Statistics and Application 统计学与应用, 2020, 9(4), 546-565
Published Online August 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.94059

Hans汉斯

Model Selection Based on AIC, BIC, CV Criteria

Junyan Wang

School of Statistics and Mathematics, Yunnan University of Finance and Economics, Kunming Yunnan
Email: 2468616148@qq.com

Received: Jul. 13th, 2020; accepted: Jul. 27th, 2020; published: Aug. 4th, 2020

Abstract

As we all know, a good model must not only have a good fit, but also have a concise form, so how to balance the accuracy and complexity of the model requires model selection. And AIC, BIC and CV can just balance the relationship of the model, which just solves the problem of current model selection. This article chooses models based on AIC, BIC and CV criteria, and models WAGE2 data. Firstly, some distribution characteristics and correlations of the data are obtained through simple statistics and statistical mapping for the 12 variables. Then, we use AIC, BIC and CV to conduct statistical modeling, select the optimal model, and use least squares method to find the fitting equation, and then explain the economic significance. Finally, in order to make the selection of the model more convincing, the optimal model was selected by repeating 1000 experiments, and compared with the previous model to obtain a consistent optimal model.

Keywords

Model Selection, Accuracy, Complexity, AIC, BIC, CV, Least Square Method

基于AIC, BIC, CV准则的模型选择

王俊艳

云南财经大学，统计与数学学院，云南 昆明
Email: 2468616148@qq.com

收稿日期：2020年7月13日；录用日期：2020年7月27日；发布日期：2020年8月4日

摘 要

众所周知，一个好的模型不仅要具有优良的拟合度，而且还要具有简洁的形式，那么究竟怎样平衡模型的精确度与复杂度呢，这就需要进行模型选择了。而AIC，BIC及CV恰好能平衡模型这种关系，恰好解决

文章引用: 王俊艳. 基于 AIC, BIC, CV 准则的模型选择[J]. 统计学与应用, 2020, 9(4): 546-565.
DOI: 10.12677/sa.2020.94059

了当下模型选择的难题。本文依据AIC, BIC及CV准则来进行模型选择，对WAGE2数据进行建模。首先对12个变量通过简单统计量及其统计作图得到数据的一些分布特征及其相关关系。接着用AIC, BIC及CV来进行统计建模，选出最优模型，并用最小二乘法求得拟合方程，然后进行经济学意义的解释。最后为了让模型的选择更具有说服力，重复1000次实验选出最优模型，与之前的模型进行比较，得到一致最优的模型。

关键词

模型选择，精确度，复杂度，AIC，BIC，CV，最小二乘法

1. 绪论

1.1. 研究背景及意义

随着社会科学的发展，模型选择是很多研究学者讨论的一个重要话题，究竟应该怎样评判一个模型的好坏呢，它的拟合优度及其参数选择怎样才更合理呢，这是一个值得研究与探索的问题。

模型选择一定伴随着参数估计的问题，有很多学者采用极大似然函数作为目标函数，这也是评判拟合优度的一个标准。为了提高模型的精度，我们可以选用较多的训练样本，但是，一般情况下，模型精度的提高伴随着另一个问题就是模型的复杂度变大了，为此，就可能出现另外一种结果，出现过度拟合的情况，因此，对于模型的选择是一个迫切需要研究的课题，要怎样在模型的精确度与复杂度之间平衡呢？AIC, BIC [1]及其 CV [3]在这方面就很好的平衡了这两者之间的关系，而本文就依托这三个模型选择的准则来选择最优模型。

1.2. 文献综述

关于模型识别国内外有很多学者做了相关研究，特别是关于 AIC, BIC 准则的模型识别。

YUHONG YANG [1]在 AIC 和 BIC 的优势可以共享吗？模型辨识与回归估计的关系这篇文章中提出在模型选择中，BIC 在选择真模型时是一致的，AIC 在估计回归函数时是最优的极大极小率。最近的一个发展方向是自适应模型选择，与 AIC 和 BIC 相比，惩罚项是数据相关的。在自适应模型选择的支持下，已经取得了一些理论和实证结果，但目前还不清楚它是否能真正共享 AIC 和 BIC 模型的结合或平均的强度，已引起越来越多的关注，这是克服模型选择不确定性的一种方法，贝叶斯模型平均值是否是估计模型的最佳方法？最小极大意义上的回归函数？我们发现，这些问题的答案基本上是否定的：对于任何一个模型选择准则都是一致的，它必须表现出次优的行为来估计覆盖率极小极大值项下的回归函数；而贝叶斯模型平均不能成为回归估计的极小极大值。Cheryl J. Flynn [2]在规范化参数选择的效率——误判模式的惩罚似然估计这篇文章中提出，在经典回归中，当最大候选模型的维数与样本量之间存在较大的相关性时，AIC 往往会选择过于复杂的模型，仿真研究表明，AIC 在使用惩罚回归时，会有一些缺点。因此，提出了使用经典校正 AIC(AICC)作为替代方案，并证明它保持了所需的渐近性质。Jun Shao [3]在交叉验证发的线性模型的选择这篇文章中提出可以通过使用遗漏 n 交叉验证，可以纠正遗漏 1 交叉验证的一致性，并且给出了使用遗漏交叉验证方法的动机、理由和一些实用性的讨论，并给出了仿真研究的结果。

1.3. 研究问题概述

本文主要的研究问题有 5 个：

1)：对给出的 12 个变量进行描述性的统计分析；

2)：依托全部数据用 AIC, BIC, CV 准则从 12 个模型中选出最优模型；

3)：对最优模型用最小二乘法进行拟合，求得参数，获得模型的方程。

4)：对所获得的最优方程进行经济学意义的解释。

5)：重复进行 1000 次实验，把数据分为训练集与测试集，用 AIC, BIC, CV 来进行模型选择，选出最优模型。

1.4. 研究思路和行文框架

本文的研究思路是先对对数据进行描述性统计分析，然后再利用 AIC, BIC 及 CV 准则来进行模型选择。

本文具体的行文安排如下：第一章绪论部分，从模型选择的研究意义出发说明研究本文的必要性，然后分析了模型选择的研究现状，并给出本文的研究思路及行文框架。第二章主要是相关知识准备，主要包括 AIC, BIC 及 CV 的简介原理及其实现步骤。第三章是对数据进行描述性的统计分析。第四章主要依托第二章的相关知识用 AIC, BIC 及 CV 这三个准则进行模型选择。先用这三个准则进行模型选择，接着再用最小二乘法对最优模型进行拟合，求得参数，获得最优方程，并且对最优方程进行经济学解释。最后，为了让结果更具有说服力，重复进行 1000 次实验，来选出最优模型。第五章为研究结论，主要是针对本文所做的分析做一个总结。

2. 相关知识准备

2.1. AIC, BIC, CV 的简介与原理

1) AIC 的简介及原理

AIC 是赤池信息准则的简称，是日本的一个统计学家赤池宏次提出的，它的用途是衡量统计模型拟合度是否优良，对多个模型做出选择判别。不仅如此，它在估计模型的复杂度方面也有很大的用途。AIC 准则主要在熵的基础上建立的，一般情况下，认为 AIC 越小，所对应的模型拟合度越好，模型越精确。[4]

AIC 的一般表达式为：

$$\mathrm{AIC} = \left(2k - 2L\right)/n \tag{1}$$

k 表示的是模型中参数的个数，L 表示的是对数似然函数，n 是样本量。

我们要想选取 AIC 最小的那个模型，需要做到两点：

一是要提高极大似然函数的拟合度，即提高模型的拟合度。

二是：要降低过度拟合的可能性，这就需要加入惩罚项，使模型的参数尽可能少。

显然，AIC 准则在合理控制了参数的同时也使得似然函数尽可能大，模型的拟合度尽可能高。

特别注意的是 AIC 的使用条件一定是在误差项服从正太分布的情况下。

2) BIC 的简介及原理

BIC 是贝叶斯信息准则的简称，是 Schwarz 提出的，它与 AIC 准则相似，也是用于模型选择。当增加参数 k 的数量时，就增加了模型的复杂度，似然函数也会增大，与 AIC 相似，也易导致过度拟合的现象。针对此现象，AIC, BIC 的处理方式相似，都引入了与参数相关的惩罚项，但是 BIC 的惩罚项会更大

一点相对 AIC 而言，因此，考虑了样本量，样本量较大时，就有效的解决了由于模型精度过高导致的复杂度也较高的问题。

BIC 的一般表达式为：

$$\mathrm{BIC} = k\ln(n) - 2\ln(L) \tag{2}$$

k 表示的是模型中参数的个数，L 表示的是对数似然函数，n 是样本量，$k\ln(n)$ 表示惩罚项。

3) CV 的简介及原理

CV 是交叉验证法的简称，它也是一种分类的统计分析方法，它的基本思想是对原始的数据集分组为两部分，训练集与验证集。先对训练集进行训练，然后再用验证集对训练的模型进行测试，进一步进行分类评价。

最常见的 CV 方法主要有 2 种：

a) 一种是将原始数据进行分组，将其中的一组数据作为验证，其余的 K-1 组作为训练集，这样就可以得到 k 个模型，一般情况 k 大于 2，分类结果还是相对有效的。

b) 另外一种方法与第一种方法的不同是将每个样本都做一次验证集，剩下的全部样本作为训练集，假设有 n 个样本，则共有 n 个模型。最终可以取分类准确率的平均数来作为分类的性能指标。几乎用上了所有样本作为训练集，最接近原始的样本，几乎没有信息损失，结果更为可靠，这种方法更受欢迎。

本文的 CV 方法采用的就是第二种方法。

2.2. AIC, BIC, CV 的实现步骤

1) AIC, BIC 模型的实现步骤

a) 计算总体的概率密度。假设 $y_i, i = 1,\cdots,n$，是来自总体 $g(y)$ 的样本。

$$y_i = X_i\beta + \varepsilon_i \tag{3}$$

其中 $y_i \sim N(X_i', \sigma^2)$，$\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$

若 y_i 的密度函数为

$$f(y_i \mid \beta,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(y_i - X_i'\beta)^2}{2\sigma^2}\right) \tag{4}$$

则似然函数为

$$L = f(y \mid \beta,\sigma^2) = \prod_1^n f(y_i \mid \beta,\sigma^2) = \prod_{i=1}^n \left\{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-(y_i - X_i'\beta)^2}{2\sigma^2}\right)\right\} \tag{5}$$

b) 计算对数似然函数。对数似然函数为

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \left\{\frac{\sum_{i=1}^n (y_i - X_i'\beta)^2}{2\sigma^2}\right\} \tag{6}$$

c) 求出参数 σ^2 的极大似然估计

$$\hat{\sigma}_2 = \frac{\sum_{i=1}^n (y_i - X_i'\beta)^2}{n} \tag{7}$$

d) 计算 β 的最小二乘估计

在正太分布的情况，β 的极大似然估计与最小二乘估计无大的区别。在本文中用的是最小二乘估计。

$$\hat{\beta}=(X'X)^{-1}X'y \tag{8}$$

e) 计算 AIC 与 BIC。将计算得到的参数估计值代入计算 AIC, BIC

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

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

其中 p 表示的是参数的数目，n 表示的是总的样本量。

2) CV 模型的实现步骤

设总的数据的样本书目为 n，y_i 为因变量，x_i 为自变量，用 CV 来选择模型。

第一步：删除 (x_1,y_1)，用 $(x_2,y_2),\cdots,(x_n,y_n)$ 作为训练集，用最小二乘法来做参数估计。

$$\hat{\beta}_{[-1]}=(X'X)^{-1}X'y \tag{11}$$

其中 $\hat{\beta}_{[-1]}$ 表示已经去除了第一个样本的参数估计值。预测误差为 $\left(y_1-x_1'\hat{\beta}_{[-1]}\right)^2$。

第二步：删除 (x_2,y_2)，用 $(x_1,y_1),(x_3,y_3),\cdots,(x_n,y_n)$ 作为训练集，用最小二乘法来做参数估计。

$$\hat{\beta}_{[-2]}=(X'X)^{-1}X'y \tag{12}$$

其中 $\hat{\beta}_{[-2]}$ 表示已经去除了第二个样本的参数估计值。

预测误差为 $\left(y_2-x_2'\hat{\beta}_{[-2]}\right)^2$。

第三步往后，依次删除一个样本，用这个删除的样本作为测试集，用剩下的样本来做参数估计，并求得预测误差。

最后一步：删除 (x_n,y_n)，用 $(x_1,y_1),\cdots,(x_{n-1},y_{n-1})$ 作为训练集，用最小二乘法来做参数估计。

$$\hat{\beta}_{[-n]}=(X'X)^{-1}X'y \tag{13}$$

其中 $\hat{\beta}_{[-2]}$ 表示已经去除了第二个样本的参数估计值。

预测误差为 $\left(y_n-x_n'\hat{\beta}_{[-n]}\right)^2$

总共进行了 n 次，则有 n 个预测误差。

计算累加的误差之和。

$$\mathrm{CV}=\sum_1^n\left(y_i-x_i'\hat{\beta}_{[-i]}\right)^2 \tag{14}$$

则 CV 就是当前模型的误差。

如果有 k 个模型，只需要比较这 k 个 CV 值，选择最小 CV 值对应的模型就是最优的模型。

3. 对数据的描述性统计分析

3.1. 数据准备

本文的数据来源于 WAGE2，共有 935 个样本，12 个变量，本文主要对这 12 个变量进行数据分析，变量说明如下表 1。

Table 1. Description of variables in wage2
表 1. WAGE2 中的变量说明

y	x1	x2	x3	x4	x5
lwage	hours	IQ	KWW	educ	exper
x6	x7	x8	x9	x10	x11
tenure	age	married	black	south	urban

3.2. 统计特征

运用 MATLAB 来进行统计分析，运行代码，得到的统计结果如下表 2~5，代码见附录 1。

1) 平均值

Table 2. Average values of variables in wage2
表 2. WAGE2 中的变量的平均值

y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11
6.779	43.929	101.28	35.744	13.468	11.564	7.2342	33.08	0.893	0.1283	0.3412	0.7176

2) 中位数

Table 3. Median of variables in wage2
表 3. WAGE2 中的变量的中位数

y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11
6.8079	40	102	37	12	11	7	33	1	0	0	1

3) 众数

Table 4. Modes of variables in wage2
表 4. WAGE2 中的变量的众数

y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11
6.9078	40	96	38	12	11	1	30	1	0	0	1

4) 方差

Table 5. Variance of variables in wage2
表 5. WAGE2 中的变量的方差

y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11
0.1774	52.19	226.58	58.351	4.8253	19.137	25.758	9.6584	0.0956	0.112	0.225	0.2028

通过(1) (2) (3) (4)平均数，中位数，众数，与方差的数据特征，我们可以发现，y，x4，x8，x9，x10，x11 的数据方差比较小，说明数据分布是较为集中的，它们的中位数与众数都是几乎相同的，说明在中位数附近的数据是较为集中的。

5) 图 1 是 12 个变量的相关系数矩阵

	y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11
y	1	−0.047	0.3148	0.3063	0.3121	0.0206	0.1859	0.1618	0.15	−0.232	−0.195	0.2038
x1	−0.047	1	0.0738	0.1139	0.091	−0.062	−0.056	0.0248	0.0326	−0.108	−0.03	0.0166
x2	0.3148	0.0738	1	0.4135	0.5157	−0.225	0.0422	−0.044	−0.015	−0.388	−0.21	0.0389
x3	0.3063	0.1139	0.4135	1	0.3881	0.0175	0.1414	0.3931	0.0899	−0.281	−0.094	0.0982
x4	0.3121	0.091	0.5157	0.3881	1	−0.456	−0.036	−0.012	−0.059	−0.179	−0.097	0.0722
x5	0.0206	−0.062	−0.225	0.0175	−0.456	1	0.2437	0.4953	0.1063	0.0558	0.0213	−0.047
x6	0.1859	−0.056	0.0422	0.1414	−0.036	0.2437	1	0.2706	0.0726	−0.078	−0.062	−0.038
x7	0.1618	0.0248	−0.044	0.3931	−0.012	0.4953	0.2706	1	0.107	−0.036	−0.029	−0.007
x8	0.15	0.0326	−0.015	0.0899	−0.059	0.1063	0.0726	0.107	1	−0.053	0.0228	−0.04
x9	−0.232	−0.108	−0.388	−0.281	−0.179	0.0558	−0.078	−0.036	−0.053	1	0.2365	0.0702
x10	−0.195	−0.03	−0.21	−0.094	−0.097	0.0213	−0.062	−0.029	0.0228	0.2365	1	−0.11
x11	0.2038	0.0166	0.0389	0.0982	0.0722	−0.047	−0.038	−0.007	−0.04	0.0702	−0.11	1

Figure 1. Correlation coefficients of variables in wage2
图 1. WAGE2 中的变量的相关系数

由 12 个变量的相关系数图可以看出，对数工资 y 与变量 x1，x9，x10 成反比，与 x2，x3，x4，x5，x6，x7，x8，x11 成正比，其中与 x2 的相关性最大，其相关系数为 0.3148，说明了对数工资与智商之间的相关性最强。

3.3. 统计作图

通过统计作图可以直观的发现数据特征，运用 MATLAB 画图，得到的统计结果如下，代码见附录 2。

1) 12 个变量的盒形图，见图 2。

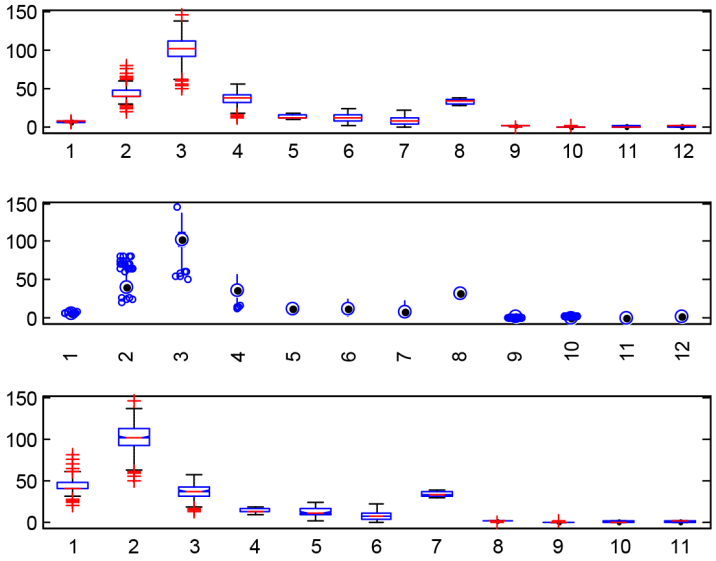

Figure 2. Box diagram of variables in wage2
图 2. WAGE2 中的变量的盒形图

上图盒形图更加直观的体现了数据的集中于分散程度，还提供了上四分位数，中位数，下四分位数的信息，能够跟家直观的反应统计特征信息。例如从上图可以观察到变量 y 与 x8，x9，x10，x11 的数据都是较为集中，并且还可以根据分位点，中位数来判断数据的分布情况。

2) 样本概率图形。

对 y 这个变量做概率图形，由前面的概率特征可知，平均值为 6.779 众数为 6.9078。我们想要求得对数工资 y 在区间[6, 7.5]概率，如下图 3 结果。

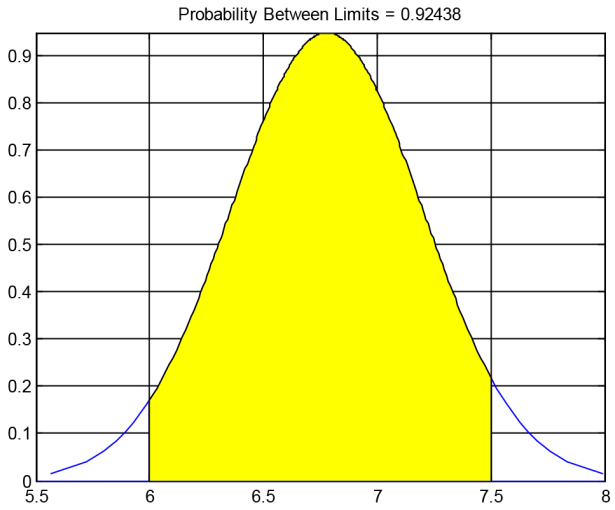

Figure 3. Sample probability graph of logarithmic wage
图 3. 对数工资的样本概率图形

可以直观的发现对数工资 y 在区间[6, 7.5]概率是 92.4%，这说明的大多数数据都集中在 6 到 7.5 这个区间，只有极小部分不在这个区间。也可以类似的去查看其他变量的数据分布所占比例。

4. AIC, BIC, CV 模型选择与分析

4.1. 问题回顾

本章用线性模型研究“对数薪水”和其他协变量之间的关系，考虑 12 个带有嵌套结构的备选模型

模型 1：$y_i = \beta_0 + \varepsilon_i$，$\varepsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right)$

模型 2：$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$，$\varepsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right)$

……

模型 12：$y_i = \beta_0 + \beta_1 x_i + \cdots + \beta_{11} x_{11} \varepsilon_i$，$\varepsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right)$

旨在用 AIC, BIC, CV 来从这 12 个模型中选择合适的模型，更好的解释对数薪水与其他协变量之间的关系，其中的变量说明见表 1。

4.2. 对 WAGE2 中的全部数据来进行模型选择

1) 本节目的：依托 WAGE2 中所有的 935 个数据，使用 CV，AIC 和 BIC，从上述 12 个备选模型里选择合适的模型。

2) 根据 2.2 节 AIC, BIC 模型的实现步骤先对全部数据进行模型选择，运行 MATLAB 代码，见附录 3，可以得到 AIC, BIC 选择模型的结果。

结果如下表 6。

Table 6. AIC values of 12 models
表 6. 12 个模型的 AIC 值

12 个模型的 AIC											
MODEL1	MODEL2	MODEL3	MODEL4	MODEL5	MODEL6	MODEL7	MODEL8	MODEL9	MODEL10	MODEL11	MODEL12
1039.3	1039.2	940.51	899.48	880.18	859.21	842.4	843.43	823.75	814.16	800.79	759.2700765

其中最小的 AIC 的值为 759.2700765，对应的模型为 12，故有 AIC 准则确定最优的模型应该为第 12 个。

Table 7. BIC values of 12 models
表 7. 12 个模型的 BIC 值

12 个模型的 BIC											
MODEL1	MODEL2	MODEL3	MODEL4	MODEL5	MODEL6	MODEL7	MODEL8	MODEL9	MODEL10	MODEL11	MODEL12
1049	1053.7	959.87	923.68	909.23	893.09	881.13	887	872.16	867.4	858.87	822.1971814

由表 7 知，其中最小的 BIC 的值为 822.1971814，对应的模型为 12，故由 BIC 准则确定最优的模型也是第 12 个。

3) 用 CV 来对全部数据处理，选出最优的模型。

用 935 个数据，每次抽出一行作为验证集，剩下的 934 个样本作为训练集，总共需要进行 935 次，运行 MATLAB 代码得到如下次结果，代码见附录 4。

将 CV 记作 12 个模型 935 次预测误差求和，则最小的 CV 值就对应最优的模型。

Table 8. CV values of 12 models
表 8. 12 个模型的 CV 值

12 个模型的 CV(935 次的预测误差求和)											
MODEL1	MODEL2	MODEL3	MODEL4	MODEL5	MODEL6	MODEL7	MODEL8	MODEL9	MODEL10	MODEL11	MODEL12
166.01	166.14	149.5	143.1	140.21	137.11	134.71	134.87	132.06	130.7	128.87	123.2714065

由上表 8 的结果知，最小的 CV 值为 123,2714065，对应的模型为第 12 个，故由 CV 的判别准则第 12 个模型是最优的。

综上：AIC, BIC, CV 这三个判别准则所选的模型都是第 12 个模型，所以第 12 个模型就是最优模型。

4) 用最小二乘法对第 12 个模型进行拟合，运行 MATLAB 代码，见附录 5，加结果如下：

第 12 个模型中 y 与 x 的散点图如下图 4。

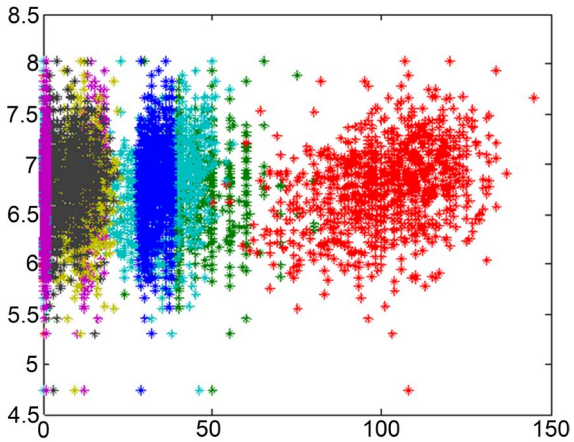

Figure 4. Scatter plot of the 12th model
图 4. 第 12 个模型的散点图

最小二乘估计的参数结果如下表 9。

Table 9. CV values of 12 models
表 9. 12 个模型的 CV 值

β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}	β_{11}
5.2797	−0.006	0.0033	0.0035	0.0492	0.0105	0.01	0.0054	0.195	−0.142	−0.081	0.1775

第 12 个模型拟合的方程如下

$$y = 5.2797 - 0.006x_1 + 0.0033x_2 + 0.0035x_3 + 0.0492x_4 + 0.0105x_5 + 0.01x_6 + 0.0054x_7 + 0.195x_8 - 0.142x_9 - 0.081x_{10} + 0.1775x_{11}$$

4.3. 经济学原理对 4.2 的结果进行解释

由 4.2 的结果知，AIC, BIC 与 CV 所选的模型都为第 12 个模型，且由最小二乘估计得到了第 12 个模型的拟合的方程

$$y = 5.2797 - 0.006x_1 + 0.0033x_2 + 0.0035x_3 + 0.0492x_4 + 0.0105x_5 + 0.01x_6 + 0.0054x_7 + 0.195x_8 - 0.142x_9 - 0.081x_{10} + 0.1775x_{11}$$

我们发现，

1) y 与变量 x1 的系数为负值，说明 lwage 与 hours 是成反比的关系，hours 每减少 1 个单位，lwage 增加 0.006 个单位，说明了现代社会对效率的要求越来越高，效率越高，报酬越多；

2) y 与变量 x2 的系数为正值，说明 lwage 与 IQ 是成正比关系，说明智商每增加一个单位，对数工资就增加 0.0033 个单位，同时也说明了智商高的人获得的工资报酬就越高；

3) y 与变量 x3 的系数为正值，说明 lwage 与 kww 是成正比关系，说明世界工作知识得分每增加一个单位，对数工资就增加 0.0035 个单位；

4) y 与变量 x4 的系数为正值，说明 lwage 与 educ 是成正比关系，说明教育每增加一个单位，对数工资就增加 0.0492 个单位；

5) y 与变量 x5 的系数为正值，说明 lwage 与 exper 是成正比关系，说明工作经验每增加一个单位，对数工资就增加 0.0105 个单位。

6) y 与变量 x6 的系数为正值，说明 lwage 与 enure 是成正比关系，说明与现任雇主共事年限每增加一个单位，对数工资就增加 0.01 个单位。

7) y 与变量 x7 的系数为正值，说明 lwage 与 age 是成正比关系，说明在一定的年龄范围内，年龄每增加一个单位，对数工资就增加 0.0054 个单位。因为在一定年龄范围内，年龄大的人相对来讲工作经验会多一点，知识储备会多一些，所以获得的工资报酬也会相对的高一些。

8) y 与变量 x8 的系数为正值，说明 lwage 与 married 是成正比关系，说明已婚每增加一个单位，对数工资就增加 0.195 个单位。

9) y 与变量 x9 的系数为负值，说明 lwage 与 black 是成反比关系，说明黑人获得的工资会更少一些，同时也说明了现代社会依然存在着种族歧视。

10) y 与变量 x10 的系数为负值，说明 lwage 与 south 是成反比关系，说明越靠近南边，工资越少，因为当前全球区域经济发展不平衡，例如南非，经济发展落后，所以工资报酬会更低一点。

11) y 与变量 x11 的系数为正值，说明 lwage 与 urban 是成正比关系，说明了生活在标准城市统计区的人工资报酬会更高一点。因为标准城市区域经济会相对发达一点，所以工资报酬会相对较高一点。

4.4. 将 WAGE2 的数据分为训练集与测试集，重新进行模型选择

在 4.2 节中，我们只对模型进行了一次选择，并不很具有说服力，所以在本小节，我们对模型重复进行 1000 次选择，采用随机抽样的方法将数据集随机地分为训练集(500 个样本)和测试集(435 个样本)，依托训练集，使用 CV，AIC 和 BIC，从 12 个备选模型里选择模型，并用选出的模型预测测试集里的测试样本，考察预测误差评价 CV，AIC 和 BIC 三种模型选择准则中哪种准则的表现最好。

运行 MATLAB 代码，得到结果，代码见附录 6。

1) 图 5 是 12 个模型与运行 1000 次得到的 AIC，维度为 12 行 1000 列。由于空间有限，仅显示部分结果。

577.7272	571.5149	565.2074	535.81	545.7028	546.3719	527.4607	541.3504
579.231	573.1337	564.5481	537.5007	547.396	548.127	528.3747	543.3397
525.7978	530.1103	508.0412	496.7615	503.7399	472.8512	488.3907	483.2091
511.2674	499.0427	491.4775	466.719	481.8172	452.3891	464.9608	462.7021
501.5725	489.2894	483.338	460.2843	468.2429	437.0215	459.3229	455.5478
483.9562	475.7496	475.5959	448.7772	462.9754	423.9022	447.6246	448.7309
479.1059	464.6041	457.0892	439.5193	455.9818	418.8158	442.5534	443.8776
480.7853	464.175	457.531	441.3358	457.9677	420.7245	443.9238	445.4655
475.8101	452.1039	448.7247	433.7384	445.7769	412.5968	433.2617	438.7975
467.1841	450.6955	441.7151	431.8083	440.5855	404.0935	428.037	433.5915
466.1974	446.9332	433.1279	421.9063	434.228	391.3456	410.8327	434.574
447.0748	423.9334	419.7492	406.936	419.5751	362.255	393.6094	411.7552

Figure 5. AIC values from 1000 runs
图 5. 运行 1000 次得到的 AIC 值

可以发现，1000 次运行结果中都是第 12 个模型的 AIC 的值最小，故选择第 12 个模型为最优模型。

下图是最小 AIC 对应模型的预测误差(即第 12 个模型的预测误差)，因为重复进行 1000 次，故有 1000 个预测误差，空间有限仅显示前 10 次的结果，见表 10。

Table 10. Prediction error of corresponding model for minimum AIC
表 10. 最小 AIC 对应模型的预测误差

最小 AIC 对应模型的预测误差(即第 12 个模型的预测误差)									
第 1 次	第 2 次	第 3 次	第 4 次	第 5 次	第 6 次	第 7 次	第 8 次	第 9 次	第 10 次
0.12283	0.12899	0.13141	0.13486	0.12908	0.14693	0.13845	0.1344	0.12962	0.11813

2) 下图是 12 个模型与运行 1000 次得到的 BIC，也仅展示部分结果，见图 6。

586.1564	579.9441	573.6367	544.2392	554.132	554.8011	535.8899	549.7797
591.8749	585.7776	577.192	550.1445	560.0399	560.7709	541.0185	555.9835
542.6562	546.9687	524.8996	513.62	520.5983	489.7096	505.2491	500.0675
532.3404	520.1157	512.5505	487.7921	502.8902	473.4621	486.0338	483.7751
526.8601	514.577	508.6257	485.572	493.5306	462.3091	484.6106	480.8355
513.4585	505.2519	505.0982	478.2795	492.4777	453.4045	477.1269	478.2332
512.8227	498.3209	490.806	473.2362	489.6987	452.5327	476.2703	477.5945
518.7167	502.1064	495.4625	479.2673	495.8992	458.6559	481.8553	483.397
517.9562	494.25	490.8708	475.8845	487.923	454.7429	475.4078	480.9436
513.5448	497.0562	488.0758	478.169	486.9461	450.4542	474.3976	479.9522
516.7727	497.5085	483.7032	472.4816	484.8033	441.9209	461.408	485.1493
501.8647	478.7233	474.5391	461.7259	474.365	417.0449	448.3993	466.5451

Figure 6. BIC values from 1000 runs
图 6. 运行 1000 次得到的 BIC 值

同样可以发现，1000 次运行结果中都是第 12 个模型的 BIC 的值最小，故选择第 12 个模型为最优模型。

下图是最小 BIC 对应模型的预测误差(即第 12 个模型的预测误差)，因为重复进行 1000 次，故有 1000 个预测误差，空间有限仅显示前 10 次的结果，见表 11。

Table 11. Prediction error of minimum BIC corresponding model
表 11. 最小 BIC 对应模型的预测误差

最小 BIC 对应模型的预测误差(即第 12 个模型的预测误差)									
第 1 次	第 2 次	第 3 次	第 4 次	第 5 次	第 6 次	第 7 次	第 8 次	第 9 次	第 10 次
0.122830573	0.128987541	0.131413267	0.134855128	0.129082835	0.146934749	0.138448017	0.134397175	0.129622311	0.118126556

3) 下图是 CV 运行 1000 次得到的预测误差，共 12 行 1000 列，篇幅限制，仅展示部分结果，见图 7。

0. 168846	0. 171537	0. 176908	0. 186094	0. 182877	0. 181891	0. 189098	0. 184025
0. 168309	0. 171006	0. 177811	0. 185575	0. 182352	0. 181379	0. 188662	0. 184208
0. 151918	0. 150422	0. 161589	0. 163537	0. 160483	0. 172521	0. 166265	0. 169566
0. 143254	0. 148286	0. 152862	0. 159919	0. 153612	0. 164939	0. 159571	0. 162177
0. 14014	0. 145082	0. 148169	0. 155493	0. 151899	0. 163173	0. 155107	0. 157558
0. 140416	0. 142937	0. 143538	0. 152199	0. 146831	0. 159571	0. 151945	0. 152326
0. 136364	0. 141592	0. 145038	0. 149568	0. 143383	0. 155801	0. 148082	0. 148805
0. 136193	0. 142493	0. 145215	0. 15012	0. 143289	0. 156146	0. 147887	0. 148566
0. 131866	0. 140103	0. 141715	0. 146051	0. 140804	0. 151982	0. 144968	0. 144455
0. 131954	0. 137841	0. 140863	0. 143857	0. 139431	0. 15149	0. 143561	0. 14305
0. 128871	0. 134885	0. 139515	0. 143136	0. 137037	0. 151439	0. 145662	0. 140629
0. 122831	0. 128988	0. 131413	0. 134855	0. 129083	0. 146935	0. 138448	0. 134397

Figure 7. Prediction error for 1000 runs of CV
图 7. 运行 1000 次 CV 的预测误差

下图是 1000 次中每次选中的模型，篇幅限制，仅展示部分结果，见表 12。

Table 12. Optimal model for each selection in 1000 times
表 12. 1000 次中每次选中的最优模型

1000 次中每次选中的最优模型									
第 1 次	第 2 次	第 3 次	第 4 次	第 5 次	第 6 次	第 7 次	第 8 次	第 9 次	第 10 次
7	11	4	6	7	12	12	1	3	12

下图是 1000 次中每个模型被选中的概率。

Table 13. Proportion of each model selected in 1000 times
表 13. 1000 次中每个模型被选中的比例

1000 次中每个模型被选中的比例											
MODEL1	MODEL2	MODEL3	MODEL4	MODEL5	MODEL6	MODEL7	MODEL8	MODEL9	MODEL10	MODEL11	MODEL12
0.114	0.093	0.111	0.088	0.052	0.062	0.043	0.038	0.036	0.042	0.093	0.228

由表 13，可以发现第 12 个模型被选中的比例最高，为 22.8%，其他模型被选中的比例明显都比较小，所以 CV 模型选中的依然是第 12 个模型。

综上：AIC, BIC, CV 三个模型选出的全部为第 12 个模型最优，故认为第 12 个模型为最优的模型。

5. 研究结论

依托前面的知识准备，通过第三章的描述性统计分析，可以清楚直观的看到数据的分布特征及其变量的相关关系。然后在第四章用 AIC, BIC, CV 准则依托全部数据来进行模型选择，最后 3 个准则选出的最优模型均为第 12 个模型，故此，对第 12 个模型用最小二乘法进行了拟合，求得参数，得到方程，进而用经济学原理对求得的第 12 个模型进行解释。但是考虑到一次实验结果并不是十分具有说服力，因此，又将实验重复 1000 次，再次看模型选择的结果，结果发现 AIC, BIC 在这 1000 次实验中一致选择的是第 12 个模型，而 CV 选择的模型依概率算是占比最高的，为 22.8%，说明在 1000 次实验中，选择第 12 个模型的次数占 228 次。而选择其他模型的比例最高占据 11.4%，故此一致认为第 12 个模型是最优的。

参考文献

[1] 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

[2] Flynn, C.J., Hurvich, C.M. and Simonoff, J.S. (2013) Efficiency for Regularization Parameter Selection in Penalized Likelihood Estimation of Misspecified Models. *Journal of the American Statistical Association*, **108**, 1031-1043. https://doi.org/10.1080/01621459.2013.801775

[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] https://blog.csdn.net/qq_30142403/article/details/80457050

附录

附录 1：统计特征

```
load('C:\Users\Administrator\Desktop\wage2.mat')
n = length(y); Data = [y,x];
M1 = mean(Data); M2 = median(Data); M3 = mode(Data);
V = var(Data); C = cov(Data); Corr = corrcoef(Data);
```

附录 2：统计作图

```
%绘制盒形图
subplot(3,1,1); boxplot(Data);subplot(3,1,2);
boxplot(Data,' plotstyle','compact');
subplot(3,1,3); boxplot(x,'notch','on');
%绘制对数工资的样本概率图形
p1 = capaplot(y,[6 7.5]); grid on; axis tight;
p2 = capaplot(y,[6.9 8.8]); grid on; axis tight;
```

附录 3：AIC, BIC 依托全数据进行模型选择

```
%AIC, BIC 做模型选择
%用全部的数据 935 个数据进行 AIC, BIC 模型选择
%[AIC,BIC,c2AIC,c2BIC] = Fun1AB(x,y)
%y 表示的是 lwage
%x 表示的是 hours,IQ,KWW,educ,exper,tenure,age,married,black,south,urban 这 11 个变量
function [AIC,BIC,c2AIC,c2BIC] = Fun1AB(x,y)
%__________________( AIC,BIC 数据准备 )_____________________________________
n = length(y); Data = [y,ones(n,1),x];
X= Data(:,2:13);
X1= X(:,1);X2 = X(:,1:2); X3= X(:,1:3); X4= X(:,1:4); X5= X(:,1:5); X6 = X(:,1:6); X7= X(:,1:7);
X8= X(:,1:8); X9= X(:,1:9);X10= X(:,1:10); X11= X(:,1:11); X12= X(:,1:12);
%____________________模型 1 的 AIC 与 BIC 构建----------------------------------
hb1 = [(X1'*X1)\(X1'*y);zeros(11,1)];%第一个模型的参数 beta，为了保证与后面 11 个模型的参数维度一致，所以补 11 行 1 列的 0
hsig21 = mean((y − X*hb1).^2);%模型 1 的 sig2 的极大似然估计
L1 = −n*log(2*pi)/2 − n*log(hsig21)/2 − n/2;%模型 1 的极大似然函数
AIC1 = −2*L1 + 2*2; %模型 1 的赤池信息准则，存储在 AIC 的第 1 行第 k 列
BIC1 = −2*L1 + log(n)*2; %模型 1 的贝叶斯信息准则，存储在 BIC 的第 1 行第 k 列
%____________________模型 2 的 AIC 与 BIC 构建 _____________________________
hb2 = [(X2'*X2)\(X2'*y);zeros(10,1)];%第二个模型的参数 beta,为了保证与 beta3 的参数维度一致，所以补 10 行 1 列的 0
hsig22 = mean((y − X*hb2).^2);%模型 2 的 sig2 的极大似然估计
L2 = −n*log(2*pi)/2 − n*log(hsig22)/2 − n/2;%模型 2 的极大似然函数
AIC2 = −2*L2 + 2*3; %模型 2 的赤赤池信息准则，存储在 AIC 的第 2 行第 k 列
```

```
BIC2 = −2*L2 + log(n)*3;%模型 2 的贝叶斯信息准则,存储在 BIC 的第 2 行第 k 列
%_____ 模型 3 的 AIC 与 BIC 构建 _____
hb3 = [(X3'*X3)\(X3'*y);zeros(9,1)];%第三个模型的参数 beta
hsig23 = mean((y − X*hb3).^2);%模型 3 的 sig2 的极大似然估计
L3 = −n*log(2*pi)/2 − n*log(hsig23)/2 - n/2;%模型 3 的极大似然函数
AIC3 = −2*L3 + 2*4; %模型 3 的赤池信息准则,存储在 AIC 的第 3 行第 k 列
BIC3 = −2*L3 + log(n)*4;%模型 3 的贝叶斯信息准则，存储在 BIC 的第 3 行第 k 列
%_____ 模型 4 的 AIC 与 BIC 构建 _____
hb4= [(X4'*X4)\(X4'*y);zeros(8,1)]; hsig24 = mean((y − X*hb4).^2);
L4 = −n*log(2*pi)/2 − n*log(hsig24)/2 − n/2;
AIC4 = −2*L4 + 2*5; BIC4= -−2*L4 + log(n)*5;
%_____模型 5 的 AIC 与 BIC 构建 _____
hb5= [(X5'*X5)\(X5'*y);zeros(7,1)];hsig25 = mean((y - X*hb5).^2);
L5= −n*log(2*pi)/2 − n*log(hsig25)/2 −n/2;
AIC5 = −2*L5 + 2*6; BIC5 = −2*L5 + log(n)*6;
%_____模型 6 的 AIC 与 BIC 构建 _____
hb6 = [(X6'*X6)\(X6'*y);zeros(6,1)];hsig26 = mean((y − X*hb6).^2);
L6 = −n*log(2*pi)/2 − n*log(hsig26)/2 − n/2;
AIC6 = −2*L6 + 2*7;BIC6 = −2*L6 + log(n)*7;
%_____模型 7 的 AIC 与 BIC 构建 _____
hb7 = [(X7'*X7)\(X7'*y); zeros(5,1)]; hsig27 = mean((y − X*hb7).^2);
L7= −n*log(2*pi)/2 - n*log(hsig27)/2 - n/2;
AIC7= -2*L7+ 2*8;BIC7 = -2*L7 + log(n)*8;
%_____ 模型 8 的 AIC 与 BIC 构建 _____
hb8= [(X8'*X8)\(X8'*y);zeros(4,1)];hsig28 = mean((y - X*hb8).^2);
L8 = −n*log(2*pi)/2 − n*log(hsig28)/2 −n/2;
AIC8 = −2*L8+ 2*9; BIC8 = −2*L8 + log(n)*9;
%_____ 模型 9 的 AIC 与 BIC 构建 _____
hb9 = [(X9'*X9)\(X9'*y);zeros(3,1)];hsig29 = mean((y - X*hb9).^2);
L9 = −n*log(2*pi)/2 − n*log(hsig29)/2 − n/2;
AIC9 = −2*L9 + 2*10; BIC9 = −2*L9 + log(n)*10;
%_____模型 10 的 AIC 与 BIC 构建 _____
hb10 = [(X10'*X10)\(X10'*y); zeros(2,1)];hsig210 = mean((y − X*hb10).^2);
L10 = −n*log(2*pi)/2 − n*log(hsig210)/2 − n/2;
AIC10    = -2*L10+ 2*11;BIC10     = -2*L10 + log(n)*11;
%_____ 模型 11 的 AIC 与 BIC 构建 _____
hb11 = [(X11'*X11)\(X11'*y);0];hsig211 = mean((y - X*hb11).^2);
L11 = −n*log(2*pi)/2 − n*log(hsig211)/2 − n/2;
AIC11 = −2*L11 + 2*12; BIC11 = −2*L11 + log(n)*12;
%_____ 模型 12 的 AIC 与 BIC 构建 _____
```

```
hb12 = [(X12'*X12)\(X12'*y)]; hsig212 = mean((y − X*hb12).^2);
L12 = −n*log(2*pi)/2 − n*log(hsig212)/2 − n/2;
AIC12 = −2*L12+ 2*13; BIC12 = −2*L12 + log(n)*13;
AIC = [AIC1,AIC2,AIC3,AIC4,AIC5,AIC6,AIC7,AIC8,AIC9,AIC10,AIC11,AIC12]
BIC = [BIC1,BIC2,BIC3,BIC4,BIC5,BIC6,BIC7,BIC8,BIC9,BIC10,BIC11,BIC12]
% _____选出 12 个模型中的最小的 AIC 与 BIC _____
[c1AIC,c2AIC] = min(AIC);%c1ATC 代表最小的 AIC,c2AIC 代表最小的 AIC 所在的位置
[c1BIC,c2BIC] = min(BIC);%c1BTC 代表最小的 BIC,c2BIC 代表最小的 BIC 所在的位置
```

附录 4：用 CV 依托全数据进行模型选择

```
%=======================CV 做模型选择=====================================
%输入数据
%load('C:\Users\Administrator\Desktop\wage2.mat')
%[CV,c1CV,c2CV] = Fun1CV(x,y)
function [CV,c1CV,c2CV] = Fun1CV(x,y)
n = length(y);%总的样本量
Data = [y,ones(n,1),x];%构造的数据矩阵
for i = 1:n
%_____ CV 数据准备 _____
yTrain = Data(:,1);yTrain(i,:) = [];yTest = Data(i,1);
xTrain = Data(:,2:13);xTrain(i,:) = []; xTest = Data(i,2:13);
%_____12 个 CV 模型的变量数据_____
X1 = xTrain(:,1); X2 = xTrain(:,1:2); X3 = xTrain(:,1:3); X4 = xTrain(:,1:4); X5 = xTrain(:,1:5);
X6 = xTrain(:,1:6); X7 = xTrain(:,1:7); X8 = xTrain(:,1:8); X9 =xTrain(:,1:9); X10 = xTrain(:,1:10);
X11 = xTrain(:,1:11);X12 = xTrain(:,1:12);
%_____ 12 个 CV 模型的预测误差_____
pe1(i)=(yTest-xTest(:,1)*((X1'*X1)\(X1'*yTrain)))^2;
pe2(i)=(yTest-xTest(:,1:2)*((X2'*X2)\(X2'*yTrain)))^2;
pe3(i) = (yTest - xTest(:,1:3)*((X3'*X3)\(X3'*yTrain)))^2;
pe4(i) = (yTest - xTest(:,1:4)*((X4'*X4)\(X4'*yTrain)))^2;
pe5(i) = (yTest - xTest(:,1:5)*((X5'*X5)\(X5'*yTrain)))^2;
pe6(i) = (yTest - xTest(:,1:6)*((X6'*X6)\(X6'*yTrain)))^2;
pe7(i) = (yTest - xTest(:,1:7)*((X7'*X7)\(X7'*yTrain)))^2;
pe8(i) = (yTest - xTest(:,1:8)*((X8'*X8)\(X8'*yTrain)))^2;
pe9(i) = (yTest - xTest(:,1:9)*((X9'*X9)\(X9'*yTrain)))^2;
pe10(i) = (yTest - xTest(:,1:10)*((X10'*X10)\(X10'*yTrain)))^2;
pe11(i) = (yTest - xTest(:,1:11)*((X11'*X11)\(X11'*yTrain)))^2;
pe12(i) = (yTest - xTest(:,1:12)*((X12'*X12)\(X12'*yTrain)))^2;
end
%_____CV 预测误差求和_____
```

```
CV=[sum(pe1);sum(pe2);sum(pe3);sum(pe4);sum(pe5);sum(pe6);sum(pe7);sum(pe8);sum(pe9);sum(pe10);sum(pe11);sum(pe12)];
[c1CV,c2CV] = min(CV);%c1CV 代表最小的 CV,c2CV 代表最小的 CV 所在的位置
```

附录 5：最小二乘法拟合第 12 个模型

```
%用最小二乘法对的 12 个模型做拟和
load('C:\Users\Administrator\Desktop\wage2.mat')
plot(X12,y,'*');n = length(y); X= Data(:,2:13);Data = [y,ones(n,1),x];
X12 = X(:,1:12);hb12= [(X12'*X12)\(X12'*y)];
```

附录 6：重复 1000 次，用 AIC,BIC,CV 来进行模型选择

```
%=================AIC,BIC 与 CV 做模型选择=====================================
%[peAIC,peBIC,AIC,BIC,c2AIC,c2BIC,c2CV,CV,CV_P] = FunModelABC(x,y)
%y 表示的是 lwage
%x 表示的是 hours,IQ,KWW,educ,exper,tenure,age,married,black,south,urban 这 11 个变量
function [peAIC,peBIC,AIC,BIC,c2AIC,c2BIC,c2CV,CV,CV_P] = FunModelABC(x,y)
randn('seed',0541);%设置随机数种子
n = length(y);%总的样本量
nTest = 435; %选择 435 个样本做测试集
nTrain = n − nTest;%用剩余的样本做训练集
n1   = length(nTrain); %为了保证与 AIC, BIC 模型的训练集维度一致,CV 模型的维度 n1 = 935 − 435 = 500
CV = zeros(12,1000);%用来存放 CV 预测误差的求和
for k = 1:1000
%_________________( AIC,BIC 数据准备 )_____________________________
Data = [y,ones(n,1),x];%构造的数据矩阵
[DataTrain, DataTest] = smplwor(Data,nTrain);%从 Data 中无放回的抽样，抽出的样本为 DataTrain 作为训练集，输出的 DataTest 是抽样剩下的样本作为测试集
yTrain = DataTrain(:,1);%训练集中的因变量 y 为训练集的第 1 列
XTrain = DataTrain(:,2:13);%训练集中的自变量 x 为训练集的第 2 到 14 列
yTest   = DataTest(:,1);%测试集中的因变量 y 为训练集的第 1 列
XTest   = DataTest(:,2:13);%测试集中的自变量 x 为训练集的第 2 到 14 列
X1= XTrain(:,1); X2= XTrain(:,1:2); X3= XTrain(:,1:3); X4= XTrain(:,1:4); X5= XTrain(:,1:5);
X6= XTrain(:,1:6);X7 = XTrain(:,1:7);X8= XTrain(:,1:8);X9= XTrain(:,1:9);X10= XTrain(:,1:10);
X11= XTrain(:,1:11);X12= XTrain(:,1:12);
%_____________________ 模型 1 的 AIC 与 BIC 构建 _____________________________
hb1 = [(X1'*X1)\(X1'*yTrain);zeros(11,1)];
hsig21 = mean((yTrain - XTrain*hb1).^2);%模型 1 的 sig2 的极大似然估计
L1 = −nTrain*log(2*pi)/2 - nTrain*log(hsig21)/2 - nTrain/2;%模型 1 的极大似然函数
AIC(1,k) = −2*L1 + 2*2;%模型 1 的赤赤池信息准则，存储在 AIC 的第 1 行第 k 列
BIC(1,k) = −2*L1 + log(nTrain)*2;
%_____________________模型 2 的 AIC 与 BIC 构建 _____________________________
```

```
hb2 = [(X2'*X2)\(X2'*yTrain);zeros(10,1)];
hsig22 = mean((yTrain − XTrain*hb2).^2);%模型 2 的 sig2 的极大似然估计
L2 = −nTrain*log(2*pi)/2 − nTrain*log(hsig22)/2 − nTrain/2; %模型 2 的极大似然函数
AIC(2,k) = −2*L2  + 2*3;%模型 2 的赤赤池信息准则，存储在 AIC 的第 2 行第 k 列
BIC(2,k) = −2*L2 + log(nTrain)*3;
%_____ 模型 3 的 AIC 与 BIC 构建 _____
hb3 = [(X3'*X3)\(X3'*yTrain);zeros(9,1)];hsig23 = mean((yTrain − XTrain*hb3).^2);
L3 = −nTrain*log(2*pi)/2 − nTrain*log(hsig23)/2 − nTrain/2;%模型 3 的极大似然函数
AIC(3,k) = −2*L3 + 2*4; BIC(3,k) = −2*L3 + log(nTrain)*4;
%_____ 模型 4 的 AIC 与 BIC 构建 _____
hb4 = [(X4'*X4)\(X4'*yTrain);zeros(8,1)];hsig24 = mean((yTrain − XTrain*hb4).^2);
L4 = −nTrain*log(2*pi)/2 − nTrain*log(hsig24)/2 − nTrain/2;
AIC(4,k) = −2*L4 + 2*5;BIC(4,k) = −2*L4 + log(nTrain)*5;
%_____ 模型 5 的 AIC 与 BIC 构建 _____
hb5 = [(X5'*X5)\(X5'*yTrain);zeros(7,1)]; hsig25 = mean((yTrain − XTrain*hb5).^2);
L5 = −nTrain*log(2*pi)/2 − nTrain*log(hsig25)/2 − nTrain/2;
AIC(5,k) = −2*L5 + 2*6; BIC(5,k) = −2*L5 + log(nTrain)*6;
%_____ 模型 6 的 AIC 与 BIC 构建 _____
hb6 = [(X6'*X6)\(X6'*yTrain);zeros(6,1)];hsig26 = mean((yTrain - XTrain*hb6).^2);
L6 = −nTrain*log(2*pi)/2 − nTrain*log(hsig26)/2 − nTrain/2;
AIC(6,k) = −2*L6 + 2*7;BIC(6,k) = −2*L6 + log(nTrain)*7;
%_____ 模型 7 的 AIC 与 BIC 构建 _____
hb7 = [(X7'*X7)\(X7'*yTrain);zeros(5,1)];hsig27 = mean((yTrain − XTrain*hb7).^2);
L7 = −nTrain*log(2*pi)/2 − nTrain*log(hsig27)/2 − nTrain/2;
AIC(7,k) = −2*L7 + 2*8;BIC(7,k) = −2*L7 + log(nTrain)*8;
%_____ 模型 8 的 AIC 与 BIC 构建 _____
hb8 = [(X8'*X8)\(X8'*yTrain);zeros(4,1)];hsig28 = mean((yTrain − XTrain*hb8).^2);
L8 = −nTrain*log(2*pi)/2 − nTrain*log(hsig28)/2 − nTrain/2;
AIC(8,k) = −2*L8 + 2*9;BIC(8,k) = −2*L8 + log(nTrain)*9;
%_____ 模型 9 的 AIC 与 BIC 构建 _____
hb9 = [(X9'*X9)\(X9'*yTrain);zeros(3,1)];hsig29 = mean((yTrain - XTrain*hb9).^2);
L9 = −nTrain*log(2*pi)/2 − nTrain*log(hsig29)/2 − nTrain/2;
AIC(9,k) = −2*L9 + 2*10; BIC(9,k) = −2*L9 + log(nTrain)*10;
%_____ 模型 10 的 AIC 与 BIC 构建 _____
hb10 = [(X10'*X10)\(X10'*yTrain);zeros(2,1)];hsig210 = mean((yTrain - XTrain*hb10).^2);
L10= −nTrain*log(2*pi)/2 − nTrain*log(hsig210)/2 − nTrain/2;
AIC(10,k) = −2*L10 + 2*11;BIC(10,k) = −2*L10 + log(nTrain)*11;
%_____ 模型 11 的 AIC 与 BIC 构建 _____
hb11 = [(X11'*X11)\(X11'*yTrain);0];hsig211 = mean((yTrain − XTrain*hb11).^2);
L11 = −nTrain*log(2*pi)/2 − nTrain*log(hsig211)/2 − nTrain/2;
```

```
AIC(11,k) = −2*L11 + 2*12;BIC(11,k) = −2*L11 + log(nTrain)*12;
%_____ 模型 12 的 AIC 与 BIC 构建 _____
hb12 = [(X12'*X12)\(X12'*yTrain)]; hsig212 = mean((yTrain − XTrain*hb12).^2);
L12 = −nTrain*log(2*pi)/2 − nTrain*log(hsig212)/2 − nTrain/2;
AIC(12,k) = −2*L12 + 2*13;BIC(12,k) = −2*L12 + log(nTrain)*13;
%_____ 关于 AIC,BIC 的 12 个模型的预测误差 _____
pe(1,k) = mean((yTest − XTest*hb1).^2); pe(2,k) = mean((yTest − XTest*hb2).^2);
pe(3,k) = mean((yTest − XTest*hb3).^2);pe(4,k) = mean((yTest − XTest*hb4).^2);
pe(5,k) = mean((yTest − XTest*hb5).^2);pe(6,k) = mean((yTest − XTest*hb6).^2);
pe(7,k) = mean((yTest − XTest*hb7).^2); pe(8,k) = mean((yTest − XTest*hb8).^2);
pe(9,k) = mean((yTest − XTest*hb9).^2); pe(10,k) = mean((yTest − XTest*hb10).^2);
pe(11,k) = mean((yTest − XTest*hb11).^2); pe(12,k) = mean((yTest − XTest*hb12).^2);
%=========================CV 做模型选择=====================================
%选用随机抽样的 500 个样本对 CV 进行数据处理，维度 n1 = length(nTrain) = 500
%选用的数据是前面 AIC,BIC 模型中随机抽样的 500 个训练集样本 DataTrain
%用 DataTrain 的数据集每次删除一个样本，用这个样本作为测试集，用剩下的样本做训练集，循环 500
次
%考察预测误差，循环过程 1000 次
for i = 1:n1
%_____ CV 数据准备 _____
y1Train = DataTrain(:,1);%取出 DataTrain 的第 1 列作为 y1Train
y1Train(i,:) = []; %删除第 y1Train 中的第 i 行
y1Test = DataTrain(i,1);%用删除的 y1Train 中的那行作为测试，记作 y1Test
x1Train = DataTrain(:,2:13);%x1Train 为中的第 2 列到 13 列
x1Train(i,:) = [];%删除 x1Train 中的第 i 行
x1Test = DataTrain(i,2:13);%用删除 x1Train 中的那行作为测试，记作 x1Test
%_____12 个 CV 模型的变量数据_____
X11 = x1Train(:,1); X21 = x1Train(:,1:2); X31=x1Train(:,1:3); X41 = x1Train(:,1:4);
X51 = x1Train(:,1:5); X61 = x1Train(:,1:6); X71 = x1Train(:,1:7); X81 =x1Train(:,1:8);
X91 = x1Train(:,1:9); X101 = x1Train(:,1:10); X111 = x1Train(:,1:11); X121 = x1Train(:,1:12);
%_____ 12 个 CV 模型的预测误差_____
pe11(i) = (y1Test − x1Test(:,1)*((X11'*X11)\(X11'*y1Train)))^2;
pe21(i) = (y1Test − x1Test(:,1:2)*((X21'*X21)\(X21'*y1Train)))^2;
pe31(i) = (y1Test − x1Test(:,1:3)*((X31'*X31)\(X31'*y1Train)))^2;
pe41(i) = (y1Test − x1Test(:,1:4)*((X41'*X41)\(X41'*y1Train)))^2;
pe51(i) = (y1Test − x1Test(:,1:5)*((X51'*X51)\(X51'*y1Train)))^2;
pe61(i) = (y1Test − x1Test(:,1:6)*((X61'*X61)\(X61'*y1Train)))^2;
pe71(i) = (y1Test − x1Test(:,1:7)*((X71'*X71)\(X71'*y1Train)))^2;
pe81(i) = (y1Test − x1Test(:,1:8)*((X81'*X81)\(X81'*y1Train)))^2;
```

```
pe91(i) = (y1Test − x1Test(:,1:9)*((X91'*X91)\(X91'*y1Train)))^2;
pe101(i) = (y1Test − x1Test(:,1:10)*((X101'*X101)\(X101'*y1Train)))^2;
pe111(i) = (y1Test − x1Test(:,1:11)*((X111'*X111)\(X111'*y1Train)))^2;
pe121(i) = (y1Test − x1Test(:,1:12)*((X121'*X121)\(X121'*y1Train)))^2;
end
%_____CV 预测误差求和_____
Loss =
[sum(pe11);sum(pe21);sum(pe31);sum(pe41);sum(pe51);sum(pe61);sum(pe71);sum(pe81);sum(pe91);sum(pe1
01);sum(pe111);sum(pe121)];
CV(:,k) = Loss;
end
% _____选出 12 个模型中的最小的 AIC 与 BIC _____
[c1AIC,c2AIC] = min(AIC); %c1ATC 代表最小的 AIC,c2AIC 代表最小的 AIC 所在的位置
[c1BIC,c2BIC] = min(BIC); %c1BTC 代表最小的 BIC,c2BIC 代表最小的 BIC 所在的位置
[c1CV,c2CV] = min(CV); %c1CV 代表最小的 CV,c2CV 代表最小的 CV 所在的位置
%_____计算 CV 的 12 个模型 1000 次中每个模型被选中的比例_____
for h = 1:12
CV_P(h) = mean(c2CV==h);
end
%_____最小 AIC,BIC,CV 对应位置的预测误差 _____
for k = 1:1000
peAIC(k) = pe(c2AIC(k),k);peBIC(k) = pe(c2BIC(k),k);peCV(k) = pe(c2CV(k),k);
end
%========== 从 x 中无放回的抽取 k 个样本 ==========
function [x, Xre, idk ] = smplwor(X,K)
[N, p] = size(X); Nk= N; x = zeros(K,p); idk= zeros(K,1);
for k = 1:K
idk(k) = unidrnd(Nk);%从 1 到 Nk 的正整数中产生 1 个均匀随机整数，作为角标
x(k,:) = X(idk(k),:);%取出原始数据 X 对应角标的那行存储在 x 的第 k 行中
X(idk(k),:) = [];%删除原始数据 X 所抽走的那行
Nk = Nk - 1;%无放回抽样，每次抽一行后都减掉 1
end
Xre = X;%把已经删除了抽样行的矩阵记作 Xre，即抽样剩下的样本矩阵的样本矩阵
```