

# 作业1 线性回归

## 1.证明题[30]

请证明线性回归中 $R^2$ 与皮尔逊相关系数 $r$ 的关系： $R^2 = r^2$ ,

其中， $R^2 = \frac{\sum_1^n (\bar{y} - \hat{y}_i)^2}{\sum_1^n (\bar{y} - y_i)^2}$ ,  $r = \frac{cov(\mathbf{x}, \mathbf{y})}{\sigma_x \sigma_y}$ 。  $\mathbf{x}, \mathbf{y}$  均为一维向量。

## 2.过拟合问题[40]（作业中请提供源代码）

利用模型 $y = \theta_1 \cdot x + \theta_0 + \varepsilon$ 生成一组仿真数据 $(\mathbf{x}, \mathbf{y})$ ，其中 $\mathbf{x}$ 服从 $N(0, 1)$  正态分布， $\theta_1 = 3.5$ ， $\theta_0 = 6$ 。残差项 $\varepsilon$ 服从正态分布 $N(0, \sigma^2)$ ，其中分别考虑 $\sigma$  等于0.5和2 两种不同情况，完成下面要求：

(1)随机生成10个训练样本，分别用线性模型、一元二次模型、一元三次模型回归这组数据，得到回归模型的参数，计算 $R^2$ ，并比较大小。

(2)可视化：请绘制数据散点图和回归曲线。

(3)再随机生成100个测试样本，用(1)中的回归模型预测 $\mathbf{y}$ 值，并比较三种模型的预测效果。

(4)将题目(1)中“随机生成10个训练样本”改为“随机生成100个训练样本”，重复(1)-(3)过程。

(5)请再重复一次(1)、(3)、(4)步骤，总结一下 $\sigma$  取值、模型复杂程度、训练样本量和训练 $R^2$ 、测试 $R^2$ 的关系。

提示：

Matlab函数：`randn()`:用于生成正态分布；`regress()`、`fitlm()`、`fit()` 等:用于回归模型；(`regress`: 多元线性回归，`fitlm`:构造线性回归模型，`fit`:拟合曲线和曲面，不局限于线性)；`plot()`:绘制散点或线；学会在matlab帮助中查找函数的相关帮助。

## 3.癌症术后生存时间[30]（作业中请提供源代码）

有一组癌症患者术后生存时间的数据，其中有3个预测变量 $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3)$ ： $x_1 = gene$ ，表示某种基因表达量(FPKM)， $x_2 = size$ ，表示肿瘤的大小（cm）， $x_3 = Gender$ (1 表示女性，0 表示男性)，目标 $y$ 是术后生存时间（天）。附件X.txt文件给出预测变量，每一行表示一个样本，每一列表示一个变量，分别是基因表达、肿瘤大小、性别；Y.txt 文件给出对应样本的生存时间。不考虑交叉项情况下，请利用最小二乘法回归上述数据，给出 $\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3$  的取值。并回答下列问题：

(1) 给定一名女性， $size = 3cm$ ， $gene = 110$ ，请预测生存时间；

(2) 如果考虑预测变量的交叉项情况，是否会有更好的预测效果？请说明理由。并回答(1)题。

提示：

Matlab函数：`load()`:输入文件；`regress()`、`fitlm()` 等:用于回归模型；