线性回归中的模型选择

任宣霏

2022 年 10 月 2 日

目录

1 预备知识

1. 韦来生《数理统计》

2. 王松桂等《线性统计模型》 1-6章

3. 李东风老师的 R 语言讲义

2 问题的提出

研究线性拟合问题。已知 y 可能与协变量 $S \triangleq \{x_i\}_{i=1}^n$ 有关,采用线性拟合模型：

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_{p-1} + e \tag{2.1}$$

其中 e 是噪声项，可以假定其均值为 0 ，方差为 σ^2 .

假如我们选取所有的 x_i 来用最小二乘法拟合，可以得到最好的拟合效果，这里的拟合优度更好是直观的。但事实上， y 可能与某些 x_i 完全不相关，选取更多的自变量个数只能使得观测值 y 与估计值 $\hat{y} = \hat{\beta_0} + \hat{\beta_1} x_1 + \ldots + \hat{\beta_{p-1}} x_n$ 之间的差值 $\sum_{k=1}^n \left(\hat{y_k} - y_k\right)^2$, i.e.残差平方和 RSS更小（假如观测到 n 组数据），但是不能很好地预测真实情况，变量选取过多也自然增加了复杂度。因此我们要选取最佳自变量组，防止过拟合，同时保证拟合效果好。目前有两种最常用方法：AIC，BIC，均处理模型选择问题，对于过多参数给予惩罚项。

2.1 Akaike information criterion

AIC 处理过拟合和欠拟合风险

Definition 2.1 (AIC). *假设模型误差服从独立正态分布，以 k 表示参数数量，L 表示似然函数的最大值，则 AIC 定义为*

$$AIC = n \ln\left(RSS_q\right) + 2q \ln n$$

增加自由参数的数目提高了拟合的优良性， AIC 鼓励数据拟合的优良性但是尽量避免出现过度拟合（Overfitting）的情况。

所以优先考虑的模型应是 AIC 值最小的那一个。赤池信息量准则的方法是寻找可以最好地解释数据但包含最少自由参数的模型。

Remarks 1. *AIC* 处理的是对于一个未知过程 f, 对比两种拟合 g_1,g_2, 且只能给出相对的好坏。如果两个拟合都很差， *AIC* 方法无法给出警告。*AIC* 选择可能是最好的模型，使得信息丢失最少。可以直接用 R 软件来计算 *AIC*.

2.2 Bayesian information criterion

Definition 2.2 (BIC). 参数含义与 *AIC* 相同。

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

Definition 2.3 (BIC'). 用残差平方和定义

$$BIC = n\ln(\frac{SSR}{n}) + 2q\ln n$$

Remark 2.4. *BIC* 中，更复杂的模型受到更多的惩罚。因此更倾向于选协变量更少的模型。

3 论文简介

上述模型中可能会依赖协变量与残差的正态假设，论文[1]提出了一种不严格依赖正态假设且不作协变量稀疏性假设的估计方法，即所谓“估计方程的方法”。

Definition 3.1 (因变量被解释的比例). 定义为

$$r^2 = \frac{Var\left\{E(Y_i|X_{i1}\dots X_{ip})\right\}}{Var\left(Y_i\right)} \tag{3.1}$$

在线性模型

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \epsilon_i \tag{3.2}$$

中，若 $E\left(\epsilon_i\right) = 0$,$E\left(\epsilon_i^2\right) = \sigma_\epsilon^2$,且互不相关，即满足 Gauss-Markov 假设，列向量 $\beta = (\beta_1 \cdots, \beta_p)^t$, X的协方差阵 $Var\left(X_{i1}, \cdots, X_{ip}\right) = \Sigma$. 则解释可以化为

$$r^2 = \frac{\beta^t\Sigma\beta}{\beta^t\Sigma\beta + \sigma_\epsilon^2} \tag{3.3}$$

Remark 3.2. 这里是我个人对于这个“解释”的理解： r^2 可以看作 $\frac{\beta^t\Sigma\beta}{\sigma_\epsilon^2}$ 的单调递增函数，分子表示由于自变量 X 的波动引起的 Y 的波动程度，分母表示随机扰动的影响。如果拟合的较好，则 Y 的振动主要由协变量 X_i 的扰动引起，因此拟合得好等价于这个“ Y 被 X 解释的程度”，即 r^2 更大!

Remark 3.3. 在这里可以和最初研究的模型选择问题联系起来，给一组样本，通过对 r^2 的估计来说明选择特定模型，拟合的好坏。希望能选取 r^2 最大的模型。

Remark 3.4. 上述只是我个人参考文章的形象化理解，是否有一个严格的数学表达？

Definition 3.5 (unbiased estimating scores). 文章里定义了一个*estimating scores,(应该就是作者定义的一个名字)* 略去中间过程，最后可以表达为

$$\frac{1}{\sigma_Y^2}\tilde{Y}\tilde{Y}^t-\left(I-\mathbf{1}_n\mathbf{1}_{\boldsymbol{n}}^t/n\right)-\left\{M-\left(I-\mathbf{1}_n\mathbf{1}_n^t/n\right)\right\}r^2 \tag{3.4}$$

其中满足：

$$\begin{aligned}\tilde{Y}&=Y-\mathbf{1}_n\bar{Y}\\ M&=\frac{1}{p}\left(Z-\mathbf{1}_n\bar{Z}^t\right)\left(Z-\mathbf{1}_n\bar{Z}^t\right)^t\\ Z&=\left(Z_{ij}\right)_{n\times p}\\ \left(Z_{i1},\cdots,Z_{ip}\right)^t&=\Sigma^{-1/2}\left(X_{i1},\cdots,X_{ip}\right)^t\end{aligned} \tag{3.5}$$

引入一个加权矩阵 W ,可以取为

$$W=(I+\lambda M)^{-1}(M-I)(I+\lambda M)^{-1}$$

用 $\hat{\sigma}_Y^2$ 估计 σ_Y^2 , i.e.

$$\begin{aligned}RSS&=y^t\left(I-X\left(X^tX\right)^{-1}X^t\right)y\\ \hat{\sigma}^2&=\frac{RSS}{n-p}\end{aligned} \tag{3.6}$$

然后用矩阵 W 给上面定义的“估计分数”加权，得到

$$tr\left(W\left[\frac{1}{\hat{\sigma}_Y^2}\tilde{Y}\tilde{Y}^t-\left(I-\mathbf{1}_n\mathbf{1}_n^t/n\right)-\left\{M-\left(I-\mathbf{1}_n\mathbf{1}_n^t/n\right)\right\}r^2\right]\right)=0$$

Remark 3.6. n 很大时，矩阵不一定收敛，但它的秩容易收敛，这里只是一个想法，后面会对简单的情况给出具体证明。比如 Σ 是固定给出的，或者直接假定 X_i 不相关，不用引入 Z_i .

从而得到 r^2 的估计：

$$\hat{r}^2 = \frac{tr\left[W\left\{\frac{1}{\hat{\sigma}_Y^2}\tilde{Y}\tilde{Y}^t - (I - \mathbf{1}_n\mathbf{1}_n^t/n)\right\}\right]}{tr\left[W\left\{M - (I - \mathbf{1}_n\mathbf{1}_n^t/n)\right\}\right]} \tag{3.7}$$

这里方差 $\hat{\sigma}_Y^2$ 由样本的经验估计 (3.6) 给出。

Remark 3.7. 这里的 M 含有 Z ,而 Z 由(3.5)给出，那么用样本估计 r^2 时，Z 中的 Σ 用样本的协方差矩阵估计吗？事实上，真实研究中可以找一些不含 Y 的数据专门来估计协方差，下面由于我未能找到漂亮的数据来练习，暂且使用样本 X 的协方差阵。这里可能会有偏差。

4 R数据模拟

这一部分我们用上面的“解释量”尝试做变量选择问题，并与传统的 AIC 方法作比较。由于我的R语言还很不熟练，第一次正式写代码做问题，一步一步慢慢来。后来计算表明，(4.1) 和 (4.2) 中的数据可能不符合原假定，得到的 $\hat{r^2}$ 估计量甚至不在 $0 \sim 1$ 中。在后面，要先实用强假设的模型计算一些符合假定的模拟数据。这些结果错误的数据暂且先放在这里，看后续会不会有什么用处。

4.1 餐馆数据

数据来自于李东风老师讲义 （回归自变量筛选） （例33.4）里附带的“餐馆营业额”可能受到“居民数”“人均餐费”“月收入”“参观数”“距离”五个自变量影响。数据大概如图(1)所示，截图不容易截全，稍后我会选择完整的25组数据做计算。

	营业额	居民数	人均餐费	月收入	餐馆数	距离
1	53.2	163.0	168.6	6004	5	6.5
2	18.5	14.5	22.5	209	11	16.0
3	11.3	88.2	109.4	1919	10	18.2
4	84.7	151.6	277.0	7287	7	10.0
5	7.3	79.1	17.4	5311	15	17.5
6	17.9	60.4	93.0	6109	8	3.6
7	2.5	53.2	21.5	4057	17	18.5
8	27.3	108.5	114.5	4161	3	4.0
9	5.9	48.7	61.3	2166	10	11.6
10	23.9	142.8	129.8	11125	9	14.2
11	69.4	214.7	159.4	13937	2	2.5
12	20.6	65.6	91.0	4000	18	12.0
13	1.9	13.2	6.1	2841	14	12.8
14	3.0	60.9	60.3	1273	26	7.8
15	7.3	21.2	51.1	2404	34	2.7
16	46.2	114.3	73.6	6109	12	3.2
17	78.8	299.5	171.7	15571	4	7.6
18	11.1	78.9	38.8	4228	11	11.0
19	8.6	90.0	105.3	3772	15	28.4

图 1: Restaurant

首先附一个直接用 *AIC* 值筛选，用R软件得到的计算结果。这里其实采用了逐步回归的方法。

```
## Start:  AIC=123.39
## 营业额 ~ 居民数 + 人均餐费 + 月收入 + 餐馆数 + 距离
##
##             Df Sum of Sq    RSS    AIC
## - 月收入     1     35.96 2189.0 121.81
## - 餐馆数     1     79.17 2232.2 122.30
## <none>                   2153.0 123.39
## - 居民数     1    199.42 2352.4 123.61
## - 距离       1    392.54 2545.6 125.58
## - 人均餐费   1    942.22 3095.2 130.47
##
## Step:  AIC=121.81
## 营业额 ~ 居民数 + 人均餐费 + 餐馆数 + 距离
##
##             Df Sum of Sq    RSS    AIC
## - 餐馆数     1     78.22 2267.2 120.69
## <none>                   2189.0 121.81
```

```
## - 距离       1    445.69 2634.7 124.44
## - 人均餐费   1    925.88 3114.9 128.63
## - 居民数     1   1133.27 3322.3 130.24
##
## Step:  AIC=120.69
## 营业额 ~ 居民数 + 人均餐费 + 距离
##
##             Df Sum of Sq    RSS    AIC
## <none>                   2267.2 120.69
## - 距离       1    404.28 2671.5 122.79
## - 人均餐费   1   1050.90 3318.1 128.21
## - 居民数     1   1661.83 3929.0 132.43
```

希望用表格中数据，计算式（4.1）的估计值。

$$\hat{r}^2 = \frac{tr\left[W\left\{\frac{1}{\hat{\sigma}_Y^2}\tilde{Y}\tilde{Y}^t - (I - \mathbf{1}_n\mathbf{1}_n^t/n)\right\}\right]}{tr\left[W\left\{M - (I - \mathbf{1}_n\mathbf{1}_n^t/n)\right\}\right]} \tag{4.1}$$

4.1.1 选择变量 X1,X2,X3,X4,X5

首先，容易计算出

$$\begin{aligned}\tilde{Y} =&(26.756, -7.944, -15.144, 58.256, -19.144,\\ &-8.544, -23.944, 0.856, -20.544, -2.544,\\ &42.956, -5.844, -24.544, -23.444, -19.144,\\ &19.756, 52.356, -15.344, -17.844, 22.456,\\ &-4.344, -15.344, -17.844, 22.456, -4.344)^t\\ \hat{\sigma_Y^2} =&113.3172\end{aligned} \tag{4.2}$$

借助对角化，对特征值求根的方法可以求出 $\Sigma^{-1/2}$, 此时令 $X_{n\times p}$ 为原数据，则 $Z_{25\times 5} = X_{25\times 5}\Sigma^{-1/2}$ 为标准化的数据，这里可以验证 $var(Z)$ 是单位阵 I_5 .

我这里取 W 中的 $\lambda = 0.1$. 文章中提到在正态假设下 λ 取 $\frac{r^2}{1-r^2}$ 可以得到最好的估计，在后面的数据模拟中提到可以先取 $\lambda = 0.1$,再用 $\lambda = \frac{\hat{r^2}}{1-\hat{r^2}}$ 作五次迭代。

直接取 $\lambda = 0.1$ 可以得到 r^2 的一个估计 $\hat{r^2} = 4.341748$, 这是五个自变量 $X1, X2, X3, X4, X5$ 都选取的结果。

4.1.2 考虑四个变量

仍然用上面的方法，任意选取四个变量估计 $\hat{r^2}$.这里考虑相对于上面的模型需要改变的地方，便是矩阵 Z 和由它影响的 M .结果如表中所示。表格中第一行表示删去某个变量，第二行是用其他四个变量得到的估计值。

	X_1	X_2	X_3	X_4	X_5
$\hat{r^2}$	4.303755	3.976004	4.375883	4.356814	4.218545

这时，可以得到结论，在去掉 X_3 以后，得到的解释量的估计值最大，因此我们选择删掉影响不大的 X_3 来作拟合。我们惊喜地发现，这和用 AIC 值来估计作出的判断相同!

4.1.3 考虑三个变量

现在我们已经使用了四个变量 X_1, X_2, X_4, X_5 拟合，得到的最大解释量为 $\hat{r^2} = 4.375883$ 继续运算，看是否有去掉某个变量能得到更好的效果。这时，我们考虑分别去掉这四个变量的拟合，列表同上。

	X_1	X_2	X_4	X_5
$\hat{r^2}$	3.945643	4.032796	4.389009	4.234587

去掉 X_4 的结果 4.389009 显著地优于表格中其他项，并且去掉以后优于选择四个变量的结果。故我们做出判断： X_4 影响不大，可以去掉！发现和 AIC 值计算得到的结论符合！泪目了……

4.1.4 考虑两个变量

可能是有点象征性地，我要说明选择前面三个变量是最优的，即不必再删去某个变量。这时，我们就要去计算只剩两个变量时的 $\hat{r^2}$ 估计值。

	X_1	X_2	X_5
$\hat{r^2}$	3.765712	4.010769	4.270146

解释的效果不如从前，结果符合我们之前用 AIC 值计算的方法!

4.2 学生身高体重年龄

同样是李东风老师讲义 （回归自变量筛选） （例33.3）中的数据，如图(2)。

	name	sex	age	height	weight
1	Alice	F	13	56.5	84.0
2	Becka	F	13	65.3	98.0
3	Gail	F	14	64.3	90.0
4	Karen	F	12	56.3	77.0
5	Kathy	F	12	59.8	84.5
6	Mary	F	15	66.5	112.0
7	Sandy	F	11	51.3	50.5
8	Sharon	F	15	62.5	112.5
9	Tammy	F	14	62.8	102.5
10	Alfred	M	14	69.0	112.5
11	Duke	M	14	63.5	102.5
12	Guido	M	15	67.0	133.0
13	James	M	12	57.3	83.0
14	Jeffrey	M	13	62.5	84.0
15	John	M	12	59.0	99.5
16	Philip	M	16	72.0	150.0
17	Robert	M	12	64.8	128.0
18	Thomas	M	11	57.5	85.0
19	William	M	15	66.5	112.0

图 2: Class

考虑体重受三个自变量：身高，年龄，性别的影响。直接用上述程序，进行 7 次计算，选择最大的 $\hat{r^2}$ 所对应的模型。

首先附一个直接用 AIC 值筛选，用R软件得到的计算结果。这里其实采用了逐步回归的方法。

```
## Start:  AIC=94.93
## weight ~ height + age + sex
##
##          Df Sum of Sq    RSS     AIC
## - age     1    113.76 1957.8  94.067
## <none>                1844.0  94.930
## - sex     1    276.09 2120.1  95.581
## - height  1   1020.61 2864.6 101.299
##
## Step:  AIC=94.07
## weight ~ height + sex
##
##          Df Sum of Sq    RSS     AIC
## - sex     1     184.7 2142.5  93.780
## <none>                1957.8  94.067
## - height  1    5696.8 7654.6 117.974
##
## Step:  AIC=93.78
## weight ~ height
##
##          Df Sum of Sq    RSS    AIC
## <none>                2142.5  93.78
## - height  1    7193.2 9335.7 119.75
```

可见，第一步筛掉了 X_2, 第二步筛掉了 X_3, 最终得到的回归模型，体重只收身高 X_1 的影响。

两到三个变量拟合的运行结果如下：

```
> r2_es(cbind(X1,X2,X3),3)
[1] 3.218943
> r2_es(cbind(X2,X3),2)
[1] 2.762563
> r2_es(cbind(X1,X3),2)#
[1] 3.223606
```

```
> r2_es(cbind(X1,X2),2)
[1] 3.141079
```

则我们首先排除 X_2 对 Y 的影响，再去考察只有一个自变量时，回归方程是否更优。

只有一个自变量时，注意到 Σ 是一个数，所以它的 $-\frac{1}{2}$ 次方可以直接代数运算！对原来的算法稍作修改，得到结果：

```
> onevar(cbind(X1),1)
[1] 3.19377
> onevar(cbind(X2),1)
[1] 2.203877
> onevar(cbind(X3),1)
[1] 0.5562381
```

可见选用 X_1 指标拟合得到的结果明显优于 X_2 和 X_3, 但不如选用两个自变量 X_1, X_3, 因此用这种方法做出的选择是选用变量 X_1, X_3 拟合，和用 AIC 作出的判断不完全相同。

这里把性别男女化为两个因子 0,1 或 1,2 来计算，且运算过程中和其他数据地位相同，所以难免会出现偏差。

4.3 模拟数据的计算

4.3.1 模型描述与模拟数据的生成

由于对于原始数据标准化等操作不失一般性，不改变问题本质，所以我们直接考虑模型：

$$Y = Z_1\alpha_1 + \cdots + Z_p\alpha_p + \epsilon \tag{4.3}$$

不妨选取 $(Z_1, \cdots, Z_p)$ 服从多元标准正态分布，且相互独立， $\epsilon \sim N(0,1)$ 并与前者相互独立。 $(Z_{i1}, \cdots, Z_{ip}), \epsilon_i$ 是上述总体的简单随机样本。

为了验证我们所提出的解释量 r^2 对于自变量选择有较好的判断能力，我们将生成一组模拟数据， $(X_{i1}, \cdots, X_{i,100}), \epsilon_i$, s.t. $i = 1, \cdots, N$, N 待定，然后用 $Y_i = X_{i1}\alpha_1 + \cdots + X_{i,20}\alpha_i + \epsilon_i$ 得到 Y_i 的模拟数值。

希望给出的模拟数据中， Y 只由受前 20 个自变量的影响。所以可以作出系数假定：

$$\alpha_i = \begin{cases} U(1,2), & i = 1, \cdots, 20 \\ 0, & otherwise \end{cases} \tag{4.4}$$

若令 $X_{100\times100}$ 是每个元素都服从 $N(0,1)$ 的随机矩阵， ϵ 是一个 $100+\times1$ 的列向量。则可以由关系式 $Y=X\alpha+\epsilon$ 得到 Y 的预测值。由此我们得到了一组满足假定的随机模拟的数据。

4.3.2 解释量 $\hat{r^2}$ 的计算

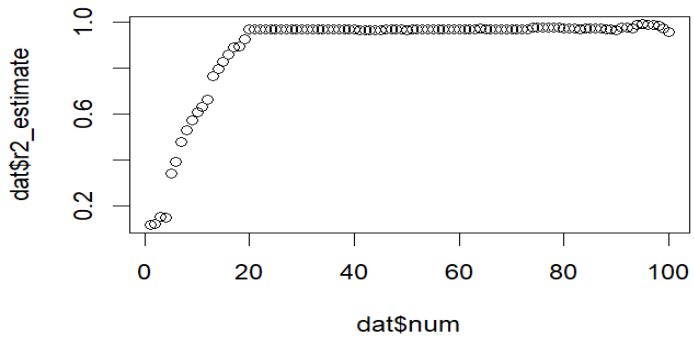
由上述模拟数据我们可以计算 $\hat{r^2}$, 利用公式

$$\hat{r}_{\lambda}^{2}=\frac{tr\left\{W_{\lambda}\left(YY^{t}-I\right)\right\}}{tr\left\{W_{\lambda}\left(M-I\right)\right\}}, \tag{4.5}$$

其中 $M=XX^{t}/p$, $W_{\lambda}=\left(I+\lambda M\right)^{-1}\left(M-I\right)\left(I+\lambda M\right)^{-1}$

首先计算符合假定模型的结果。取原始数据 X 的前 20 列计算。注意这里需要将 Y 进行标准化操作。得到的模拟结果：

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	1	[1]	0.09736878	21	[1]	0.9797035	41	[1]	0.9784787	61	[1]	0.9792852	81	[1]	0.9809385
2	2	[1]	0.2261301	22	[1]	0.979466	42	[1]	0.9781549	62	[1]	0.9787416	82	[1]	0.98163
3	3	[1]	0.2661301	23	[1]	0.9792034	43	[1]	0.9784784	63	[1]	0.9802377	83	[1]	0.9805502
4	4	[1]	0.2896485	24	[1]	0.9791103	44	[1]	0.9781704	64	[1]	0.9806611	84	[1]	0.9812378
5	5	[1]	0.3360683	25	[1]	0.9789663	45	[1]	0.977806	65	[1]	0.9808029	85	[1]	0.9801707
6	6	[1]	0.4170056	26	[1]	0.9786821	46	[1]	0.9773962	66	[1]	0.9803394	86	[1]	0.978874
7	7	[1]	0.4858603	27	[1]	0.978516	47	[1]	0.9771864	67	[1]	0.9798279	87	[1]	0.9789306
8	8	[1]	0.5121125	28	[1]	0.9782302	48	[1]	0.9768349	68	[1]	0.979298	88	[1]	0.9791218
9	9	[1]	0.5147337	29	[1]	0.9781337	49	[1]	0.9764161	69	[1]	0.9786302	89	[1]	0.9787803
10	10	[1]	0.5098731	30	[1]	0.9778468	50	[1]	0.978028	70	[1]	0.9780274	90	[1]	0.9767642
11	11	[1]	0.5246863	31	[1]	0.9776236	51	[1]	0.9775937	71	[1]	0.9792927	91	[1]	0.9787475
12	12	[1]	0.5342902	32	[1]	0.9777589	52	[1]	0.9774212	72	[1]	0.9803543	92	[1]	0.9762516
13	13	[1]	0.561797	33	[1]	0.9775856	53	[1]	0.9780346	73	[1]	0.9799794	93	[1]	0.9732936
14	14	[1]	0.638979	34	[1]	0.9777226	54	[1]	0.9775834	74	[1]	0.981253	94	[1]	0.9696846
15	15	[1]	0.7181378	35	[1]	0.9774795	55	[1]	0.9775336	75	[1]	0.9806219	95	[1]	0.963649
16	16	[1]	0.7302372	36	[1]	0.9771356	56	[1]	0.977101	76	[1]	0.9799877	96	[1]	0.9698897
17	17	[1]	0.8146078	37	[1]	0.9768335	57	[1]	0.9778186	77	[1]	0.980948	97	[1]	0.9598654
18	18	[1]	0.8693412	38	[1]	0.977088	58	[1]	0.9775117	78	[1]	0.9806014	98	[1]	0.9669369
19	19	[1]	0.927848	39	[1]	0.9767949	59	[1]	0.9776317	79	[1]	0.9812573	99	[1]	0.9468272
20	20	[1]	0.979949	40	[1]	0.9770107	60	[1]	0.9796098	80	[1]	0.9803434	100	[1]	0.9289494
21															



r^2 的含义为自变量 X 变化对于 Y 可以做出的解释的比例，是相较于随机误差而言自变量的影响程度大小，当自变量恰好选择为真实模型时，解释量几乎为 1 受到一点随机扰动的影响，若选择的自变量较少，解释不充分，剩余的波动都归于随机误差，所以解释量下降。

当模型选择完全，再添加多余的自变量时，解释的程度不会增加。

5 进一步尝试

由上面的初步模拟可以得到 $\hat{r^2}$ 随自变量个数增加的图像，如图（4.3.2）所示，可以看出在有效自变量缺失时， $\hat{r^2}$ 的估计值下降，当选取全部有效变量之后达到(有扰动意义下的)最大值，若增加多余变量基本不变。类似于 AIC 的想法，想通过增加和变量个数有关的惩罚项来得到一种选取最优模型的算法。由于 r^2 越大越好，因此对于变量个数增加我们需要给一个负的惩罚项，以避免做出错误判断。

假设选择变量的个数过少，再增加一项会使得 $\hat{r^2}$ 有一个比较明显的增加，因此为了体现出这一点，惩罚项与这个增加相比不宜更大，否则会埋没自变量对于因变量解释更充分的作用。

若已经达到最优，再增加自变量后 $\hat{r^2}$ 会基本不变，如果视为不变则惩罚项取任意数都能够保证排除更多变量的方案。但由于随机因素，变量的增加可能对于 $\hat{r^2}$ 有一个随机扰动的影响，比如在达到最优以后，再加一个自变量，可能由于随机因素使得显示出来的 $\hat{r^2}$ 更大，因此增加的惩罚项 $f(p)$ 应该大于这个扰动的影响，这样可以做到当有多余项出现时，利用惩

罚项的惩罚作用把变量更多的方案排除。

根据上述表述，我们希望得到关于自变量个数 p 的一个函数 $f(p)$ 作为惩罚项，且满足：

$$Q_1 > f(q) > Q_2, \tag{5.1}$$

其中 Q_1 是为了防止选择过少产生，大概意思是当自变量不充足时增加一个自变量使解释量 r^2 的增加量，类似的， Q_2 为自变量充足时的随机扰动效应，惩罚项的最初目的便是使自变量过多时加以限制。

需要注意的是，在真实情况是有 r 个自变量对因变量产生影响，且影响程度近似相同（上述模拟中采用系数全部为 $1 \sim 2$ 的随机数保证）的模型下，用全部这些自变量的得到的 r^2 值接近 1, 每个自变量对于解释量 r^2 的贡献是 $\frac{1}{r}$ 的量级。注意到当 r 较大时（趋于 ∞ ），每个自变量对于因变量的影响都很小，可能添加一个自变量对于 r^2 的影响甚至不如随机误差的扰动，因此难以抉择，我们暂且不考虑这种情况。当 r 较小时，每个自变量对于因变量的影响较大,我们可以明显的看出如何选取自变量使得解释最优，且不多选。下面我将计算一些 r 较小的模拟数据。

Remark 5.1. 这里解释量全部用 r^2 表示，有真实影响的自变量个数我选用了字母 r ，但愿不会产生混淆。

5.1 真实影响自变量个数较少的随机模拟

同样是上一节计算过的模型，这时我们设真实影响的自变量个数为 $r = 5$ ，仍然考虑 100 个数据时的情况

每个自变量的引入都会使得估计值有一个较大的增长，但是这里会发现问题：当选用模型自变量个数较大，和样本数量相当时，估计值会有一个很大的波动，如图（3）我做了五十次模拟，每次都随机产生 X , 扰动 ϵ 以及模型系数，并绘图得到结果。这里我放一个最后波动最夸张的图。

当样本数量远大于模型中自变量数量 p 时，得到的结果和所想基本一致，如图(4).

由论文中证明得到的估计量 $\hat{r^2}$ 在大样本下的渐进分布，我们自然希望样本量 n 足够大。下面进行稍一般性的模拟：将 5 个自变量随意放到 100 个中间。结果如图(5)所示。由图可以看出，在这个简化模型下，这种方法已经可以对于自变量选择做出较好判断，基本不会出错。但由于自变量系数和误差产生的随机性，增加一个有效自变量使解释量的增大和无效量引起的波动可能还不能得到一个较好的区分。

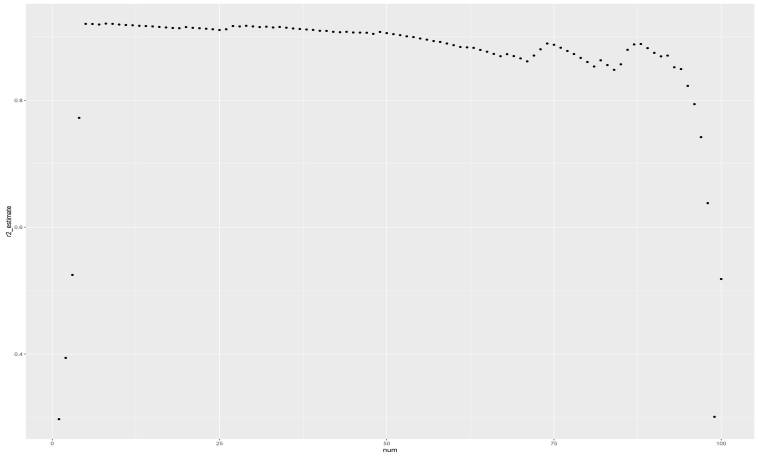

r2_estimate
0.8
0.6
0.4
0
25
50
75
100
num

图 3: 小样本

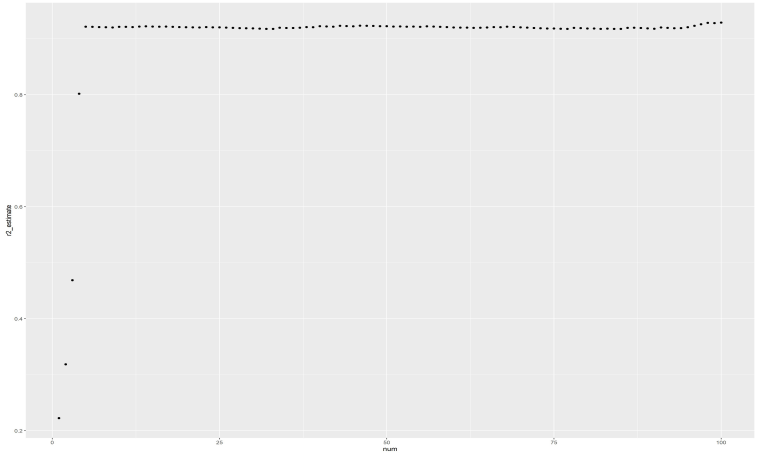

r2_estimate
0.8
0.6
0.4
0.2
0
25
50
75
100
num

图 4: 大样本

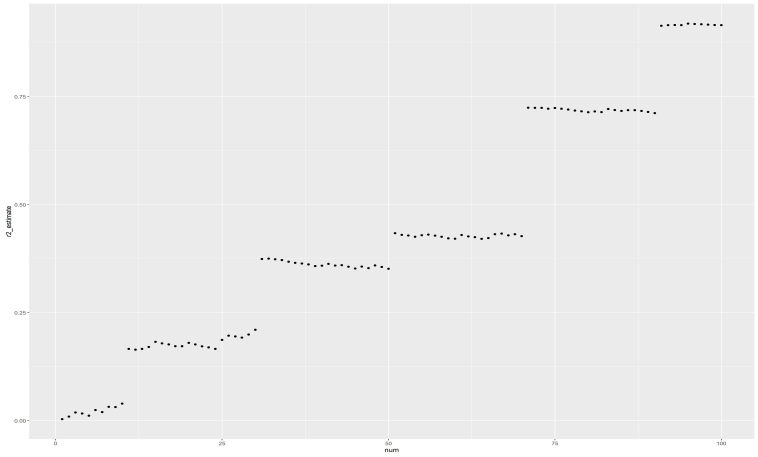


图 5: 有效自变量随意落在全部可能有影响的量中

参考文献

[1] Hua Yun Chen. Statistical inference on explained variation in high-dimensional linear model with dense effects. *arXiv preprint arXiv:2201.08723*, 2022.