DAY 1-30《数学建模姜启源版》

CH10.1——统计回归模型：牙膏的销售量

【建立回归模型的主要目的之一就是对因变量进行预测】

1、建模过程总结：

- （1）根据已知数据，从常识及经验进行分析与假设，辅以作图（散点图）
- （2）决定回归变量（多个）及其函数形式（线性的或二次的等等）
- （3）软件求解（如MATLAB统计工具箱regress/restool等等）
- （4）统计分析R^2,F,p,s^2的大小对模型整体的评价
- （5）判断每个回归系数置信区间是否包含零点——检验对应回归变量对因变量的影响是否显著（包含表示不显著）
- （6）模型改进：如添加交互项、完全二次多项式等等

2、牙膏的销售量实例

问题：市场调查收集到了过去30个销售周期（每个销售周期为4周）本公司生产的牙膏的销售量、销售价格、投入的广告费用、同期其他厂家生产的同类牙膏的市场平均销售价格。

要求：分析牙膏销售量与其他因素（或者是因素间作用）的关系，为制订价格策略和广告投入策略提供数据依据。

分析与假设：作为生活必需品，顾客更多的是关注不同品牌之间的价格差异。故将本公司销售价格和其他厂家平均价格作**差**取为一个影响因素。

- 牙膏销售量y
- 其他厂家价格x_3
- 本公司价格x_4
- 价格之差$x_1=x_3-x_4$
- 广告投入费用x_2

  基本模型：由y对x_1或x_2的散点图得出

【线性模型】

$$y=\beta_0+\beta_1x_1+\varepsilon$$

【二次函数模型】

$$y=\beta_0+\beta_1x_2+\beta_2x_2^2+\varepsilon$$

【回归模型】

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_2^2+\varepsilon$$

- 回归变量：x_1、x_2
- 回归系数：β_0、β_1、β_2、β_3

  模型求解：MATLAB统计工具箱regress

```
[b,bint,r,rint,stats]=regress(y,x,alpha)
```

其中输入 y 为模型(3)中 y 的数据(n 维向量,$n=30$),x 为对应于回归系数 $\boldsymbol{\beta}=(\beta_0,\beta_1,\beta_2,\beta_3)$ 的数据矩阵$[1\quad x_1\quad x_2\quad x_2^2]$($n\times4$ 矩阵,其中第 1 列为全 1 向量),*alpha* 为置信水平 α(缺省时 $\alpha=0.05$);输出 b 为 $\boldsymbol{\beta}$ 的估计值,常记作 $\hat{\boldsymbol{\beta}}$, bint 为 b 的置信区间,r 为残差向量 $\boldsymbol{y}-\boldsymbol{x}\hat{\boldsymbol{\beta}}$, rint 为 r 的置信区间,stats 为回归模型的检验统计量,有 4 个值,第 1 个是回归方程的决定系数 R^2(R 是相关系数),第 2 个是 F 统计量值,第 3 个是与 F 统计量对应的概率值 p,第 4 个是剩余方差 s^2.

结果分析：

表 2 模型(3)的计算结果

参数	参数估计值	参数置信区间
β_0	17.324 4	[5.728 2, 28.920 6]
β_1	1.307 0	[0.682 9, 1.931 1]
β_2	−3.695 6	[−7.498 9, 0.107 7]
β_3	0.348 6	[0.037 9, 0.659 4]
$R^2=0.905\,4$ $F=82.940\,9$ $p<0.000\,1$ $s^2=0.049\,0$		

结果分析 表 2 显示,$R^2=0.905\,4$ 指因变量 y(销售量)的 90.54% 可由模型确定,F 值远远超过 F 检验的临界值,p 远小于 α,因而模型(3)从整体来看是可用的.

表 2 的回归系数给出了模型(3)中 $\beta_0,\beta_1,\beta_2,\beta_3$ 的估计值,即 $\hat{\beta}_0=17.324\,4,\hat{\beta}_1=1.307\,0,\hat{\beta}_2=-3.695\,6,\hat{\beta}_3=0.348\,6$. 检查它们的置信区间发现,只有 β_2 的置信区间包含零点(但区间右端点距零点很近),表明回归变量 x_2(对因变量 y 的影响)不是太显著的,但由于 x_2^2 是显著的,我们仍将变量 x_2 保留在模型中.

销售量预测：把回归系数的估计值代入模型（3）得到预测方程：

$$\hat{y}=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2x_2+\hat{\beta}_3x_2^2$$

补充说明：

回归模型的一个重要应用是,对于给定的回归变量的取值,可以以一定的置信度预测因变量的取值范围,即预测区间. 比如当 $x_1=0.2,x_2=6.5$ 时可以算出①,牙膏销售量的置信度为95%的预测区间为[7.823 0,8.763 6],它表明在将来的某个销售周期中,如公司维持产品的价格差为 0.2 元,并投入 650 万元的广告费用,那么可以有 95% 的把握保证牙膏的销售量在 7.823 到 8.763 6 百万支之间. 实际操作时,预测上限可以用来作为库存管理的目标值,即公司可以生产(或库存)8.763 6 百万支牙膏来满足该销售周期顾客的需求;预测下限则可以用来较好地把握(或控制)公司的现金流,理由是公司对该周期销售 7.823 百万支牙膏十分自信,如果在该销售周期中公司将牙膏售价定为3.70元,且估计同期其他厂家的平均价格为 3.90 元,那么董事会可以有充分的依据知道公司的牙膏销售额应在 7.823 × 3.7 ≈ 29 百万元以上.

模型改进：

【1】用x_1和x_2的乘机表示交互作用加入模型（3）

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_2^2+\beta_4x_1x_2+\varepsilon \qquad (5)$$

在这个模型中,y 的均值与 x_2 的二次关系为$(\beta_2+\beta_4x_1)x_2+\beta_3x_2^2$,由系数 β_2,β_3,β_4 确定,并依赖于价格差 x_1.

- 比较结果：

表 3 模型(5)的计算结果

参数	参数估计值	参数置信区间
β_0	29.113 3	[13.701 3, 44.525 2]
β_1	11.134 2	[1.977 8, 20.290 6]
β_2	−7.608 0	[−12.693 2, −2.522 8]
β_3	0.671 2	[0.253 8, 1.088 7]
β_4	−1.477 7	[−2.851 8, −0.103 7]
$R^2=0.920\,9$ $F=72.777\,1$ $p<0.000\,1$ $s^2=0.042\,6$		

---

① 具体计算参见[91](4.77)~(4.79)式,用 MATLAB 统计工具箱中现成的程序结果与此不同.

表 3 与表 2 的结果相比 R^2 有所提高,说明模型(5)比模型(3)有所改进. 并且,所有参数的置信区间,特别是 x_1,x_2 的交互作用项 x_1x_2 的系数 β_4 的置信区间不包含零点,所以有理由相信模型(5)比模型(3)更符合实际.

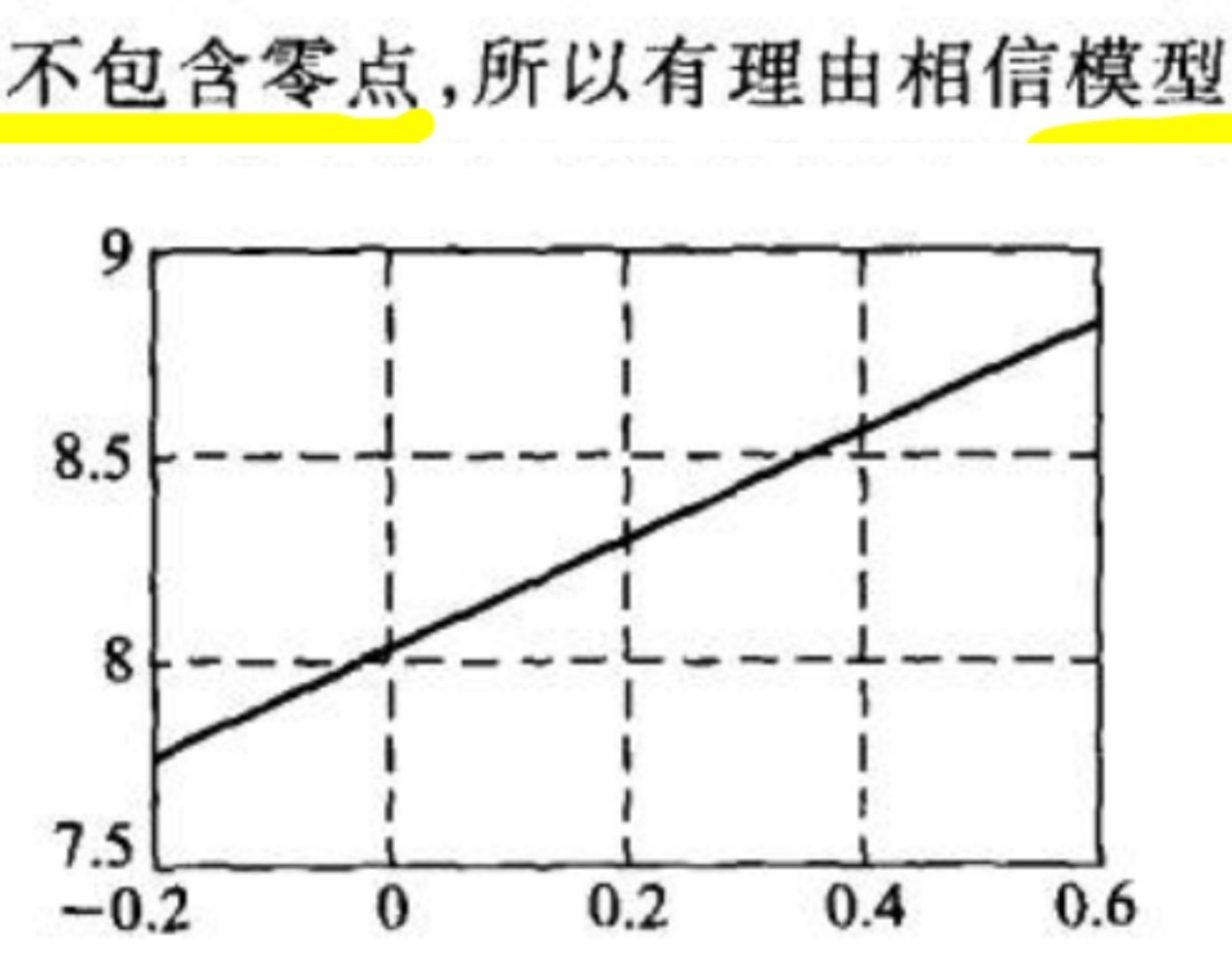

图 3 模型(3)$\hat{y}$ 与 x_1 的关系

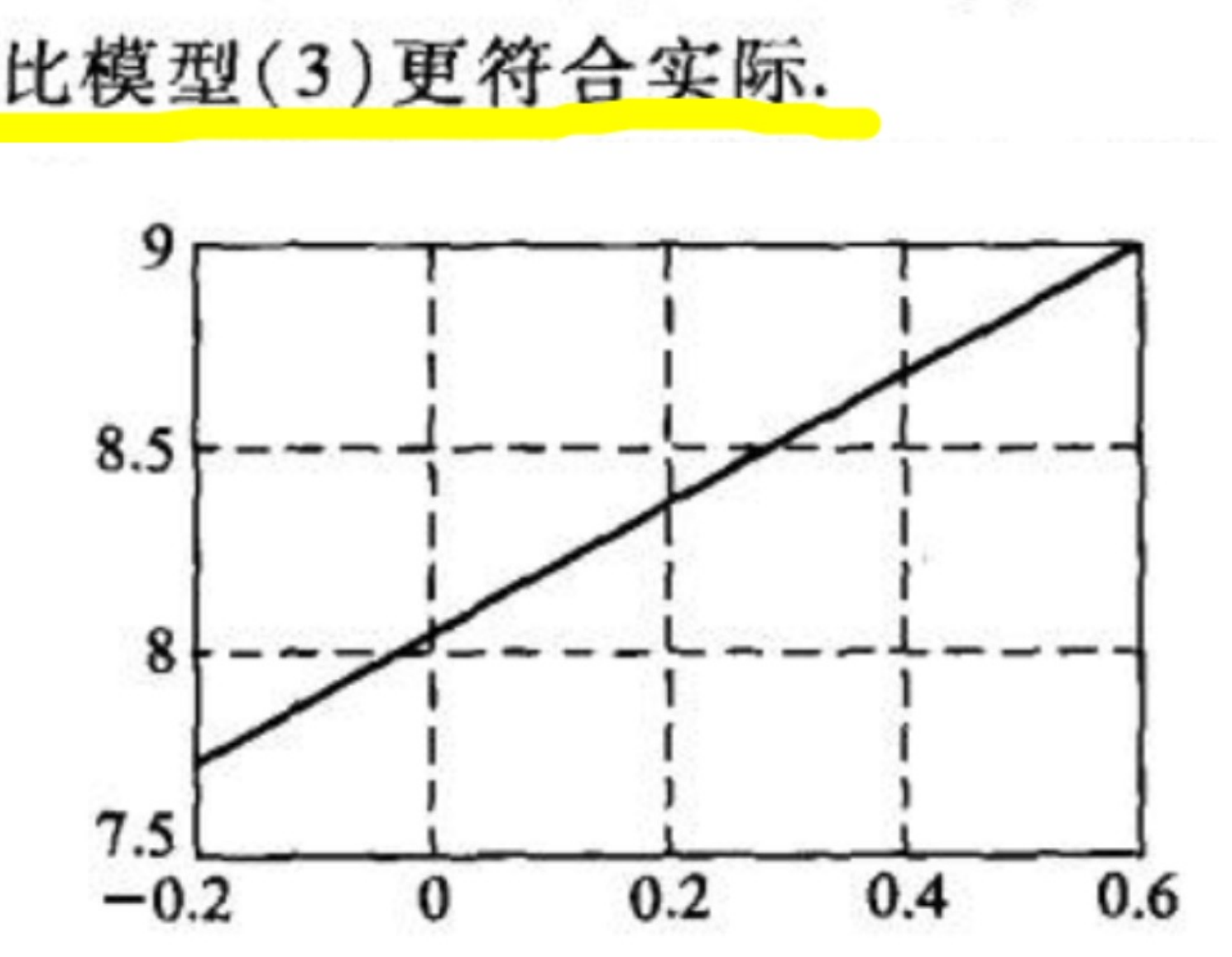

图 4 模型(5)$\hat{y}$ 与 x_1 的关系

在保持价格差 $x_1=0.2$ 元不变的条件下,分别对模型(3)和(5)中牙膏销售量的均值 $\hat{y}$ 与广告费用 x_2 的关系作图,见图 5 和图 6.

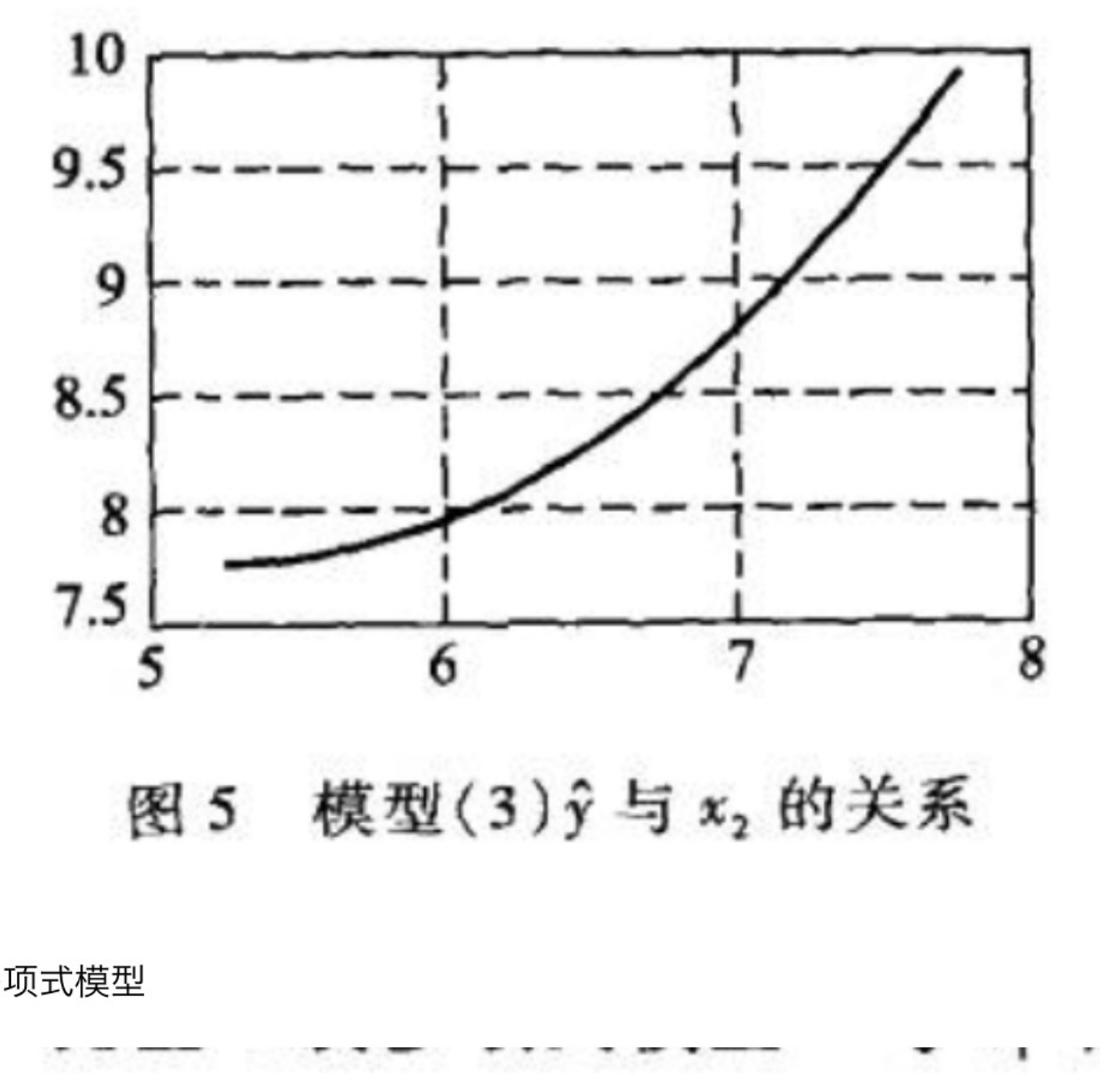

图 5 模型(3)$\hat{y}$ 与 x_2 的关系

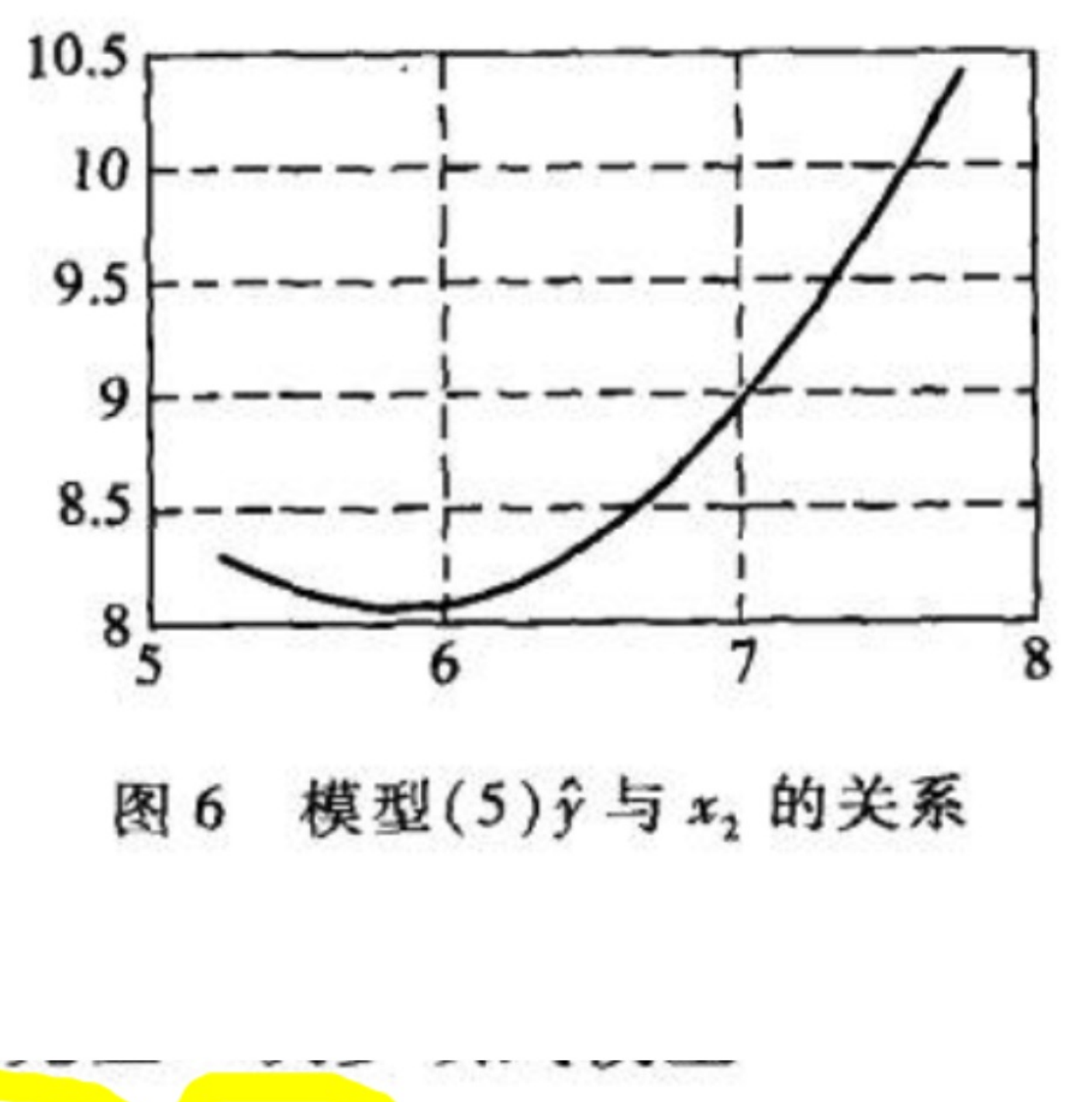

图 6 模型(5)$\hat{y}$ 与 x_2 的关系

- 结果总结：

【2】完全二次多项式模型

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_1x_2+\beta_4x_1^2+\beta_5x_2^2+\varepsilon \qquad (10)$$

相比,模型(5)只少 x_1^2 项,我们不妨增加这一项,建立模型(10). 这样做的好处之一是 MATLAB 统计工具箱中有直接的命令 rstool 求解,并且以交互式画面给出 y 的估计值 $\hat{y}$ 和预测区间. 这个命令的输出如图 8,从左下方的输出 Export 可

- 软件求解得到回归系数的估计值：

$$\hat{\beta}=(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2,\hat{\beta}_3,\hat{\beta}_4,\hat{\beta}_5)=(32.098\,4,14.743\,6,$$

$$-8.636\,7,-2.103\,8,1.107\,4,0.759\,4)$$

- 比较结果：和【1】相差不大

用鼠标移动交互式画面中的十字线,或在图下方的窗口内输入,可改变 x_1 和 x_2 的数值,图中当 $x_1=0.2,x_2=6.5$ 时,左边的窗口显示 $\hat{y}=8.302\,9$,预测区间为 $8.302\,9\pm0.255\,8=[8.047\,1,8.558\,7]$. 这些结果与模型(5)相差不大.

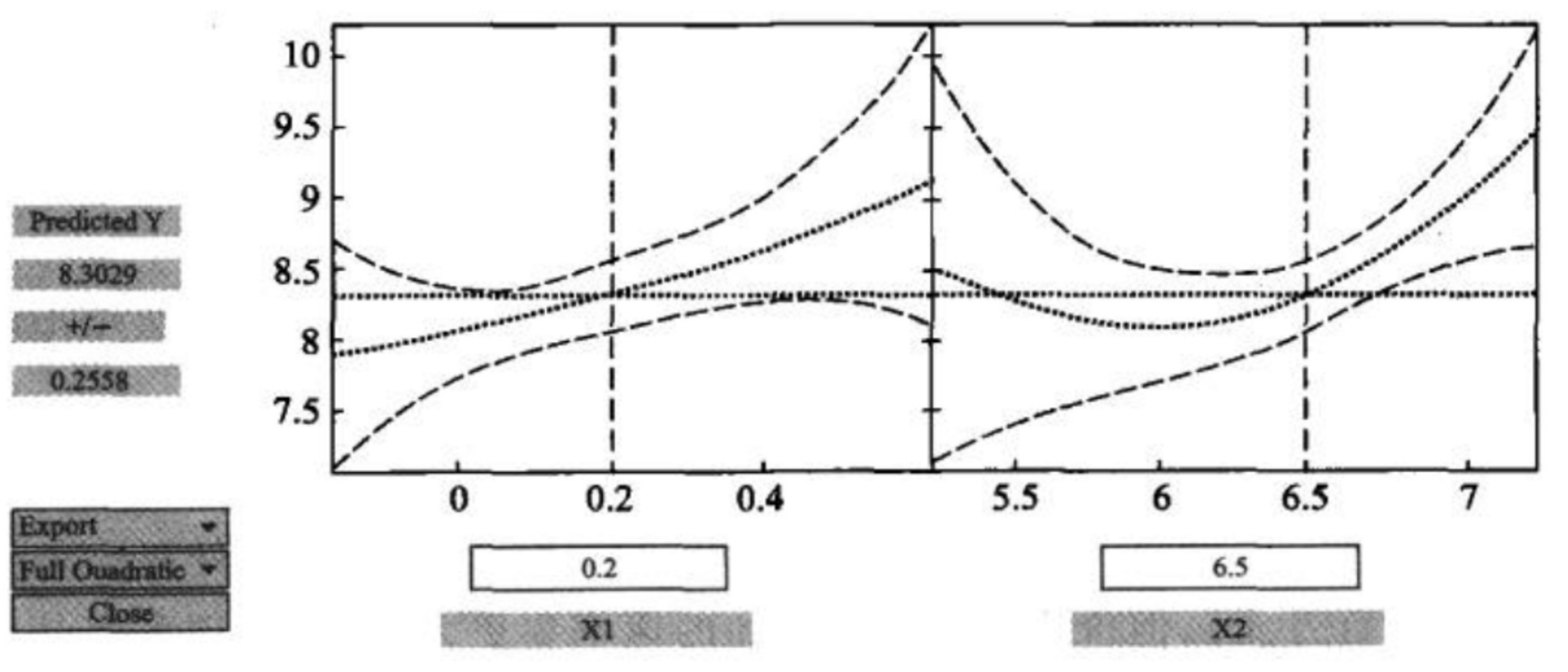

图 8 完全二次多项式模型(10)的输出