Advances in Applied Mathematics 应用数学进展, 2021, 10(1), 351-358
Published Online January 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.101040

Hans汉斯

基于AIC，BIC和CV准则对对数工资的探究

王　悦

云南财经大学统计与数学学院，云南　昆明
Email: 43784856@qq.com

收稿日期：2020年12月25日；录用日期：2021年1月19日；发布日期：2021年1月28日

摘　要

AIC信息准则，BIC信息准则和Cross-Validation (交叉验证法，简称CV)是统计学中模型选择和评价的重要工具。本文为研究对数工资与周平均时长、IQ得分、世界工作知识得分、受教育年限等11个协变量之间的关系，从无协变量的模型经每次引入一个协变量共产生了12个备选的线性模型。基于AIC，BIC，CV三个准则分别在全样本上选择最优模型。在训练集上选最优模型，在测试集上进行误差分析对三个准则进行评判。其中，三个准则对应的值越小越好。在全样本上，基于三个准则选择的模型均是11个协变量共存的线性模型。将全样本分为训练集和测试集，基于AIC，BIC和CV三个准则，经过1000次运算，选择11个协变量共存的模型的概率分别为100%，99%，100%。对对数工资的探究，最优线模型为11个协变量共存的模型，并且三个准则的表现无明显差异。

关键词

AIC，BIC，CV，最优模型

Exploration of Logarithmic Wages Based on AIC, BIC and CV Criteria

Yue Wang

College of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan
Email: 43784856@qq.com

Received: Dec. 25th, 2020; accepted: Jan. 19th, 2021; published: Jan. 28th, 2021

Abstract

AIC information criterion, information criterion and Cross-Validation (cross-validation method, abbreviated as) are important tools for model selection and evaluation in statistics. This article is

文章引用: 王悦. 基于 AIC, BIC 和 CV 准则对对数工资的探究[J]. 应用数学进展, 2021, 10(1): 351-358.
DOI: 10.12677/aam.2021.101040

to study the relationship between logarithmic wages and 11 covariates such as average weekly length, IQ score, world work knowledge score, years of education, etc. A total of 12 covariates are generated from the model without covariates and one covariate is introduced each time. The optimal models were selected on the full sample based on the three criteria of AIC, BIC, and CV. The optimal model is selected on the training set, and error analysis is performed on the test set to judge the three criteria. For all three criteria, their corresponding values are as small as possible. In the full sample, the models selected based on the three criteria are all linear models with 11 covariates coexisting. The full sample is divided into training set and test set. Based on the three criteria of AIC, BIC and CV, after 1000 operations, the probability of selecting a model with 11 covariates coexisting is 100%, 99%, and 100%, respectively. In the exploration of logarithmic wages, the optimal line model is a model in which 11 covariates coexist, and there is no significant difference in the performance of the three criteria.

Keywords

AIC, BIC, CV, Optimal Model

1. 引言

1.1. 研究背景及意义

随着社会经济的快速发展，人们在解决温饱问题后，对物质的追求和精神的追求越来越高。所以，人们在享受生活便利的同时，在一定程度上增加了经济压力。所以对工资因素的研究是有必要的，供人们根据自己的实际情况选择工作环境提供一个参考。在本文中，探究对数工资与周平均时长、人种、居住地、工作方位等因素之间的关系。

1.2. 相关知识介绍

1.2.1. 多元线性回归分析

多元线性回归研究的是一个变量与多个变量之间的关系。

1.2.2. 最小二乘法

最小二乘法用于估计多元线性回归模型中协变量的系数。

1.2.3. AIC 信息准则

AIC 信息准则是衡量统计模型拟合优良性的一种标准，可以权衡所估计模型的复杂度和此模型拟合数据的优良性, 全称是最小信息量准则[1]。

1.2.4. BIC 信息准则

BIC 准则全称贝叶斯信息准则与 AIC 信息准则相似，用于模型选择[2]。

1.2.5. Cross-Validation

Cross-Validation 根据模型的预测能力选择模型的一种方法[3]。将样本分为训练集和测试集，在训练集上进行模型选择，在测试集上预测误差。

1.3. 文献综述

Jun Shao [3]在 1993 年通过交叉验证法选择选择线性模型。虽然交叉验证法测试为一个样本的很受欢迎，也比较方便，但是存在一定的缺陷，随着样本量的增加，最优模型的概率不趋近于 1。因此 Jun Shao [3]做了与去一交叉验证法相对的验证法，训练集样本数量减少，测试集样本量增多，有效的弥补了去一交叉验证法的缺陷[3]。而当预测变量的样本量相对于总的样本量较大时，AIC，BIC 可能会出现倾向于过度拟合的问题，因此 Yuhong Yang [4]提出了修正的 AIC 准则，即 AICc，能有效避免由于变量过多导致的过度拟合的现象[4]。

2. 相关理论介绍

2.1. 最小二乘法

基于多元回归模型的最小二乘法，方法是使得真实值与预测值之间的差距达到最小。

Step 1. 模型的建立

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_px_p+\varepsilon \tag{1}$$

在这里，模型中的 $x_1,x_2,\cdots,x_p$ 是协变量，未知参数 $\beta_0,\beta_1,\cdots,\beta_p$ 模型参数，ε 是服从正态分布的随机误差项。

Step 2. 参数估计

记目标函数为如下(2)式：

$$\phi\left(\beta_0,\beta_1,\cdots\beta_p\right)=\min\left\{\sum_{i=1}^{n}\left[y_i-\left(\hat{\beta}_0+\hat{\beta}_1x_{i1}+\hat{\beta}_2x_{i2}+\cdots+\hat{\beta}_px_{ip}\right)\right]^2\right\},i=1,2,\cdots,n \tag{2}$$

用(2)式对未知参数 $\beta_0,\beta_1,\cdots,\beta_p$ 求偏导，并令其等于 0 可得如下(3)式：

$$\begin{cases}\partial\phi/\partial\beta_0=-2\sum_{i=1}^{n}\left[y_i-\left(\beta_0+\beta_1x_{i1}+\beta_2\beta x_{i2}+\cdots+\beta_px_{ip}\right)\right]=0\\ \partial\phi/\partial\beta_1=-2\sum_{i=1}^{n}\left[y_i-\left(\beta_0+\beta_1x_{i1}+\beta_2\beta x_{i2}+\cdots+\beta_px_{ip}\right)\right]x_{i1}=0\\ \vdots\\ \partial\phi/\partial\beta_p=-2\sum_{i=1}^{n}\left[y_i-\left(\beta_0+\beta_1x_{i1}+\beta_2\beta x_{i2}+\cdots+\beta_px_{ip}\right)\right]x_{ip}=0\end{cases} \tag{3}$$

通过(3)式可得未知参数的估计表达式如下(4)式：

$$\hat{\beta}=\left(X'X\right)^{-1}X'y \tag{4}$$

其中，$\hat{\beta}=\left(\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p\right)'$，$X=\left(x_1,x_2,\cdots,x_p\right)$，$x_i=\left(x_{i1},x_{i2},\cdots,x_{ip}\right),i=1,2,\cdots,n$。

2.2. AIC 信息准则

Step 1. 模型的建立

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip}+\varepsilon_i,\varepsilon_i\sim N\left(0,\sigma^2\right) \tag{5}$$

Step 2. AIC 准则的建立

对于 Step1 中的模型，$y_i,i=1,2,\cdots,n$ 是相互独立的，所以其极大似然函数为：

$$L=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{\frac{-\left[y_i-\left(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip}\right)\right]^2}{2\sigma^2}\right\} \tag{6}$$

根据(6)式，其对数似然函数为：

$$\ln L=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln\sigma^2-\sum_{i=1}^{n}\frac{\left[y_i-\left(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip}\right)\right]^2}{2\sigma^2} \tag{7}$$

通过极大化模型的对数似然函数(7)式可得：

$$\sigma^2=\sum_{i=1}^{n}\frac{\left[y_i-\left(\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_px_{ip}\right)\right]^2}{n} \tag{8}$$

$$\mathrm{AIC}=-2\ln L\left(\beta,\sigma^2\mid y\right)+2(p+1) \tag{9}$$

(9)式为 AIC 信息准则的判别式，其中未知参数 β 由最小二乘法得到，p 为未知变量的个数。

2.3. BIC 信息准则

BIC 准则的表达式的建立类似于 AIC。

$$\mathrm{BIC}=-2\ln L\left(\beta,\sigma^2\mid y\right)+(p+1)\ln(n) \tag{10}$$

2.4. Cross-Validation

在本文中，采用的是每次剔除一个样本并将其作为测试集，剩余的作为训练集的 Cross-Validation。将其在测试集上的误差的平法和作为模型选择的标准。

Step 1. 将数据划分为训练集和测试集两部分。

剔除一个样本，将其作为测试集，将剩余的 $n-1$ 个样本作为训练集。

Step 2. 计算测试集上的误差的平方如下(11)式

$$pe_i=\left\{y_i-\frac{x'_{\neq i}x_{\neq i}}{x_{\neq i}y}\right\}^2 \tag{11}$$

Step 3. 重复 Step1 和 Step2

重复 Step1 和 Step2，将得到 $n-1$ 个误差的平方。并将这 $n-1$ 个误差的平方和作为选模型的准则。

$$pe=\sum_{i=1}^{n}pe_i \tag{12}$$

即为一个模型对应的 pe 越小，说明该模型越好。

3. 实验

3.1. 数据的探索性分析

3.1.1. 数据介绍

本文研究的是对数工资(lwage)与 11 变量之间的关系。这 11 个变量分别为每周的平均时数(hours)、IQ 得分(IQ)、世界工作分数知识(KWW)、教育年限(educ)、工作的年限(exper)、目前雇主年资(tenure)、岁数(age)、是否已婚(married)、是否是黑人(black)、是否位于南方(south)、是否住在 SMSA(urban)。并且把这 11 个协变量分别记为 $x_1,x_2,\cdots,x_{11}$，将响应变量 lwage 记为 y。

3.1.2. 通过协变量的协方差阵的变量的箱线图探究数据

协变量之间的相关系数矩阵。表 1 为协变量的系数矩阵每一列绝对值的最大值和对应的变量(除本身外)。

Table 1. Correlation coefficients between variables
表 1. 变量间的相关系数

变量	绝对值	真实值
(x_1, x_3)	0.1139	0.1139
(x_2, x_4)	0.5157	0.5157
(x_3, x_2)	0.4135	0.4135
(x_4, x_2)	0.5157	0.5157
(x_5, x_7)	0.4953	0.4953
(x_6, x_7)	0.2706	0.2706
(x_7, x_5)	0.4953	0.4953
(x_8, x_7)	0.1070	0.1070
(x_9, x_2)	0.3879	−0.3879
(x_{10}, x_9)	0.2365	0.2365
(x_{11}, x_{10})	0.1099	−0.1099

从表 1 可知，x_2 和 x_4 的相关系数是最大的为 0.5157，x_5 和 x_7 相关系数为 0.4953，x_3 和 x_2 的相关系数为 0.4135。其余相关系数较小，说明在做线性模型时，当 x_2 、x_4 和 x_3 共存或 x_5 和 x_7 共存时，模型可能会存在一定的共线性。

3.2. 模型的建立

在本文中，共验证了 12 线性模型。从无变量入选模型经每次入选一个变量，一共有 12 个线性模型。分别为：

模型 1：$y_i = \beta_0 + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$

模型 2：$y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$

模型 3：$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$

模型 4：$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$

模型 5：$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$

模型 6：$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$

…

模型 12：$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \cdots + \beta_{11} x_{i11} + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2)$

3.3. 根据所有的 935 个样本点，使用 AIC、BIC 和 CV 做模型选择

通过 MATLAB 计算 12 个模型对应的 AIC、BIC 和 CV 值。

3.3.1. 基于 AIC 准则和 BIC 准则的模型选择

根据全部样本，采用最小二乘法确定模型参数，并根据(9)式，通过 MATLAB 软件可计算出每个模型对应的 AIC 值和 BIC 值如下表 2。

由表 2 可知，第 12 个模型对应的 AIC 值为 759.270，对应的 BIC 的值为 822.197。在所有模型中模型 12 对应的 AIC 值和 BIC 值是最小的，因此通过 AIC 准则和 BIC 准则，选择的最优模型均是第 12 个模型。

Table 2. AIC value and BIC value of each model
表 2. 各模型的 AIC 值和 BIC 值

模型编号	AIC 值	BIC 值
1	1039.275	1048.956
2	1039.188	1053.709
3	940.511	959.873
4	899.478	923.681
5	880.183	909.226
6	859.210	893.094
7	842.402	881.127
8	843.434	886.999
9	823.755	872.160
10	814.156	867.402
11	800.787	858.874
12	759.270	822.197

3.3.2. 基于 CV 值的模型选择

通过 MATLAB 软件可计算出每个模型对应的 CV 值如下表 3。

Table 3. CV value corresponding to each model
表 3. 各模型对应的 CV 值

模型编号	CV 值
1	166.011
2	166.139
3	149.499
4	143.105
5	140.211
6	137.106
7	134.706
8	134.865
9	132.057
10	130.697
11	128.865
12	123.271

由表 3 可知，第 12 个模型对应的 CV 的值最小为 123.271，在所有模型中是最小的。因此，通过交叉验证，选择的最优模型是第 12 个模型。

3.4. 根据训练集和测试集的模型选择

在本文中，将数据集随机地分为训练集和测试集，其中训练集有 500 个样本，测试集有 435 个样本。

在训练集上做模型选择，在测试集上做误差分析。在本节实验中共做 1000 次实验。

下表 4 为每一次实验过程中，所选出的 12 个模型中最小的 AIC 值、BIC 值以及 CV 值。在这里，只展示部分结果。其中，C1AIC 为 12 个模型中 AIC 值的最小值。C1BIC 为 12 个模型中 BIC 值的最小值。C1CV 为 12 个模型中 CV 值的最小值。括号中的数字代表的是模型序号。

Table 4. The corresponding AIC, BIC and CV values of each model on the training set
表 4. 各模型在训练集上对应的 AIC、BIC 以及 CV 值

C1AIC 值	C1BIC 值	C1CV 值
341.95 (12)	396.74 (12)	58.04 (12)
469.66 (12)	524.45 (12)	74.838 (12)
408.15 (12)	462.94 (12)	66.08 (12)
431.62 (12)	486.41 (12)	62.315 (12)
378.73 (12)	433.52 (12)	62.32 (12)
456.61 (12)	511.40 (12)	72.82 (12)
451.68 (12)	506.47 (12)	72.29 (12)
386.75 (12)	441.54 (12)	63.51 (12)
419.63 (12)	474.42 (12)	61.63 (12)
368.37 (12)	423.16 (12)	61.17 (12)

由表 4 可以看出，模型 12 对应的 AIC 值、BIC 值以及 CV 值是最小的。

下表 5 为在 1000 次实验过程中，三个准则选择模型 12 的概率。

Table 5. Selection of Model 12 based on three criteria
表 5. 基于三个准则对模型 12 的选择情况

准则类别	AIC	BIC	CV
选择模型 12 的概率	1	0.9900	1

由表 5 可知，基于 AIC 准则选择模型 12 的概率为 100%；基于 BIC 准则选择模型 12 的概率为 99%；基于交叉验证选择模型的 12 的概率为 100%。因此，基于在本文中选用的模型是模型 12。

3.5. 结合经济学解释 3.3 节和 3.4 节选出的最优模型

由 3.3 节可知，在全部样本下，AIC 准则、BIC 准则以及 CV 选择的均是第 12 个模型，基于最小二乘法确定第 12 个模型，如下：

$$\begin{aligned} y = {} & 5.2797 - 0.0056x_1 + 0.0033x_2 + 0.0035x_3 + 0.0492x_4 + 0.0105x_5 \\ & + 0.0100x_6 + 0.0054x_7 + 0.1950x_8 - 0.1419x_9 - 0.815x_{10} + 0.054x_{11} \end{aligned} \tag{13}$$

由(13)式可知，在其他条件不变的情况下，当 x_1 每增加一个单位，即每周平均时数每增加一个单位，对数工资 y 平均减少 0.0056 单位。当 x_2 每增加一个单位，即 IQ 每增加一个单位，对数工资 y 平均增加 0.0033 个单位。当 x_3 增加一个单位，即世界工作分数知识每增加一个单位，对数工资平均增加 0.0035 个单位。当 x_4 每增加一个单位，即教育年限每增加一个单位，对数工资平均增加 0.0492 个单位。当 x_5 每增加一个单位，即工作的年限每增加一个单位，对数工资平均增加 0.0105 个单位。当 x_6 每增加一个单位，

即目前雇主年资每增加一个单位，对数工资平均增加 0.0100 个单位。当 x_7 每增加一个单位，即岁数每增加一个单位，对数工资平均增加 0.0054 个单位。当 $x_8=1$，即已婚，对数工资平均增加 0.1950 个单位。当 $x_9=1$，即是黑人，对数工资平均减少 0.1419 个单位。当 $x_{10}=1$，即是住在南方，对数工资平均减少 0.0815 个单位。当 $x_{11}=1$，即居住 SMSA，对数工资平均增加 0.1775 个单位。

4. 结论

在本文中，依据 WAGE2.XLS 数据集，对对数工资和其他协变量之间的线性关系进行探究。从无协变量引入模型经每次入选一个协变量共产生了 12 个备选线性模型。

1) 在所有的 935 个样本上，从 12 个备选线性模型中，基于 AIC、BIC 和 CV 准则，选择的最优模型均是第 12 个模型，即为引入所有的协变量。通过最小二乘法可以确定第 12 个模型。

2) 将 935 个样本分为训练集和测试集，其中训练集有 500 个样本，测试集有 435 个样本。在训练集上，基于 AIC、BIC 和 CV 准则，从 12 个备选线性模型中选择最优模型。经过 1000 次运算，基于 AIC 信息准则选中第 12 个模型的概率为 100%，基于 BIC 信息准则选择第 12 个模型的概率为 99%，基于 CV 准则选中第 12 个模型的概率为 100%。所以，针对对数工资与其余变量之间关系的这个数据集，三个准则的效果是一样的。

参考文献

[1] https://baike.baidu.com/item/AIC/10910647

[2] https://blog.csdn.net/lynnucas/article/details/47947943

[3] Shao, J. (1993) Linear Model Selection by Cross-Validation. *Journal of the American statistical Association*, **88**, 486-494. https://doi.org/10.1080/01621459.1993.10476299

[4] Yang, Y. (2005) Can the Strengths of AIC and BIC Be Shared? A Conflict between Model Indentification and Regression Estimation. *Biometrika*, **92**, 937-950. https://doi.org/10.1093/biomet/92.4.937