线性回归中的模型选择准则
——$Model\ Trade\ off\ in\ Linear\ regression$

经济学系　郑丽珊 & 宋悦溪

March 15, 2019

$Foundations\ of\ Statistical\ Learning$ 已介绍数据拟合和模型的抉择问题，本文就其中一个视角—以线性回归为例—再现模型的 $tradeoff$，并引入常见的模型选择准则探讨人们如何进行权衡决策。同时，联系实际谈谈对机器学习的看法。

1　简单回顾-经典线性回归模型 OLS

1.1　基础假设

a. 线性：

$$Y = X\beta^o + \varepsilon$$

其中，$X_{n\times k} = (X_1, ..., X_n)'$，$Y_{n\times 1} = (Y_1, ..., Y_n)'$，$\varepsilon_{n\times k} = (\varepsilon_1, ..., \varepsilon_n)'$,$X_t = (1, X_{1t}, ..., X_{kt})'$

b. 严格外生性：

$$E(\varepsilon_t \mid X) = E(\varepsilon_t \mid X_1, ..., X_t, ..., X_n) = 0, t = 1, ..., n$$

c. 非奇异性：

(1) $X'X = \sum X_t X_t'$ 是非奇异的。

(2) 当 $n \to \infty$,$X'X$ 的最小特征值 $\lambda_{min}(X'X) \to \infty$ 的概率为 1。

d. 球形误差方差：

(1) 条件同方差：$E(\varepsilon_t^2 \mid X) = \sigma^2 > 0$，$t = 1, ..., n$

(2) 条件不相关：$E(\varepsilon_t\varepsilon_s \mid X) = 0$,$t \neq s$ ，$t, s \in 1, ..., n$

1.2 建立模型

为简便公式，本文选择 L2 函数，最小化残差和，即：

$$min\ SSR(\beta) = min\ (Y - X\beta^{'})^{'}(Y - X\beta^{'}) = min\ \sum(Y_t - X_t\beta)^2$$

通过计算，可得系数的 OLS 估计量：

$$\widehat{\beta} = arg\ minSSR(\beta) = (X^{'}X)^{-1}X^{'}Y$$

结合公式：

$$\widehat{Y}_t = X_t^{'}\widehat{\beta}$$

推出估计残差值：

$$e_t = Y_t - \widehat{Y}_t = (X_t^{'}\beta^o + \varepsilon_t) - X_t^{'}\widehat{\beta} = \varepsilon_t - X_t^{'}(\widehat{\beta} - \beta^o)$$

2 拟合优度指标

线性回归模型对数据拟合程度的优劣是我们十分关心的问题，因此我们通常使用如下两个指标来度量拟合优度，即辨别估计的回归线拟合真实 Y 值分布的好坏。

2.1 决定系数 R^2

决定系数的定义为

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum e_t^2}{\sum(Y_t - \overline{Y})^2}$$

其中 $\overline{Y} = \frac{1}{n}\sum Y_t$ 是样本均值。

R^2 的一个重要性质是，对于任何给定的随机样本，R^2 是线性回归模型中解释变量数目的非减函数。换一种说法讲，模型中解释变量 X 的数目越多，R^2 越大，无论新增加的 X_t 对 Y_t 是否有真正的解释力，均是如此，具体证明如下。

假设有如下两个线性回归模型，$Y_t = X_t^{'}\beta + \varepsilon_t$ 和 $Y_t = \widetilde{X}_t^{'}\gamma + u_t$。其中，

$$X_t = (1, X_{1t}, ..., X_{kt})^{'}$$

$$\widetilde{X}_t = (1, X_{1t}, ..., X_{kt}, X_{(k+1)t}, ..., X_{(k+q)t})^{'}$$

设 R_1^2 和 R_2^2 分别为两个模型的决定系数，e 是 Y 对 X 回归的残差向量，$\widetilde{e}$ 是 Y 对 $\widetilde{X}$ 回归的残差向量。根据拟合优度 R^2 定义可知，

$$R_1^2 = 1 - \frac{e^{'}e}{\sum(Y_t - \overline{Y})^2}$$
$$R_2^2 = 1 - \frac{\widetilde{e}^{'}\widetilde{e}}{\sum(Y_t - \overline{Y})^2}$$

因为 OLS 估计量 $\widehat{\gamma} = (\widetilde{x}^{'}\widetilde{x})^{-1}\widetilde{x}^{'}Y$ 是使扩展模型 $Y_t = \widetilde{X}_t^{'}\gamma + u_t$ 的 $SSR(\gamma)$ 最小化的最优解，因此对于任意的 $\gamma \in R^{K+q}$，有

$$\widetilde{e}^{'}\widetilde{e} = \sum(Y_t - \widetilde{X_t^{'}}\widehat{\gamma})^2 \leq \sum(Y_t - \widetilde{X_t^{'}}\gamma)^2$$

选择

$$\gamma = (\widehat{\beta}^{'}, 0^{'})^{'}$$

其中 $\widehat{\beta} = (X^{'}X)^{-1}X^{'}Y$ 是第一个回归模型 $Y_t = X_t^{'}\beta + \varepsilon_t$ 的 OLS 估计量，则有

$$\widetilde{e}^{'}\widetilde{e} \leq \sum(Y_t - \sum\widehat{\beta}_j X_{jt} - \sum 0 \cdot X_{jt})^2 = \sum(Y_t - X_t^{'}\widehat{\beta})^2 = e^{'}e$$

因此，有

$$R_1^2 \leq R_2^2$$

通过以上证明我们可以发现，对于被解释变量相同的模型，解释变量个数越多，R^2 越大，即使这些新增加的解释变量对被解释变量并没有真正的解释力，R^2 也会有所增加。鉴于此，在比较有相同被解释变量但不同数目解释变量的两个回归时，选择有更大 R^2 值的模型必须谨慎。

2.2 修正的决定系数 $\overline{R}^2$

由上述论证可知，要比较两个模型的 R2 项，必须考虑模型中解释变量 X_t 的个数。因此我们提出一个新的拟合优度指标-—修正的决定系数 $\overline{R}^2$，适用于更广泛的范围用来测度模型拟合情况。“修正”的含义为对 R^2 定义中平方和所涉及的自由度进行校正，因此定义

$$\overline{R}^2 = 1 - \frac{\sum e_t^2/(n-k)}{\sum(Y_t - \overline{Y})^2/(n-1)}$$

其中，k 表示模型中包括截距项在内的参数个数，n 为样本容量。容易得到，R^2 与 $\overline{R}^2$ 之间的关系为

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k}$$

不难发现：对于任意 $k > 1$，$\overline{R}^2 < R^2$。这就说明随着解释变量 X 个数的增加，修正的 R^2 比未修正的 R^2 增加得慢些。

2.3 两个 R^2 指标的比较

两种 R^2 均度量了回归线的拟合优度，即回归模型对被解释变量的变动解释的比例。但不论是用修正还是未修正的决定系数来比较不同模型好坏，都必须注意样本容量 n 以及被解释变量的相同性，否则不可进行比较。R^2 大可能说明模型拟合效果好，但也可能是因为模型中的解释变量多，即使其中的部分变量并没有任何解释力。这也是这两个指标的缺陷之一。

此外，R^2 测度的是一种关联性，与因果关系无关，有时候变量间的因果关系很弱或基本不存在时，也有可能获得较大的 R^2 值表明它们具有相似的变化趋势。也就是说，R^2 不能成为判断模型设定是否正确的标准，R^2 高不意味着模型设定合理，而正确设定的模型也不一定具有很高的 R^2。因此我们不能执着地追求更大的 R^2 值，而是该更多地关注解释变量与被解释变量的理论或逻辑关系，以及统计显著性。

3 模型选择准则

经过拟合优度指标的讨论，可以看出模型选择中存在两难抉择：一方面，所选择的模型在给定样本中对数据拟合程度要高（即样本内预测），往往要求添加更多的变量，因为解释变量越多，模型的系统偏差越小；另一方面，一个拟合模型在给定回归元值情况下对回归子未来值的预测（即样本外预测），往往要求较低的复杂性，因为给定随机样本容量下，参数越多，参数估计的准确性就越差。

因此，在模型的拟合优度与其复杂性（由回归元个数来判断）之间有一种权衡取舍的关系。基于此，模型准则不仅要最小化残差平方和 SSR（或提高 R^2 的值），同时也要对包含回归元个数的增加量进行惩罚。一般来说，常见的线性回归模型选择准则有 AIC 和 BIC 准则。

3.1 AIC 准则

AIC 的定义为

$$AIC = In(s^2) + \frac{2K}{n}$$

其中

$$s^2 = \frac{e^{'}e}{n-K}$$

式中，K 是自变量 X_t 的数量，s^2 是残差方差的估计量。AIC 第一项 $ln(s^2)$ 测度的是模型的拟合优度，第二项 $\frac{2K}{n}$ 测度的是模型的复杂程度。

3.2 BIC 准则

BIC 定义为

$$BIC = In(s^2) + \frac{Kln(n)}{n}$$

其中，BIC 第一项 $ln(s^2)$ 测度模型的拟合优度，第二项 $\frac{Kln(n)}{n}$ 测度模型的复杂程度。

3.3 AIC 准则 $v.s$ BIC 准则

3.3.1 相同点

在目标选择上，两种信息准则都试图在模型的拟合优度 $ln(s^2)$ 与尽量少用参数之间进行权衡；在比较两个或多个模型时，具有最低的 AIC 值 (BIC 值) 的模型效果最优；在使用范围上，两个准则不仅适用于样本内预测，还适用于预测一个回归模型在样本外的表现。

3.3.2 不同点

AIC 准则和 BIC 准则的区别主要在于对模型复杂度的惩罚方式不同，BIC 对模型复杂度施加的惩罚比 AIC 更严厉。因此，在模型偏好上，AIC 准则倾向于选择具有最优预测能力的模型，BIC 准则倾向于更加简单的线性回归模型; 在目标选择上，AIC 准则选择的参数更多，BIC 准则倾向于选择准确的 K 值。

综上所述，在采用模型选择准则时，应该根据样本量的情况细细斟酌。比如，当样本量趋于无穷时，应优先考虑 BIC 准则。因为 AIC 准则在时间序列很长的情况下，相关信息会更加分散，增加自变量个数才会提高拟合优度。但在实际中，当样本大小趋于无穷时，AIC 准则选择的拟合模型并不会收敛于真实模型，而且会使拟合模型具有比真实模型更多的未知变量个数。

4 模型设定误差

在大样本条件下，BIC 更接近于真实模型，AIC 准则倾向于接受过多参数的模型，这容易引发对模型参数个数的讨论。在统计学中，包含过多参数的模型被称为过度拟合 ($overfitting$)；而包含过少参数的模型被称为不足拟合（$underfitting$）。为了清楚认识到两种设定误差所带来的后果，本文将举例说明。

4.1 模型拟合不足（漏掉一个有关变量）

假设真实模型是

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t$$

但出于某种原因，拟合的模型为:

$$Y_t = \alpha_1 + \alpha_2 X_{2t} + v_t$$

漏掉 X_3 的后果将是:

a. 若漏掉的 X_3 和 X_2 存在相关性，一般不再有 $E(X_{2t}v_t) = 0$，导致 OLS 估计量不再是的 α 一致估计，同时 α_1 和 α_2 是有偏误的，往往这种偏误不会随着样本容量的增大而消失; 若 X_3 和 X_2 不存在相关性，尽管估计量无偏，但估计量仍有偏误。

b. 误差的方差将不被正确估计。同时，通常的置信区间和假设检验程序，对于所估计参数的统计显著性，容易得出误导性结论，且结果往往不可靠。

4.2 模型过度拟合（包含一个无关变量）

假设真实模型是

$$Y_t = \beta_1 + \beta_2 X_{2t} + u_t$$

但出于某种原因，拟合的模型为:

$$Y_t = \alpha_1 + \alpha_2 X_{2t} + \alpha_3 X_{3t} + v_t$$

多包含无关变量 X_3 将导致的后果有:

a. 拟合模型的全部参数的 OLS 估计量都是无偏而又一致的, 误差方差的估计是正确的, 通常的置信区间和假设检验程序仍然有效;

b. 然而通常情况下，拟合模型的参数估计量是非有效的，方差往往大于真实模型参数的方差。

4.3 评价

综上，权衡两种模型设定误差的后果，我们在建立模型时应避免漏掉有关变量，因为模型拟合不足导致拟合的 OLS 估计量既偏误且非一致；而过度拟合的唯一代价是，系数方差的估计值变大，导致对参数进行概率推断的精度降低。通过利弊权衡，人们往往会产生这样一个想法：宁可出现过度拟合，也不要出现拟合不足的现象。这种思想虽然

看似可取，但选择变量的个数最好是适量的，不偏不倚。

那么，现实中学者是否有方法来避免过度拟合呢？通常，学者会用以下两种方法：

a. 手工选取特征（$model\ selection\ algorithm$），留下重要的特征，减少变量的个数。但这种方法的缺点是会丢弃一些有用的信息。

b. 正则化（$regularization$）。保持所有的特征，但是减少系数的数量。当我们有很多特征，并且每个特征对结果 y 都有或多或少的一点影响的时候，这种方法则会表现地很好。

5 机器学习与模型抉择

近期热门的学科——机器学习 ($Machine\ Learning$)，是一门多领域交叉学科，涉及概率论、统计学、算法复杂度理论等多门学科。它是运用计算机模拟或实现人类的学习行为，以不断获取新的知识或技能。最优模型就是指可以很好地拟合已有数据集，并且正确预测未知数据的模型。而机器学习可以通过反复“学习”，帮助我们在模型空间时选择出最优模型。

科技进步让我们迈入能更准确预测未来的大数据时代，但是与此同时，科技的进步也成为我们预测未来的最大变数。很多人在学术研究中倾向于选择使用复杂的可以包含全部信息的模型。因为被普遍认同的逻辑是，信息越多，人们对未来的预测越准确，从而越能做出更合理的决策。但是上述讨论告诉我们，使用全部数据过度拟合出来的模型是十分危险的。

在信息爆炸的今天，海量的信息不断涌现，但是其中大部分可能并没有价值，只是“噪声”而已。如果我们简单地使用机器学习的方法将模型覆盖到全部信息，就难以捕捉到现实世界中真正存在的客观规律，同时也无法对未来进行有效预测。

总而言之，模型的复杂程度并不是决定模型好坏的唯一因素。在建立模型时，我们要重视数据背后存在的经济逻辑，避免出现拟合不足或过度拟合的情况。

6 参考文献

[1] 达摩达尔 · N · 古扎拉蒂 (2005). 计量经济学基础. 中国人民大学出版社.

[2] 洪永淼 (2011). 高级计量经济学. 高等教育出版社.

[3] 杰弗里 · M · 伍德里奇 (2009). 计量经济学导论. 中国人民大学出版社.

[4] 安格里斯特，皮施克 (2012). 基本无害的计量经济学. 格致出版社.