2-6
拟合优度检验

主讲人：范国斌

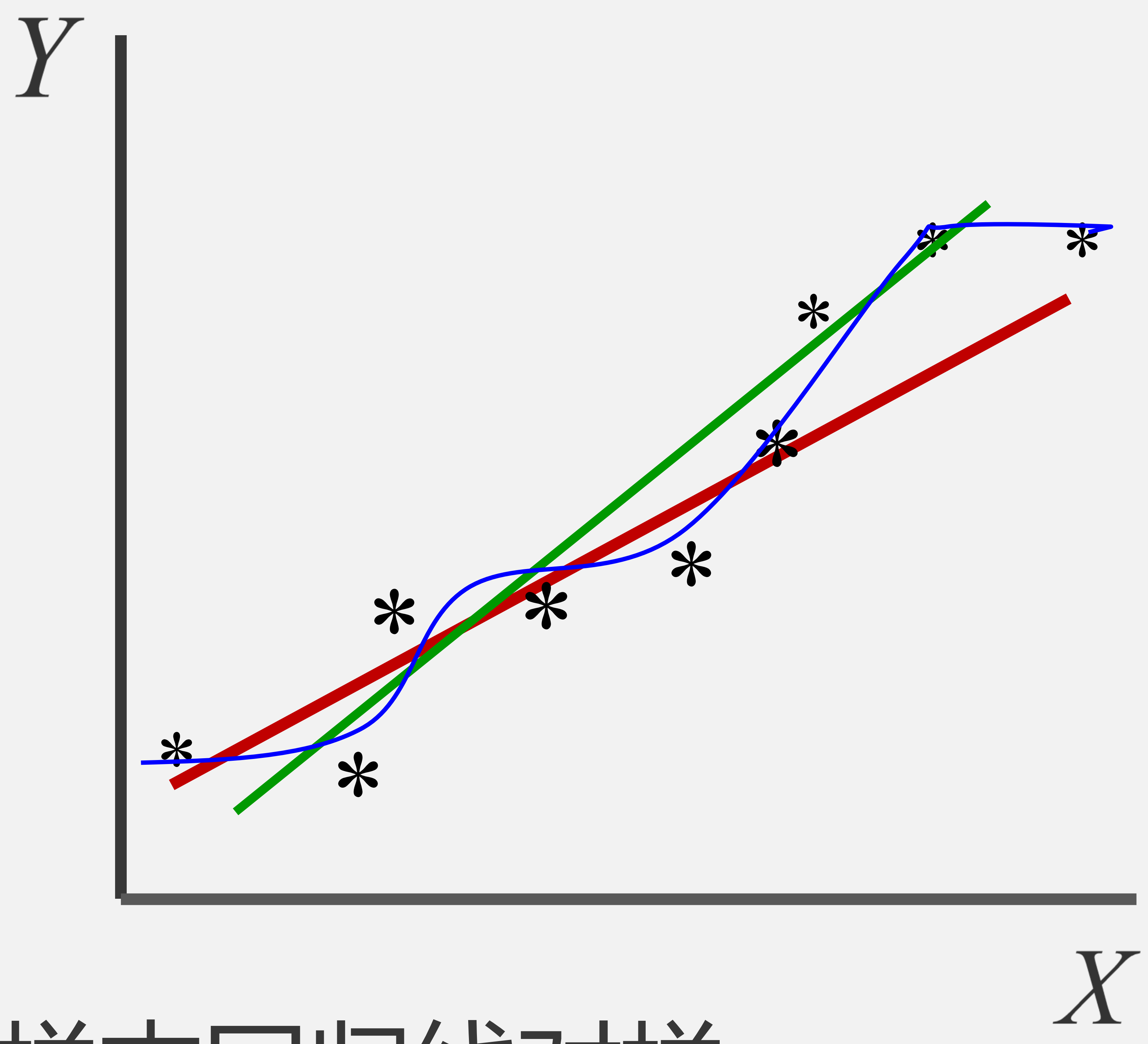
拟合优度的度量

概念：

样本回归线是对样本数据的一种拟合。

●不同的模型（不同函数形式)可拟合出不同的回归线。

●相同的模型用不同方法估计参数，可以拟合出不同的回归线。



拟合的回归线与样本观测值总是有偏离。样本回归线对样本观测数据拟合的优劣程度称为拟合优度。

如何度量拟合优度呢？

拟合优度的度量建立在对 Y 的总变差分解的基础上。

总变差的分解

分析Y的观测值 Y_i 、估计值 $\hat{Y}_i$ 与平均值 $\overline{Y}$ 有以下关系

$$Y_i - \overline{Y} = (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i)$$

将上式两边平方加总，可证得（提示：交叉项 $\sum(\hat{Y}_i - \overline{Y})e_i = 0$ ）

$$\sum(Y_i - \overline{Y})^2 = \sum(\hat{Y}_i - \overline{Y})^2 + \sum(Y_i - \hat{Y}_i)^2$$

（**TSS**） （**ESS**） （**RSS**）

或者表示为 $\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2$

总变差的分解

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2$$

总变差$\sum y_i^2$（**TSS**）：被解释变量**Y**的观测值与其平均值的离差平方和（总平方和）(说明 Y 的变动程度）

解释了的变差$\sum \hat{y}_i^2$(**ESS**）：被解释变量**Y**的估计值与其平均值的离差平方和（回归平方和）

剩余平方和$\sum e_i^2$（**RSS**）：被解释变量观测值与估计值之差的平方和（未解释的平方和）

可决系数

以TSS同除总变差等式两边：

$$\frac{\sum(Y_i-\bar{Y})^2}{\sum(Y_i-\bar{Y})^2}=\frac{\sum(\hat{Y}_i-\bar{Y})^2}{\sum(Y_i-\bar{Y})^2}+\frac{\sum(Y_i-\hat{Y}_i)^2}{\sum(Y_i-\bar{Y})^2}$$

或

$$1=\frac{\sum\hat{y}^2}{\sum y_i^2}+\frac{\sum e_i^2}{\sum y_i^2}$$

定义：回归平方和（解释了的变差ESS）$\sum\hat{y}_i^2$ 在总变差（TSS）$\sum y_i^2$ 中所占的比重称为可决系数，用 r^2 或 R^2 表示:

$$R^2=\frac{\sum\hat{y}^2}{\sum y_i^2}\quad\text{或}\quad R^2=1-\frac{\sum e_i^2}{\sum y_i^2}$$

可决系数的作用

可决系数越大，说明在总变差中由模型作出了解释的部分占的比重越大，模型拟合优度越好。反之可决系数越小，说明模型对样本观测值的拟合程度越差。

可决系数的特点：

- 可决系数取值范围：$0 \le R^2 \le 1$
- 随抽样波动，样本可决系数 R^2 是随抽样而变动的随机变量
- 可决系数是非负的统计量

可决系数使用原则

- 切勿因为R^2的高或低轻易地肯定或否定一个模型：
  - 视数据类型和样本容量
  - 视研究目的不同
  - 描述性判断而非显著性判断
- 可以比较不同模型的R^2但有前提：
  - 样本相同
  - 被解释变量相同
- R^2具有两层含义，R^2高意味着：
  - 样本回归线对样本数据的拟合程度较高
  - 所有解释变量联合起来对被解释变量的影响程度较高

拓展至多元线性回归模型

多元回归的拟合优度检验

多重可决系数：在多元回归模型中，由各个解释变量联合起来解释了的Y的变差，在Y的总变差中占的比重，用 R^2 表示与简单线性回归中可决系数 r^2 的区别只是 $\hat{Y}_i$ 不同

多元回归中 $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i} + \cdots + \hat{\beta}_k X_{ki}$

多重可决系数可表示为

$$R^2 = \frac{ESS}{TSS} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2} = \frac{TSS - RSS}{TSS} = 1 - \frac{\sum e_i^2}{\sum y_i^2}$$

(注意:红色字体是与一元回归不同的部分)

修正的可决系数

思想：

可决系数只涉及变差，没有考虑自由度。如果用自由度去校正所计算的变差，可纠正解释变量个数不同引起的对比困难。

回顾:

自由度：

统计量的自由度指可自由变化的样本观测值个数，它等于所用样本观测值的个数减去对观测值的约束个数。

可决系数的修正方法

总变差 TSS $=\sum(Y_i-\bar{Y})^2=\sum y_i^2$ 自由度为 **n-1**

解释了的变差 ESS $=\sum(\hat{Y}_i-\bar{Y})^2$ 自由度为 **k-1**

剩余平方和 RSS$=\sum(Y_i-\hat{Y}_i)^2=\sum e_i^2$ 自由度为 **n-k**

修正的可决系数为

$$\bar{R}^2=1-\frac{\sum e_i^2/(n-k)}{\sum y_i^2/(n-1)}=1-\frac{n-1}{n-k}\frac{\sum e_i^2}{\sum y_i^2}=1-\frac{n-1}{n-k}(1-R^2)$$

修正的可决系数 $\bar{R}^2$ 与可决系数 R^2 的关系

已经导出： $\bar{R}^2=1-(1-R^2)\dfrac{n-1}{n-k}$

注意：

可决系数 R^2 必定非负，但所计算的修正可决系数 $\bar{R}^2$ 有可能为负值

解决办法：若计算的 $\bar{R}^2<0$，规定 $\bar{R}^2$ 取值为0

$$\frac{1-\overline{R}^2}{1-R^2} = \frac{n-1}{n-k} \Rightarrow \overline{R}^2 \leq R^2$$

修正可决系数的特点

- 修正后$R^2 \leq R^2$，且随着解释变量个数增加两者差距变大。
- 修正后R^2与R^2同增同减（在其他条件不变的前提下），具有同样的两层含义。
- 修正后R^2不再是解释变量个数的不减函数，而要视正面影响（对拟合优度贡献）和负面影响（自由度损失）的相对大小。
- 修正后R^2也只能做描述性判断。
- 修正后R^2使用原则与R^2相同。