模式识别作业 1 线性回归

数学系 2014012118 李博扬

1、 证明题

$$r^2 = \frac{[\sum_1^n (x_i - \bar{x})\,(y_i - \bar{y})]^2}{\sum_1^n (x_i - \bar{x})^2 \sum_1^n (y_i - \bar{y})^2}$$

$$\frac{R^2}{r^2} = \frac{[\sum_1^n (x_i - \bar{x})^2]\,[\sum_1^n (\hat{y_i} - \bar{y})^2]}{[\sum_1^n (x_i - \bar{x})\,(y_i - \bar{y})]^2}$$

代入

$$\hat{y_i} = \widehat{\beta_1} x_i + \widehat{\beta_0} \qquad \bar{y} = \widehat{\beta_1}\bar{x} + \widehat{\beta_0} \qquad \widehat{\beta_1} = \frac{\sum_1^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_1^n (x_i - \bar{x})^2}$$

得到

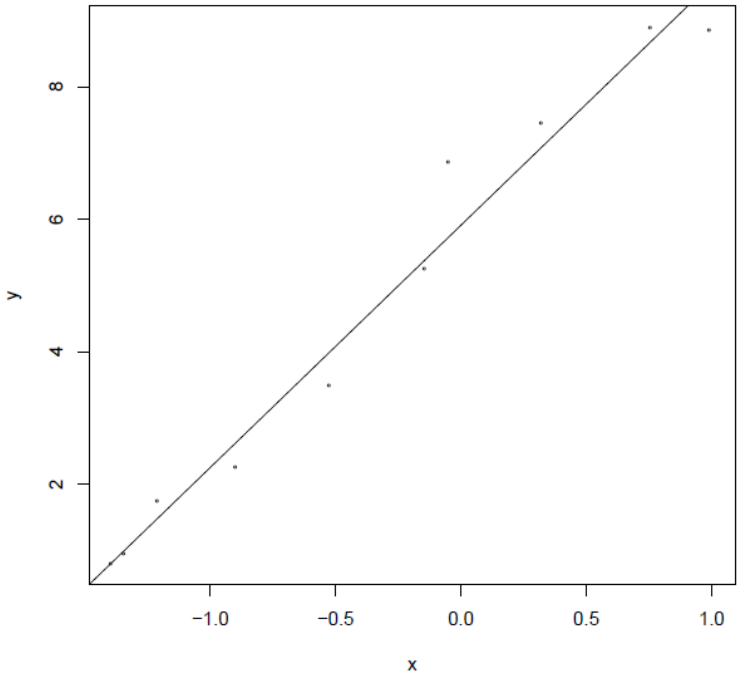
$$\frac{R^2}{r^2} = 1$$

证毕。

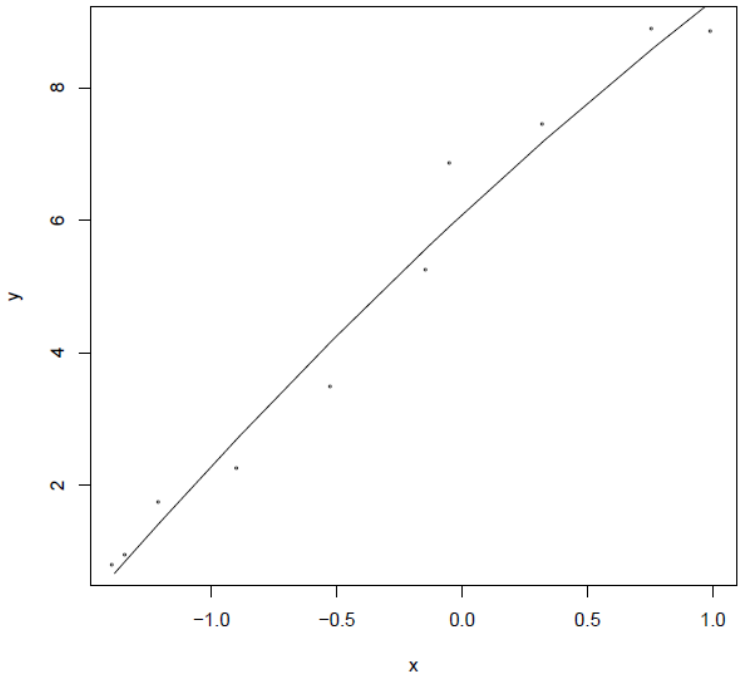
2、 过拟合问题

（1） 线性R^2=0.9741<二次R^2=0.9767<三次R^2= 0.9875

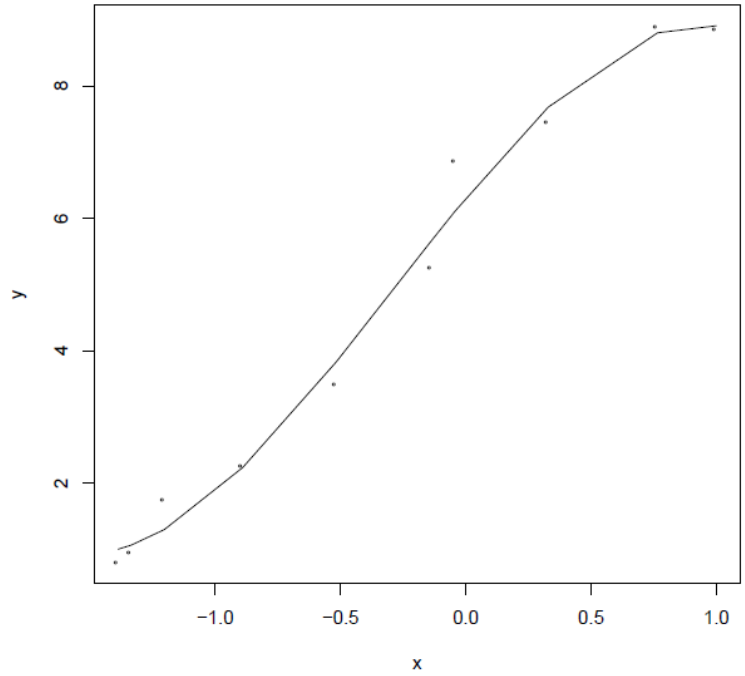
（2） 线性拟合：



一元二次拟合：



一元三次拟合：



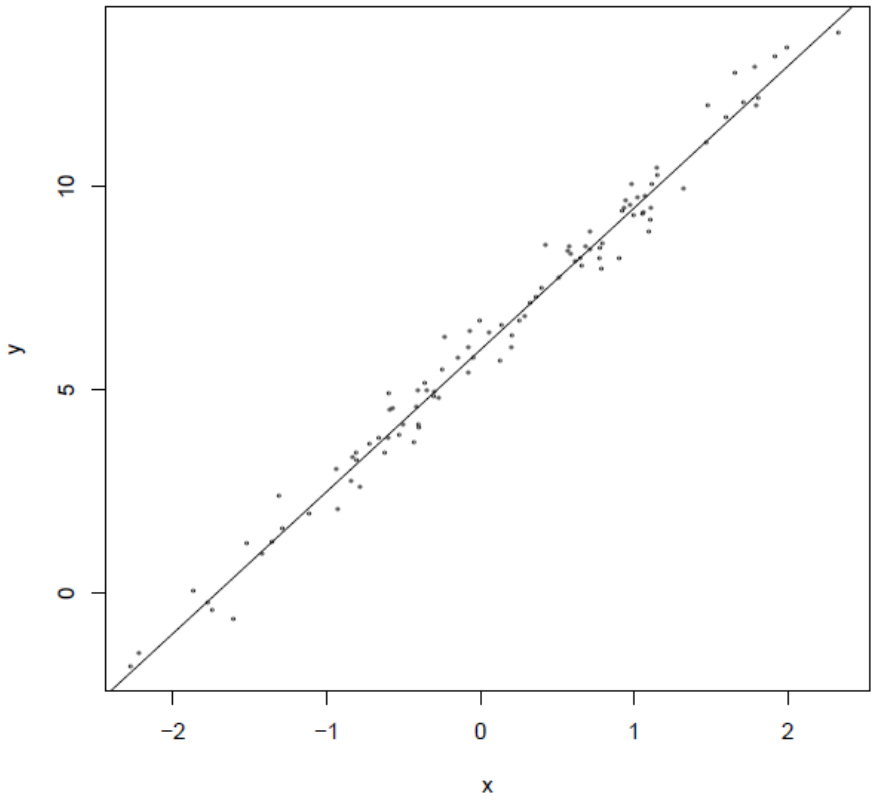
（3） 生成了 100 个样本得到拟合曲线上对应的 y 值，计算与真值差值的平方和得下表

	β_3	β_2	β_1	β_0	误差平方和
线性模型			3.6615	5.9095	27.78231
一元二次模型		-0.2815	3.5181	6.0837	40.96393

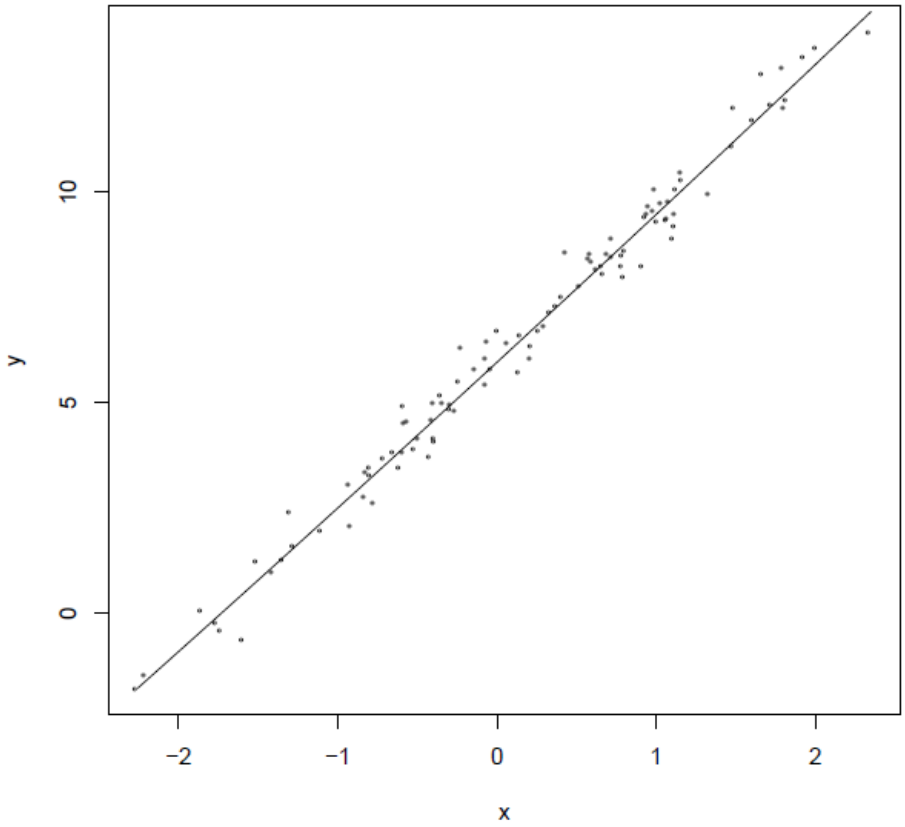
一元三次模型	-1.0693	-0.9256	4.6020	6.3065	581.0765

（4） 线性R^2= 0.9779<二次R^2= 0.9782<三次R^2= 0.9787

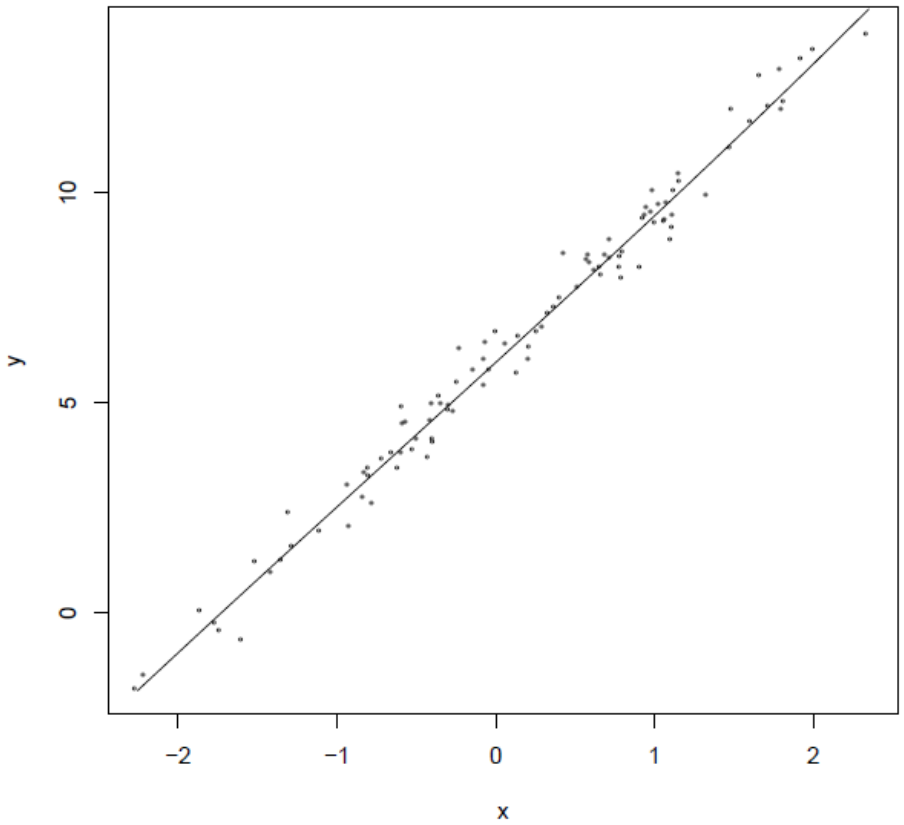
线性模型



一元二次模型：



一元三次模型



随机生成 100 个样本在拟合的曲线上找出，并计算与真值的差的平方和如下表：

	β_3	β_2	β_1	β_0	误差平方和

线性模型			3.4857	5.9857	27.92746
一元二次模型		0.05176	3.48351	5.94237	29.46914
一元三次模型	0.05017	0.05571	3.37299	5.94409	29.33832

（5） 以上为$\sigma = 0.5$的情况，若$\sigma = 2$，重复实验可以发现：R^2减小、误差平方和增加。若模型复杂度上升，则二者均上升。若训练集规模增大，则R^2增大、误差平方和减小，预测更准确。

3、癌症术后生存时间

（1）

拟合参数	β_3	β_2	β_1	β_0
参数值	61.4084	-161.5401	-0.7537	726.0731

预计存活 219.9542 天

（2）

参考项	R^2
无交叉项	0.2304
有交叉项	0.605

从R^2来看，有交叉项预测更精确。更详细数据见压缩包中“HW01_3_3 回归数据对比.txt”文件。