Universidade Federal do Paraná

Laboratório de Estatística e Geoinformação - LEG

CE071 - Análise de Regressão

Seleção de Variáveis

Introdução

- No processo de construção do modelo, um dos problemas mais importantes é escolher quais variáveis devem permanecer no modelo.
- A seguir discutiremos alguns critérios estatísticos para fazer isso.
- Contudo, lembre-se que a construção do modelo é uma tarefa realizada em conjunto com o pesquisador da área (quem tem uma opinião sobre a importância de cada variável).

Colinearidade

- Duas variáveis X_1 e X_2 são ditas exatamente colineares se $c_1X_1 + c_2X_2 = c_0$ para algumas constantes c_i.

- Entretanto, na análise de dados é mais comum haver variáveis aproximadamente colineares do que exatamente colineares.

- Assim, duas variáveis X_1 e X_2 são ditas aproximadamente colineares se $c_1X_1 + c_2X_2 \approx c_0$.

- A colinearidade entre é medida através do quadrado do coeficiente de correlação entre as variáveis, r_{12}^2.

  - Colinearidade exata ocorre quando $r_{12}^2 = 1$;
  - Mas, se $r_{12}^2 \approx 1$, devemos nos preocupar.

Colinearidade

- Quando temos p variáveis explanatórias, $X_1, \ldots, X_p$, elas são ditas (aproximadamente) colineares se $c_1 X_1 + \ldots + c_p X_p \approx c_0$.

- Para um j específico,

$$X_j \approx \frac{c_0}{c_j} - \sum_{i \neq j} \frac{c_i}{c_j} X_i$$

- Esta expressão corresponde (aproximadamente) a uma regressão com intercepto $\frac{c_0}{c_j}$.

- Portanto, as variáveis $X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p$ explicam X_j se o coeficiente de determinação R_j^2 for próximo de um.

- Se as variáveis são exatamente colineares, a solução das equações normais não é única.

Colinearidade

- A colinearidade (aproximada) tem um efeito na precisão dos coeficientes de regressão.
- Com efeito, numa regressão com duas variáveis explanatórias e intercepto,

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{1 - r_{12}^2} \frac{1}{S_{xx_j}}, \quad j = 1, 2.$$

- Então obtemos a menor variância quando $r_{12}^2 = 0$ e a variância aumenta (inflaciona) na medida em que r_{12}^2 se aproxima de um.
- Uma expressão equivalente para p variáveis explanatórias é,

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{1 - R_j^2} \frac{1}{S_{xx_j}}, \quad j = 1, 2.$$

em que $\frac{1}{(1-R_j^2)}$ é conhecido como *Fator de inflação da variância (FIV)*.

- De forma que o *FIV* representa o aumento na variância devido à correlação entre as variáveis explanatórias.

Métodos de seleção de variáveis

- Será discutido o seguinte problema: temos p variáveis explanatórias e queremos escolher o(s) modelo(s) de regressão linear múltipla com base em todas ou subconjuntos dessas variáveis.

- Supondo que o intercepto sempre faz parte da regressão, então temos 2^p possíveis modelos.

- Isto significa que com $p = 4$ temos que analisar 16 regressões.

- E com $p = 10$ teríamos que analisar 1024 ajustes.

- Nesta última situação é útil dispor de mecanismos para escolher modelos evitando ajustar todas as possibilidades.

Métodos de seleção de variáveis

- Seja $SQRes_k$ a soma de quadrados de resíduos de um ajuste com k termos: um intercepto e $(k-1)$ variáveis explanatórias.

- A seguir apresentamos alguns critérios para escolha do modelo:

  1. **Quadrado médio da regressão,** s^2: Sendo o estimador não viciado de σ^2, preferimos modelos com baixo s^2.
  2. **Coeficiente de Determinação Ajustado,** $\bar{R}^2$**:** É preferível ao R^2 já que não necessariamente aumenta quando incluímos variáveis no modelo.
  3. **Estatística** C_p **de Mallows:** Preferível aos dois critérios anteriores. Esta estatística é definida como

$$C_k = \frac{SQRes_k}{s^2} + 2k - n,$$

preferimos os modelos tais que $C_k < k$, e entre estes, aqueles com menores valores de C_k.

Métodos de seleção de variáveis

1. **Critério de Informação de Akaike, AIC**: A ideia é fazer um compromisso entre grau de ajuste (o qual aumenta quando aumentamos variáveis) e tamanho do modelo ou complexidade.

$$AIC = n \ln(SQRes_k/n) + k,$$

preferimos o modelo com menor AIC.

2. **Critério de Informação Bayesiano, BIC:** Já que com bastante frequência o AIC conduze a escolher o modelo com maior número de termos, Schwarz (1978) propôs o BIC como

$$BIC = n \ln(SQRes_k/n) + k \ln(n),$$

aqui também preferimos o modelo com menor BIC.

3. **Validação Cruzada:** Consiste em dividir a amostra em dois grupos: uma amostra *treinamento* e *teste*. O modelo é ajustado com base na amostra de treinamento e são calculados os valores preditos dos casos na amostra de teste. Depois comparamos os valores preditos com os valores reais (segundo algum critério).

Exemplo

- Nos dados temos 4 variáveis explicativas $(X_1, \ldots, X_4)$ e uma resposta (Y)

X_1	X_2	X_3	X_4	Y
7	26	6	60	78.5
1	29	15	52	74.3
11	56	8	20	104.3
11	31	8	47	87.6
7	52	6	33	95.9
11	55	9	22	109.2
3	71	17	6	102.7
1	31	22	44	72.5
2	54	18	22	93.1
21	47	4	26	115.9
1	40	23	34	83.8
11	66	9	12	113.3
10	68	8	12	109.4

Exemplo

Variáveis	SQR	QMR	R^2	C_p	AIC	BIC
Constante	2715.76	226.31	0	442.99	263.95	275.67
X_1	1265.69	115.06	0.53	202.55	156.51	170.73
X_2	906.34	82.39	0.67	142.49	112.07	122.25
X_3	1939.40	176.31	0.29	315.15	239.83	261.61
X_4	883.87	80.35	0.67	138.73	109.30	119.22
X_1, X_2	57.90	5.79	0.98	2.68	9.50	10.83
X_1, X_3	1227.07	122.71	0.55	198.09	194.68	221.79
X_1, X_4	74.76	7.48	0.97	5.50	11.87	13.52
X_2, X_3	415.44	41.54	0.85	62.44	65.90	75.08
X_2, X_4	868.88	86.89	0.68	138.23	137.85	157.05
X_3, X_4	175.74	17.57	0.94	22.37	27.88	31.76
X_1, X_2, X_3	48.11	5.35	0.98	3.04	9.90	11.78
X_1, X_3, X_4	50.84	5.64	0.98	3.50	10.44	12.42
X_1, X_2, X_4	47.97	5.33	0.98	3.02	9.86	11.73
X_2, X_3, X_4	73.81	8.20	0.97	7.34	15.17	18.05
X_1, X_2, X_3, X_4	47.86	5.98	0.98	5.00	12.91	16.04

Métodos Sequenciais Automáticos

- Mesmo com um número moderado de variáveis explanatórias, digamos 10, é conveniente dispor de algum roteiro para examinar um subconjunto de todas as possíveis regressões.

- Neste sentido, na literatura tem sido proposto métodos automáticos, os quais comparam basicamente a SQReg extra.

- Estudaremos três métodos:

  - **Seleção Forward**: começamos com o menor modelo (somente intercepto) e vamos adicionando variáveis uma a uma;
  - **Eliminação Backward**: começamos com o modelo completo e vamos eliminando variáveis uma a uma;
  - **Seleção Stepwise**: funciona como *Forward*, mas em cada passo é avaliada a entrada de variáveis incluídas anteriormente.

Seleção Forward

- Começamos com o menor modelo (somente intercepto) e vamos adicionando variáveis uma a uma. Em cada passo:

1. O modelo base tem q termos. Ajustamos as $(p-q)$ regressões (correspondentes as potenciais $(p-q)$ variáveis explanatórias) e registramos as somas de quadrados da regressão de cada potencial modelo.

2. Identificar a variável X_i, cuja incorporação ao modelo base proporciona o maior valor de $SQReg$, denominado $SQReg_{q+1} = SQReg_q + SQReg_{extra}$.

3. Teste se β_i é significante através de:

$$F_0 = \frac{SQReg_{extra}/1}{SQRes_{q+1}/(q+1)}.$$

Exemplo

Variáveis	SQR	QMR	R^2	C_p	AIC	BIC
Constante	2715.76	226.31	0	442.99	263.95	275.67
X_1	1265.69	115.06	0.53	202.55	156.51	170.73
X_2	906.34	82.39	0.67	142.49	112.07	122.25
X_3	1939.40	176.31	0.29	315.15	239.83	261.61
X_4	883.87	80.35	0.67	138.73	109.30	119.22
X_1, X_2	57.90	5.79	0.98	2.68	9.50	10.83
X_1, X_3	1227.07	122.71	0.55	198.09	194.68	221.79
X_1, X_4	74.76	7.48	0.97	5.50	11.87	13.52
X_2, X_3	415.44	41.54	0.85	62.44	65.90	75.08
X_2, X_4	868.88	86.89	0.68	138.23	137.85	157.05
X_3, X_4	175.74	17.57	0.94	22.37	27.88	31.76
X_1, X_2, X_3	48.11	5.35	0.98	3.04	9.90	11.78
X_1, X_3, X_4	50.84	5.64	0.98	3.50	10.44	12.42
X_1, X_2, X_4	47.97	5.33	0.98	3.02	9.86	11.73
X_2, X_3, X_4	73.81	8.20	0.97	7.34	15.17	18.05
X_1, X_2, X_3, X_4	47.86	5.98	0.98	5.00	12.91	16.04

Eliminação Backward

- Começamos com o modelo completo e vamos eliminando variáveis uma a uma. Em cada passo:

1. Seja o modelo base formado por q termos. Ajustamos as $(q-1)$ regressões e registramos as correspondentes SQReg.

2. Identificar a variável X_i, cuja eliminação do modelo base proporciona o maior valor de $SQReg$, denominado $SQReg_{q-1} = SQReg_q - SQReg_{extra}$.

3. Teste se β_i é significante através de:

$$F_0 = \frac{SQReg_{extra}/1}{SQRes_q/(q)}.$$

Stepwise

- Proceder como *Forward*, mas avaliando em cada passo a conveniência de retirar qualquer uma das variáveis presentes no modelo. Por exemplo:

Variáveis	SQR	QMR	R^2	C_p	AIC	BIC
Constante	2715.76	226.31	0	442.99	263.95	275.67
X_1	1265.69	115.06	0.53	202.55	156.51	170.73
X_2	906.34	82.39	0.67	142.49	112.07	122.25
X_3	1939.40	176.31	0.29	315.15	239.83	261.61
X_4	883.87	80.35	0.67	138.73	109.30	119.22
X_1, X_2	57.90	5.79	0.98	2.68	9.50	10.83
X_1, X_3	1227.07	122.71	0.55	198.09	194.68	221.79
X_1, X_4	74.76	7.48	0.97	5.50	11.87	13.52
X_2, X_3	415.44	41.54	0.85	62.44	65.90	75.08
X_2, X_4	868.88	86.89	0.68	138.23	137.85	157.05
X_3, X_4	175.74	17.57	0.94	22.37	27.88	31.76
X_1, X_2, X_3	48.11	5.35	0.98	3.04	9.90	11.78
X_1, X_3, X_4	50.84	5.64	0.98	3.50	10.44	12.42
X_1, X_2, X_4	47.97	5.33	0.98	3.02	9.86	11.73
X_2, X_3, X_4	73.81	8.20	0.97	7.34	15.17	18.05
X_1, X_2, X_3, X_4	47.86	5.98	0.98	5.00	12.91	16.04

Conclusão

- Diferentes métodos sequenciais não necessariamente conduzem ao mesmo modelo.
- Mais ainda, em geral nada garante que o *melhor* subconjunto de regressoras seja escolhido.
- Na prática, é melhor identificar vários *bons modelos* candidatos para serem examinados, do que aceitar sem criticas o modelo escolhido pelo computador.
- É importante considerar os modelos no contexto da aplicação prática, e a natureza das variáveis.
- Adicionalmente, os modelos escolhidos devem ser validados, utilizando técnicas de diagnóstico, e submetidos à avaliação junto ao pesquisador antes de se tirar conclusões.