

回归：拟合优度与模型选择

司继春

2024年3月

总平方和的分解

- 问题： y 的方差中有多少是可以被 x 解释的。
- 回忆方差的分解：

$$\mathbb{V}(y) = \mathbb{V}[\mathbb{E}(y|x)] + \mathbb{E}[\mathbb{V}(y|x)]$$

- 记 y 的方差的分子为总平方和（total sum of squares）：

$$TSS = \sum_{i=1}^N (y_i - \bar{y})^2$$

- 分解（过程见讲义）：

$$TSS = ESS + RSS$$

其中：

$$ESS = \hat{Y}' M_0 \hat{Y} = \sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2 = \sum_{i=1}^N (\hat{y}_i - \bar{y})^2$$

$$RSS = \hat{e}' \hat{e} = \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

总平方和的分解

$$ESS = \sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2 = \sum_{i=1}^N (\hat{y}_i - \bar{y})^2$$

$$RSS = \hat{e}'\hat{e} = \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

- 回归平方和 (explained sum of squares) : x 可以解释的部分
- 残差平方和 (residual sum of squares) : x 不能解释的部分

R^2 的定义

定义:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

R^2 度量了使用 x 对 y 进行预测时, x 可以解释多少部分的 y 的方差。 R^2 一定是在 $[0, 1]$ 之间的。

- 当 $R^2 = 0$ 时, 对应于 $RSS = TSS$, 此时 $\hat{y}_i = \bar{y}$, 也就是说不管有没有 x , 对 y 的最优预测都是 \bar{y} , 意味着 x 对 y 完全没有任何解释能力。
- 而如果 $R^2 = 1$, 此时 $ESS = TSS$, 即 $\hat{y}_i = y_i$, 达到了完美拟合。

R^2 的取值范围

- TSS 完全可以看作是只包含常数项的 RSS
- 如果回归中包含常数项，那么预测结果不可能比只有常数项更差，从而必然会有 $RSS < TSS$ 。
- 但是，如果回归中不包含常数项，得到的预测可能效果会非常差（比只用常数项预测的结果还差），从而可能会出现 $TSS < RSS$ 的情况，从而得到 $R^2 < 0$ 。
- 所以， $R^2 \in [0, 1]$ 的前提条件是回归中包含常数项！

变量增加时的 R^2

- 如果我们在回归中添加新的解释变量，由于出现了更多的信息，因而 R^2 值不会降低。
- 们假设 x_1 是一个 $N \times 1$ 维列向量，而 X_2 为 $N \times K$ 维其他解释变量，线性回归：

$$Y = \beta_1 x_1 + X_2 \beta_2 + u$$

的最小二乘估计分别为 $\hat{\beta}_1, \hat{\beta}_2$ 。

- 令残差： $\hat{u} = Y - \hat{\beta}_1 x_1 - X_2 \hat{\beta}_2$ ，那么拟合优度

$$R^2 = 1 - \frac{\hat{u}'\hat{u}}{TSS}$$

变量增加时的 R^2

- 如果只对 X_2 做回归，即：

$$Y = X_2\gamma + e$$

记其最小二乘估计为 $\hat{\gamma}$ ，残差为 $\hat{e} = Y - X_2\hat{\gamma}$ ，那么其 R^2 为：

$$R_2^2 = 1 - \frac{\hat{e}'\hat{e}}{TSS}$$

变量增加时的 R^2

定理

在上述回归模型中， $R_2^2 \leq R^2$ 。

目标函数为：

$$\min_{\beta_1, \beta_2} (Y - \beta_1 x_1 + X_2 \beta_2)' (Y - \beta_1 x_1 + X_2 \beta_2)$$

而如果只对 X_2 做回归，目标函数为：

$$\min_{\gamma} (Y - X_2 \gamma)' (Y - X_2 \gamma)$$

相当于限制 $\beta_2 = \gamma, \beta_1 = 0$ ，从而必然有： $\hat{u}'\hat{u} \leq \hat{e}'\hat{e}$

变量增加时的 R^2

定理

R^2 和 R_2^2 满足：

$$R^2 = R_2^2 + [\text{Corr}(\hat{\epsilon}, \hat{\epsilon})]^2 (1 - R_2^2)$$

其中 $\hat{\epsilon}$ 为回归 $x_1 = X_2\delta + \epsilon$ 的残差，即 $\hat{\epsilon} = x_1 - X_2\hat{\delta}$ 。

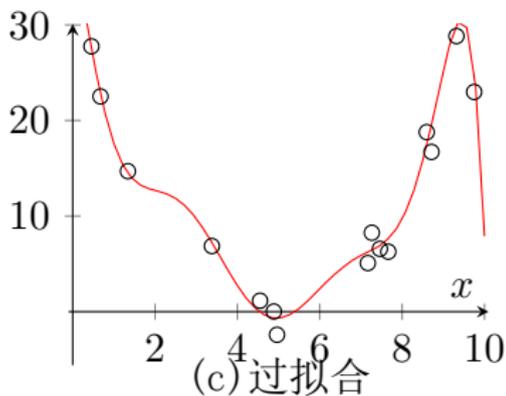
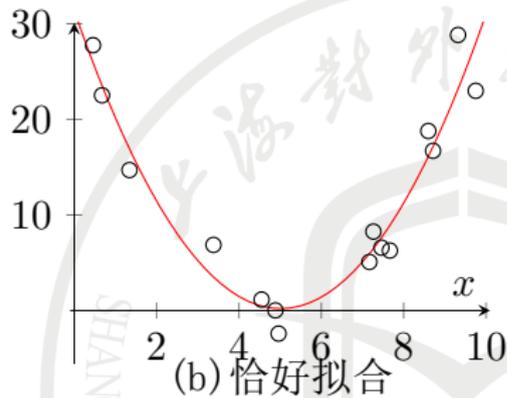
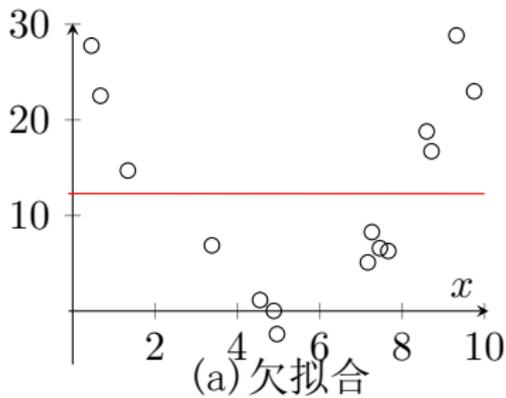
证明见讲义。其中 $\text{Corr}(\hat{\epsilon}, \hat{\epsilon})$ 度量了 Y 和 x_1 同时不能被 X_2 解释的部分相关系数，即排除 X_2 影响之后的相关系数，因而也被称为 Y 和 x_1 的偏相关系数（partial correlation）。

模型选择

从预测的角度，我们不仅仅需要对样本内进行预测并达到比较好的预测效果，更希望当有新的样本进来时，也达到非常好的预测效果。通常有三种拟合情况：

- 欠拟合 (under-fitting)：没有发现数据中本来有的 pattern
 - 通常起源于太『小』的模型
 - 比如本该是二次方关系，错误设定为线性关系
 - 即使在现有样本内都表现不好，更不要说样本外预测
- 过拟合 (over-fitting)：发现了本来没根本有的 pattern
 - 通常起源于太『大』的模型
 - 比如本来应该是二次方关系，但是使用了10阶多项式拟合
 - 样本内预测非常好，但是样本外预测非常差
- 恰好拟合：具有非常好的样本内以及样本外预测效果

欠拟合、过拟合与恰好拟合



过拟合问题

- 通常我们以 R^2 等指标为基础，更容易挑出 R^2 更大的模型来
- 但是 R^2 大并不代表样本外预测效果好，容易出现过拟合问题
- 过拟合会导致样本外预测效果很差
 - 我发现每次看国足都会都输球
 - 是否可以使用「我有没有看球」这个指标预测国足输赢？

调整后的 R^2

- 拟合优度, $R^2 = 1 - \frac{RSS}{TSS}$:
 - 度量了 y 中可预测部分 (\hat{y} 的方差) 占总的方差 (y 的方差) 的百分比
 - R^2 越大, 误差部分的方差越小, 拟合效果越好
 - R^2 大不代表得到了因果关系
- 调整后的 R^2 : $\bar{R}^2 = 1 - \frac{RSS/(N-K)}{TSS/(N-1)}$
 - \bar{R}^2 与 R^2 相比, 分别使用了残差 \hat{u} 和 y 的方差的无偏估计
 - R^2 随着解释变量的增加而单调变大, 而 \bar{R}^2 不会

信息准则

- 信息准则

- 赤池信息准则 (Akaike information creterion, AIC) : Akaike (1974) 通过Kullback - Leibler距离, 定义了如下信息准则:

$$AIC = -2\text{Log_Likelihood} + 2K$$

- 贝叶斯信息准则 (Bayesian information creterion, BIC) : Schwarz (1978) 通过贝叶斯法则提出了如下信息准则:

$$BIC = -2\text{Log_Likelihood} + \ln(N) K$$

线性回归中的信息准则

- 对于线性回归模型，如果使用极大似然估计，得到：

$$AIC = N \ln \left(\frac{\sum_{i=1}^N \hat{u}_i^2}{N} \right) + N + 2K$$

$$BIC = N \ln \left(\frac{\sum_{i=1}^N \hat{u}_i^2}{N} \right) + N + \ln(N) K$$

- 此外，小样本条件下可以使用调整的AIC，AIC_c（Hurvich和Tsai，1989）：

$$AIC_c = AIC + \frac{2K(K+1)}{N-K-1}$$

训练集、验证集和测试集

- 不管是欠拟合还是过拟合，都会导致样本外预测的误差变大，因而我们可以只使用一部分样本进行估计，而在另外一部分样本中检验模型的预测能力。
- 一种简单的做法是区分训练集（training data set）和验证集（validation data set）
 - 使用训练集的数据估计模型，并在验证集上计算预测误差的度量。
 - 最好的模型即在验证集上预测性能达到最好的模型。
- 此外，由于模型的选择使用了训练集和测试集，因而需要独立的另外一些样本用于最终模型性能的评估，这个独立于训练集和验证集的数据集即测试集（test data set），如果强调预测结果，可以在测试集上进行测试并展示。

模型选择

多项式阶数选择

在以下程序中，我们首先产生了一个伪数据集：

$$y = e^x + u$$

其中 $x \sim U(0, 3)$, $u \sim N(0, 4)$ 。接着，我们从 x 的一次方开始，逐渐向回归中添加 x^2, x^3, \dots, x^{10} 对模型进行拟合，并计算 R^2 、 \bar{R}^2 、AIC、BIC 以及留一验证的结果。

(model_selection.do cross_validation_reg.do)

模型选择

| 多项式阶数 | (1) R^2 | (2) \bar{R}^2 | (3) AIC | (4) BIC | (5) AICC | (6) MSE (CV) |
|-------|--------------|--------------------|---------------|---------------|---------------|-----------------|
| 1 | 0.8230 | 0.8194 | 242.83 | 245.83 | 242.26 | 7.471 |
| 2 | 0.9244 | 0.9212 | 201.49 | <u>207.23</u> | 202.01 | 3.268 |
| 3 | 0.9296 | <u>0.9240</u> | <u>200.60</u> | 208.25 | <u>201.49</u> | <u>3.187</u> |
| 4 | 0.9300 | 0.9238 | 201.60 | 211.16 | 202.97 | 3.295 |
| 5 | 0.9302 | 0.9223 | 203.47 | 214.95 | 205.43 | 3.445 |
| 6 | 0.9303 | 0.9205 | 205.46 | 218.84 | 208.12 | 3.656 |
| 7 | 0.9317 | 0.9202 | 206.43 | 221.73 | 209.94 | 3.628 |
| 8 | 0.9319 | 0.9205 | 208.27 | 225.48 | 212.77 | 3.586 |
| 9 | 0.9319 | 0.9186 | 210.24 | 229.36 | 215.88 | 3.819 |
| 10 | 0.9319 | 0.9186 | 212.24 | 233.28 | 219.19 | 3.854 |

作业

- 1.6、1.7、1.8、1.11、1.14、1.15

