$a = -0.7595, \quad b = 0.6505$$

计算距离平方和:

$$D_{\mathrm{TLS}} = \left\| \begin{bmatrix} -3 & -1.2 \\ -1 & -3.2 \\ 1 & 2.8 \\ 0 & 1.8 \\ 3 & -0.2 \end{bmatrix} \begin{bmatrix} -0.7595 \\ 0.6505 \end{bmatrix} \right\|_2^2 = 12.2918$$