Seleção de Modelos de Regressão - Parte 2

Fernado Bispo, Jeff Caponero

Sumário

Apresentação

O relatório desta semana está dividido em duas atividades. Na segunda atividade, se buscou determinar a o melhor modelo para determinar o PSA de pacientes com câncer de próstata, utilizando técnicas de colinearidade das variáveis de regressão linear múltipla.

Atividade 1

Atividade 2

Introdução

Com base nos dados disponibilizados no *dataset* "prostate" (do pacote *faraway*), que apresenta dados referentes ao câncer de prostata de 97 pacientes, cujo objetivo foi identificar os pacientes com indicação de prostatectomia tooal. As informações disponíveis na base de dados referem-se a:

- **lcavol**: é o volume do câncer (logarírimo);
- **lweight**: é o peso (massa) da próstata (logarítimo);
- **age**: é a idade do paciente (anos);
- **lbph**: é a quantidade de hiperplasia benigna na próstata (logarítimo);
- **svi**: é a invasão da vesícula seminal (porcentagem);
- **icp**: é a penetralção capsular(logarítimico);
- **gleason**: é o Índice de Gleason;
- **pgg45**: é a porcentagem do Índice de Gleason de valor 4 ou 5 (porcentagem);
- **lpsa**: é o valor do PSA (logarítimico);

Fonte: Andrews, D.F.; Herzberg, A.M. (1985): Data. New York: Springer-Verlag.

Com base nestes dados, objetiva-se:

Utilizar os métodos R^2, backward e stepward (Akaike e Bayesiano) como métodos de seleção de covariáveis para determinar o "melhor" ajuste de um modelo de regressão linear múltiplo.

Resultados

Análise descritiva dos dados

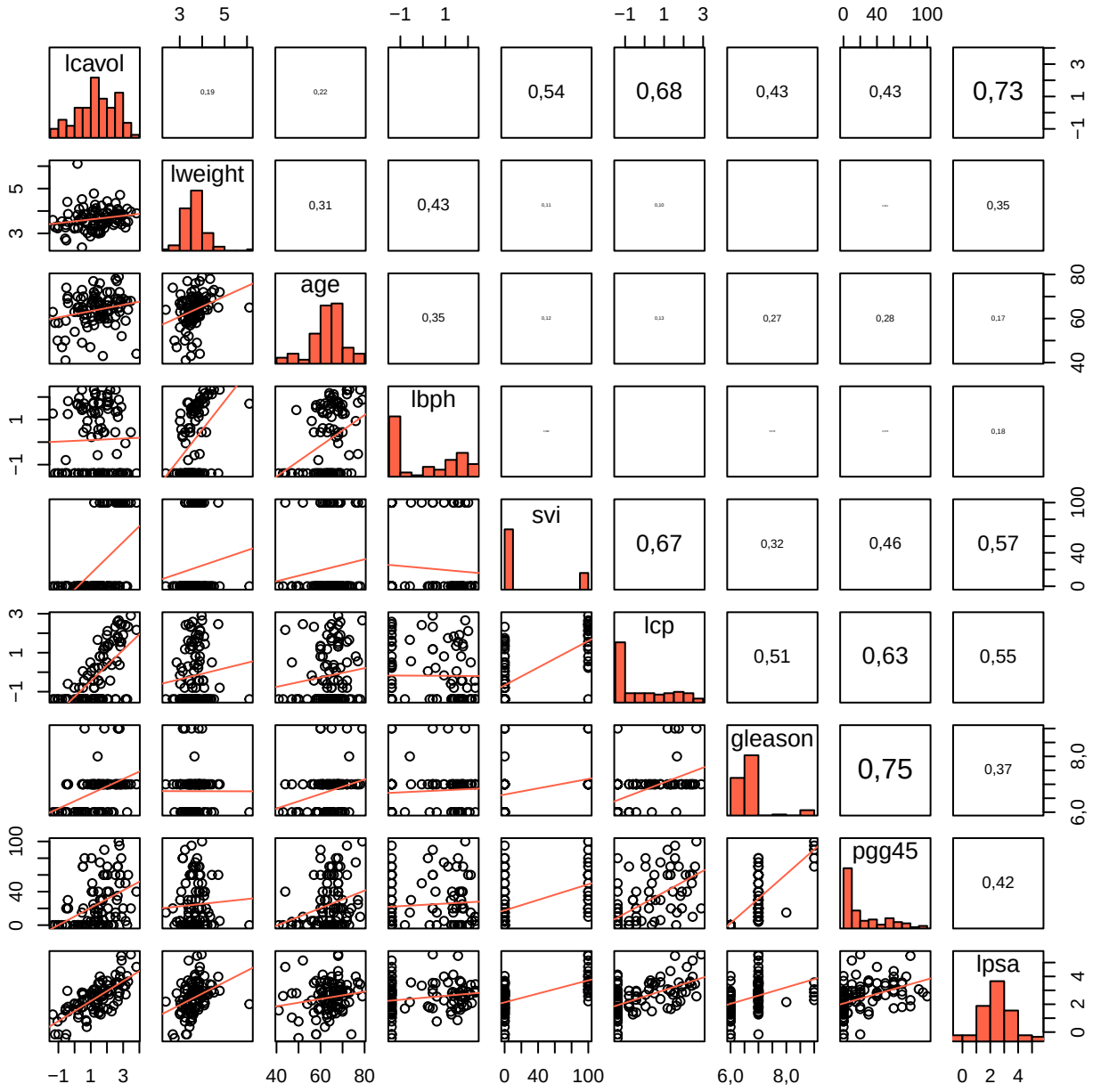
É possivel realizar uma descrição prévia dos dados por meio de medidas de resumo e da matriz de correlação como vê-se a seguir:

Tabela 1: Medidas Resumo dos dados

	Mín	Q1	Med	Média	Q3	Máx	Desv.padrão	CV	Assimetria	Curtose
Hiperplasia Benigna	-1,39	-1,39	0,30	0,10	1,56	2,33	1,45	14,46	0,13	-1,75
Idade	41,00	60,00	65,00	63,87	68,00	79,00	7,45	0,12	-0,80	0,96
Índice de Gleason	6,00	6,00	7,00	6,75	7,00	9,00	0,72	0,11	1,22	2,36
Índice de Gleason de valor 4 ou 5	0,00	0,00	15,00	24,38	40,00	100,00	28,20	1,16	0,94	-0,37
Invasão da Vesicular	0,00	0,00	0,00	21,65	0,00	100,00	41,40	1,91	1,36	-0,16
Penetralção Capsular	-1,39	-1,39	-0,80	-0,18	1,18	2,90	1,40	-7,80	0,71	-1,01
Peso da Próstata	2,37	3,38	3,62	3,65	3,88	6,11	0,50	0,14	1,18	5,02
PSA	-0,43	1,73	2,59	2,48	3,06	5,58	1,15	0,47	0,00	0,43
Volume do Câncer	-1,35	0,51	1,45	1,35	2,13	3,82	1,18	0,87	-0,24	-0,60

Como os valores analisados de diversas variáveis é obtido a partir do logarítimo dos valores reais destas variáveis observa-se na Tabela 1 alguns valores negatívos sem que isso corresponda a valores impossíveis.

Figura 1: Correlograma das variáveis dos dados.



Os diagramas de dispersão da Figura 1, aparentemente há uma baixa relação linear entre diversos pares de variáveis, o que indica a princípio uma menor colinearidade entre elas. A análise doo coeficiente de correlação de pearson revela que o volume do câncer é a variável com maior correlação com o PSA (variável resposta) e que há uma forte colinearidade entre as duas variáveis que avaliam o Índice de Gleason, como esperado.

Modelo de Regressão Linear Múltipla

O modelo de regressão múltipla obtido pode ser representado por:

Y_i =0,669 + 0,587 X_{1i}+ 0,454 X_{2i} -0,02 X_{3i}+ 0,107 X_{4i}+ 0,008 X_{5i} -0,105 X_{6i}+ 0,045 X_{7i}+ 0,005 X_{8i}

Onde:
Y_i - PSA;
X_{1i} - Volume do Câncer;
X_{2i} - Peso da Próstata;
X_{3i} - Idade;
X_{4i} - Hiperplasia Benigna;
X_{5i} - Invasão da Vesicular;
X_{6i} - Penetralção Capsular;
X_{7i} - Índice de Gleason;
X_{8i} - Índice de Gleason de valor 4 ou 5;

Interpretando-se o modelo pode-se dizer que para cada variável, fixadas as demais condições (*Ceteris Paribus*), o valor do PSA: aumenta em 0,587 para cada unidade do volume do câncer; aumenta em 0,454 para cada unidade do peso da próstata; reduz em 0,02 a cada ano do paciente; aumenta em 0,107 a cada unidade de hiperplasia benigna, aumenta em 0,008 a cada unidade d e invasão vesicular, reduz em 0,105 a cada unidade de penetração capsular; aumneta em 0,045 a cada unidade do índice de Gleason; e aumenta 0,005 a cada unidade porcentual do índice de Gleason entre 4 e 5. Neste modelo o coeficiente de determinação calculado foi de $R^2 = 0{,}655$, o que denota que 65,5% da variância dos dados é explicada pelo modelo. Pode-se calcular o coeficiente de determinação ajustado igual a $R_a^2 = 0{,}619$.

Tabela 2: Intervalos de Confiança para os parâmetros estimados no MRLS.

	LI^1	LS^2
$\hat{\beta}_0$	-1,907	3,246
$\hat{\beta}_1$	0,412	0,762
$\hat{\beta}_2$	0,117	0,792
$\hat{\beta}_3$	-0,042	0,003
$\hat{\beta}_4$	-0,009	0,223
$\hat{\beta}_5$	0,003	0,013
$\hat{\beta}_6$	-0,286	0,075
$\hat{\beta}_7$	-0,268	0,358
$\hat{\beta}_8$	-0,004	0,013

Legenda:
[1] LI: Limite Inferior (2,5%)
[2] LS: Limite Superior (97,5%)
[*] Nível de Significância de 5%.

A Tabela 2 apresenta os intervalos de confiança calculados os coeeficientes da equação acima com 95% de confiança.

Seleçâo de Modelo

As seleções de modelo foram realizadas por meio do pacote `olsrr` da linguagem R.

Eliminação backward baseada no teste F

Por meio da função `ols_step_backward_p()` verificou-se a possibilidade de eliminação de variáveis com base no teste F. Após os cálculos verificou-se que a variável: "Índice de Gleason" foi descartada do modelo. Gerando um modelo com $R^2 = 0{,}654$ e um $R^2_{aju} = 0{,}627$.

Seleção stepwise baseada no critério de informação de Akaike

Por meio da função `ols_step_both_p()` verificou-se a possibilidade de inclusão de variáveis com base no critério de informação de Akaine. Após os cálculos verificou-se que as variáveis: "Volume do Câncer", "Peso da Próstata" e "Invasão Vesicular" são as que melhor descrevem o modelo por esse critério. Gerando um modelo com $R^2 = 0{,}654$ e um $R^2_{aju} = 0{,}627$.

Seleção stepwise baseada no critério de informação Bayesiano

Por meio da função `ols_step_both_p()` verificou-se a possibilidade de inclusão de variáveis com base no critério de informação Bayesiano. Após os cálculos verificou-se o mesmo resultado que com o critério de informação de Akaine, logo o modelo com $R^2 = 0{,}654$ e um $R^2_{aju} = 0{,}627$.

Todas as regressões possíveis, baseado no R2 ajustado

Por meio da função `ols_step_best_subset()`, utilizando-se o critério baseado no coeficiente de determinação ou do coeficiente de determinação ajustado, verificou-se que o modelo com todas as variáveis é o que melhor explica a variável "PSA". Entretanto o ganho de *performance* com o aumento do número de variáveis não justifica o aumento da complexidade do modelo. Verificou-se que o aumento de uma nova variável partindo-se do modelo obtido na seleção pelo critério *stepwise* obteve ou ganho de menos de 1% de explicação da variância da variável resposta. Desta forma, o modelo que utiliza apenas as variáveis "Volume do Câncer", "Peso da Próstata" e "Invasão Vesicular" parece ser melhor que todos os demais.

Conclusão

A utilização de diversos critérios para otimizar o modelo de regressão linear múltipla avaliado, permite verificar que por mais de um critério pode se chegar a um mesmo modelo de explicação do PSA de pacientes com câncer na próstada, qual seja, aquele baseado tão

somente nas variáveis “Volume do Câncer”, “Peso da Próstata” e “Invasão Vesicular” já que este parece ser melhor que todos os demais modelos testados.