Modelos Estatísticos Utilizados na Análise de Dados de Reservatórios de Petróleo

Rejane dos Santos Brito
Depto. de Informática e Estatística, DEINFO, UFRPE,
52171-900, Recife, PE
E-mail: janesbrito@gmail.com.

Getúlio José Amorim do Amaral
Depto. de Estatística, DE, UFPE,
50740-540, Recife, PE
E-mail: gjaa@feller.de.ufpe.br.

Resumo Apresentamos a proposta de um modelo que solucione o problema abordado em estudos geológicos referente a poços de petróleo. Afim de analisar o conjunto de dados, faz-se uso de modelagem estatística tendo como ferramentas a regressão logística e a análise discriminante. Os modelos de classificação em pauta permitem realizar a avaliação de formações que definem a capacidade produtiva e a valoração das reservas de óleo e gás de poços de petróleo. Portanto, baseado nos tipos de rochas, sendo estas encontradas em cada nível de profundidade de um poço, pretende-se viabilizar a obtenção de informações mais precisas sobre os tipos de rochas com o intuito de reduzir os custos até então existentes quando se deseja analisar os poços de petróleo.

Palavras-chave Análise de resíduos; curvas ROC; modelos de classificação.

Introdução

A análise deste trabalho basea-se num conjunto de dados obtidos de poços de petróleo tal que, através da regressão logística e da análise discriminante, propõe-se um modelo preditivo para identificar tipos de rochas (litologias) favoráveis na acumulação de petróleo.

Os modelos de classificação abordados permitem auxiliar na avaliação da capacidade produtiva e valoração das reservas de óleo e gás de poços de petróleo.

A justificativa para essa abordagem surge da extrema necessidade das empresas petrolíferas obterem informações de tipos de rochas (aqui rotulada como "fácies") através da perfuração de poços, sendo estas caracterizadas como "rochas reservatório" e "rochas não-reservatório". A primeira categoria apresenta como propriedade a capacidade de acumular e produzir petróleo. Em contra partida, a segunda categoria (não-reservatório) tem propriedade oposta à primeira.

O uso da análise discriminante tem como objetivo obter a diferenciação das fácies considerando a existência de dois grupos que representam estas, para então propor um modelo que melhor discrimine os grupos. No caso da técnica de regressão logística, é possível classificar os tipos de litologias referentes às fácies reservatório e não-reservatório devido ao relacionamento existente entre a variável resposta binária e as variáveis explicativas que representam as curvas de perfis elétricos, os diferentes poços e o zoneamento destes. Como variáveis importantes desses modelos, capazes de classificar as fácies, são usados perfis geofísicos que são comuns aos poços e aos demais poços do campo onde desejamos estimar os tipos litológicos.

Avalia-se assim a adequabilidade dos diferentes tipos de modelos a fim de propor um modelo final. A utilização das técnicas de diagnóstico permite a identificação das observações que sejam influentes nas estimativas dos parâmetros dos modelos. Em particular, utilizaremos as técnicas de diagnóstico em nosso modelo de regressão logística.

O banco de dados tem um total de 1615 amostras de perfis geofísicos referentes às informações litológicas de três poços de petróleo. A partir do processamento e interpretação dos perfis geofísicos, são obtidas informações importantes a respeito das rochas contidas nos poços, como: litologia, espessura, porosidade, prováveis fluidos existentes nos poros e saturações (Thomas *et al*, 2001, p. 121). O termo "fácies", e seus derivados, é informal e normalmente usado para definir categorias segundo um critério previamente estabelecido.

Modelo de Regressão Logística Para Respostas Binárias

Dado que $\pi(\mathbf{x})$ varia entre 0 e 1, uma simples representação linear $\mathbf{x}^T\boldsymbol{\beta}$ para $\pi(\cdot)$ sobre todo o intervalo de $\mathbf{x}$ é impossível. O fato de ser impossível decorre da ocorrência de um determinado evento ser uma

função não linear das variáveis explicativas. Por essa razão, realiza-se a linearização através do uso da transformação logística $g(\pi)$, conhecida por *logit*, cujo parâmetro canônico é dado por

$$\eta = g(\pi) = \log[\pi/(1-\pi)], \qquad (1)$$

em que a razão entre π e $1-\pi$ é denominada razão de chances.

Segundo Vittinghoff *et al.*(2005, p. 162) um dos mais significantes benefícios do modelo logístico é que os coeficientes de regressão são interpretados como o logaritmo da razão de chances.

O modelo geral de regresão logística é dado por

$$\log\left\{\frac{\pi_i}{1-\pi_i}\right\} = \beta_1 + \beta_2 x_2 + \cdots + \beta_k x_k, \qquad (2)$$

em que $\mathbf{x} = (1, x_2, \ldots, x_k)^T$ contém os valores observados das $(p-1)$ covariáveis.

Pretende-se analisar o relacionamento entre uma variável resposta binária, cuja representação apreende-se em indicar a ocorrência ou não de fácies reservatório, e a utilização de um conjunto de variáveis explicativas as quais representam curvas de perfis elétricos, diferentes poços e zoneamento destes. Por fim, deseja-se propor um modelo que prediza a variável observada de forma satisfatória.

Análise Discriminante

Na análise discriminante linear, assumimos que existem M grupos ou classes, cujas probablilidades são denotadas por $\pi_1, \ldots, \pi_M$, e uma variável categórica z é associada a cada obervação $\mathbf{x}$ sendo essa denotada como classe ou membro do grupo; isto é, se $z = i$, então a observação pertence ao π_i, $i \in 1, \ldots, M$.

Fisher (1936), ao analisar o problema de discriminação entre as M populações, teve como principal objetivo encontrar uma função linear $\boldsymbol{b}'\mathbf{x}$ de forma a maximizar a razão entre a soma dos quadrados entre os grupos e a soma dos quadrados dentro dos grupos. Ele não assume normalidade das observações populacionais, entretanto, assume implicitamente que a matriz de covariância das populações sejam iguais.

No caso em que a função discriminante é aplicada a duas populações, temos que a regra de classificação se resume a

$$w = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}^{-1}[\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2)], \qquad (3)$$

cuja denominação freqüente é dada pela função de classificação de Anderson (Johnson e Wichern, 1992, pp. 524), devido ao fato de existir equivalência entre a regra de classificação de Fisher e a regra da mínima estimativa do custo de erro com probabilidades a *priori* iguais e custos de erro de classificação iguais.

A aplicação da regra de classificação para dois grupos utilizando a função discriminante linear de Fisher, que foi definida em (3), indica que

$$w > 0 \Rightarrow \mathbf{x} \in \boldsymbol{\pi}_1,$$
$$w < 0 \Rightarrow \mathbf{x} \in \boldsymbol{\pi}_2.$$

Ao estudar o relacionamento da variável fácies com as demais variáveis, tal que essa variável seja representante de duas classes distintas, considera-se o grupo π_1 como sendo o representante da variável fácie não-reservatório e o grupo π_2 o representante da fácie reservatório.

Para a aplicação da discriminante linear logística, Anderson (1982) diz que o modelo discriminate logístico é uma descrição exata numa variedade de situações que incluem: densidades de classes condicionais normais multivariadas com matrizes de covariância iguais; distribuições discretas multivariadas seguindo um modelo log-linear com iguais termos de iteração entre os grupos; e a combinação de situações anteriores, ou seja, ambas variáveis contínuas e categóricas descrevendo cada amostra.

Análise de Dados em Reservatório de Petróleo

As variáveis utilizadas para a análise estatística das fácies são os perfis DT (sônico), GR (raios gama), ILD (indução), RHOB (densidade) e NPHI (neutrônico). Adiciona-se ainda as variáveis nomeadas "poço" e "zona" para que as observações sejam classificadas por diferentes poços e faixas nos tempos geológicos equivalentes. Usamos, portanto, como variável resposta a variável fácies e como variáveis explicativas as variáveis DT, GR, ILD, RHOB, NPHI, Poço e Zona.

O uso das variáveis *Poço* e *Zona* é justificado pela necessidade de descrevermos corretamente os tipos de litologia dado a profundidade do poço, sendo assim, a variável *Zona* segue uma ordem que depende da profundidade e esta ainda define as possíveis litologias a ocorrerem na mesma. Em outras palavras, temos que para cada poço existe a presença dos diferentes tipos de zona, e ainda, em cada zona há os característicos tipos de litologia.

A variável raios gama *GR* mede primariamente a radioatividade natural das rochas, ou seja, a radioatividade total da formação geológica. É utilizada para a identificação da litologia, a identificação de minerais radioativos e para o cálculo do volume de argilas ou argilosidade. Como regra geral, quanto mais radioativa a rocha menor a sua granulometria.

A variável indução *ILD* fornece a leitura aproximada da resistividade da rocha através da medição de campos elétricos e magnéticos induzidos. De forma geral, rochas porosas com óleo têm resistividade alta e com água salgada a resistividade é baixa.

A variável densidade *RHOB* é uma medida de densidade eletrônica que detecta os raios gama defletidos pelos elétrons dos elementos das rochas, após terem sido emitidos por uma fonte colimada situada dentro do poço. Além da densidade das rochas, *RHOB* permite o cálculo da porosidade e a identificação das zonas de gás. É utilizado também como apoio à sísmica para o cálculo do sismograma sintético.

Uma outra variável definida como sônico é o *DT* cuja finalidade é medir a diferença nos tempos de trânsito de uma onda mecânica através das rochas. É utilizado para estimativas de porosidade, correlação poço a poço, estimativas do grau de compactação das rochas ou estimativas das constantes elásticas, detecção de fraturas e apoio às sísmicas para a elaboração do sismograma sintético.

A variável porosidade neutrônica *NPHI* é uma medida de densidade da rocha. É medida sob o aspecto da emissão de nêutrons. Os perfis mais antigos medem a quantidade de raios gama de captura após excitação artificial através de bombardeio dirigido de nêutrons rápidos. Os mais modernos medem a quantidade de nêutrons epitermais e/ou termais da rocha após bombardeio. São utilizados para estimativa de porosidade, litologia e detecção de hidrocarbonetos leves ou gás.

Modelo de Regressão Logística

O modelo proposto está representado pela equação (4). Como pode ser visto nesse modelo, houve a necessidade de aplicar uma transformação na variável ILD através da aplicação do logaritmo pois essa variável, originalmente, não apresenta um bom ajuste nas caudas do modelo. Em se tratando de geoestatística para análise espacial dos dados, os geólogos aplicam uma transformada para a variável ILD. A transformada aplicada utiliza-se da distribuição log-normal tri-paramétrica pelo fato de existir casos em que a distribuição log-normal bi-paramétrica não é simétrica e, por conseguinte, não é log-normal. Maiores detalhes sobre a distribuição log-normal tri-paramétrica podem ser obtidos através de Hill (1963), Hirose (1997), Wingo (1984), entre outros.

$$\begin{aligned}\Im \;=\; & \beta_0 + \beta_1 \times GR + \beta_2 \times \log(ILD) + \\ & + \beta_3 \times RHOB + \beta_4 \times Poço + \beta_5 \times Zona + \\ & + \beta_6 \times Poço \times Zona.\end{aligned} \tag{4}$$

Para uma análise das estimativas dos parâmetros do modelo proposto utilizamos a razão de chances.

Estando definido o modelo e sendo ele também o que apresenta menor $AIC = 1130$, o necessário agora é analisar os resíduos através dos métodos de diagnóstico para verificar a adequacidade do modelo proposto.

Segundo Hosmer e Lemeshow (1989, p. 157), uma conseqüência prática ao avaliar os pontos de alavanca na regressão logística é que para interpretar corretamente um valor particular de alavanca, nós precisamos saber se o valor estimado de probabilidade é menor que 0.1 ou maior que 0.9. Se a probabilidade estimada estiver entre 0.1 e 0.9 então a alavanca dará um valor que pode ser referido como distância. Assim, estaríamos utilizando da mesma consideração da regressão linear onde a alavanca é uma função monótona de incremento da distância da matriz de covariância padronizada para a média. Quando um estimador de probabilidade não está nos limites do intervalo $(0.1, 0.9)$, então o valor da alavanca não pode ser considerado medida de distância no sentido que isto implica altos valores.

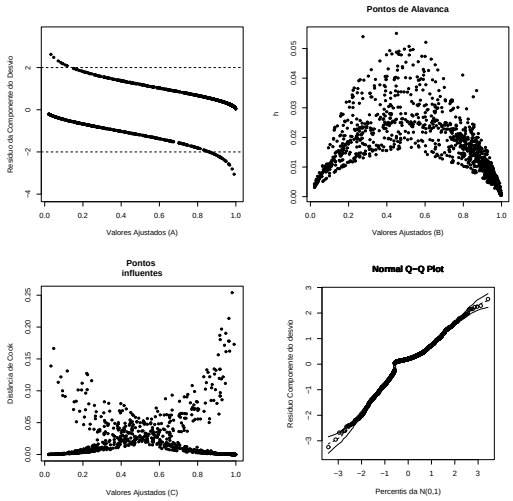

Figura 1: Análise dos resíduos do modelo ajustado.

As figuras de 1(A) a 1(C) apresentam alguns gráficos de diagnóstico considerados por Hosmer e Lemeshow (1989, pp. 160-161) como gráficos base para a análise de diagnóstico do modelo de regressão logística. Na figura 1(A) temos o gráfico referente aos resíduos do componente desvio padronizado com relação aos valores ajustados de forma a apresentar também a distribuição dos resíduos no intervalo $[-2; 2]$. Ao verificarmos o gráfico da figura 1(B), referente aos pontos de alavanca versus valores ajustados, observa-se a aparente existência de dispersão de alguns pontos. As situações consideradas na tabela 1 não correspondem a situação apresentada no gráfico quando analisamos o intervalo de $\hat{\pi}_i$. Hosmer e Lemeshow (1989, p. 161) consideram, para a análise do ponto de alavanca, o uso do valor crítico da distribuição qui-quadrado com 1 grau de liberdade ao nível de significância de 5%. Sendo assim, devido aos valores dos pontos de alavanca serem menores que o valor crítico de 3.84, então não há

como considerar presença de valores que afetem o ajuste do modelo. O gráfico da distância de Cook, referente à figura 1(C), tem como função descrever a influência de pontos no modelo. Assim, para a figura 1(C), não há indícios de pontos influentes por causa dos baixos valores encontrados ao calcular a distância de Cook.

Tabela 1: Possíveis valores para as medidas de diagnóstico $R^*_{P_i}$, $R^*_{D_i}$, LD_i e h_i com cinco regiões definidas segundo as probabilidades ajustadas.

Diagnóstico	Probabilidade Ajustada		
	0.0 - 0.1	0.1 - 0.3	0.3 - 0.7
$R^*_{P_i}$ ou $R^*_{D_i}$	Menor/Maior	Moderado	Moderado a menor
LD_i	Menor	Maior	Moderado
$\hat{h}_{ii}$	Menor	Maior	Moderado a menor
	0.7 - 0.9	0.9 - 1.0	
$R^*_{P_i}$ ou $R^*_{D_i}$	Moderado	Menor/Maior	
LD_i	Maior	Menor	
$\hat{h}_{ii}$	Maior	Menor	

Apesar das informações descritas pelos gráficos, ao verificar o desvio do modelo temos que este apresenta valor baixo ao relacioná-lo com os graus de liberdade. Ou seja, há indícios de subparametrização devido ao fato do desvio ser $D(\mathbf{y};\hat{\boldsymbol{\mu}}) = 1100$ e os graus de liberdade do modelo corresponderem a 1614. O gráfico normal de probabilidades com envelopes para o resíduo de componente do desvio (figura 1(D)) não apresenta indícios de problemas sérios da suposição de distribuição binomial para a variável resposta, pois a maioria dos pontos apresentam-se dentro do envelope.

Ao utilizarmos um ponto de corte em 50% para o modelo logístico observamos uma taxa de erro em torno de 15.47%. Na tabela 2 consta a matriz de confusão para o modelo logístico.

Tabela 2: Matriz de confusão para o modelo de regressão logística.

População Verdadeira	Classificação Realizada	
	0	1
0	318	112
1	138	1047

Toma-se como passo seguinte, avaliar o poder de classificação do modelo, com todas as observações presentes, através do método simples de classificação logístico e também através do método de curva ROC.

Segundo Louzada e Martinez (2000), quando o teste sob investigação produz uma resposta sob a forma de uma variável categórica ordinal ou contínua, emprega-se uma regra de decisão baseada em buscar um ponto de corte que resume tal quantidade em uma resposta dicotômica. Desta forma, para diferentes pontos de corte dentro da amplitude dos possíveis valores que o teste sob investigação produz, pode-se estimar sensibilidades e especificidades. Sabe-se ainda que a curva ROC tem como vantagem a avaliação de métodos de diagnóstico através do seu gráfico. Baseado nisto, deseja-se avaliar o desempenho do modelo logístico, considerado aqui como um teste de diagnóstico, de acordo com o conjunto de suas possíveis respostas através da sua representação visual direta. Na figura 2 está descrita a curva ROC para o modelo logístico proposto, onde se observa que a curva caracteriza uma boa descrição do poder de classificação do modelo. Através da regra dos trapézios, a área sob a curva ROC é estimada em 0.91 indicando que o modelo tem boa capacidade preditiva. A curva ROC não utiliza conjunto de teste e treinamento em suas análises de modelos.

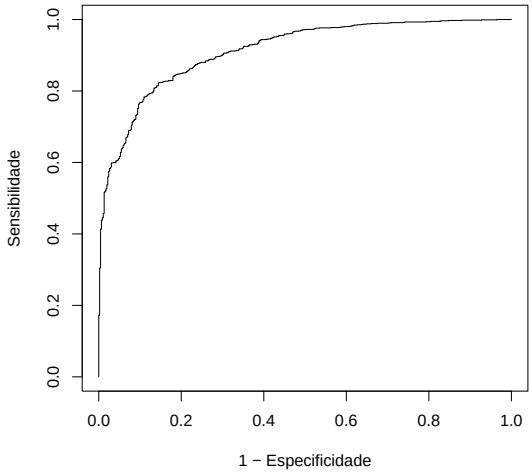

Figura 2: Curva ROC do modelo logístico proposto.

Tabela 3: Matriz de confusão para o modelo de regressão logística usando o conjunto de teste.

População Verdadeira	Classificação Realizada	
	0	1
0	116	53
1	37	332

Através de uma amostra de teste podemos estimar de forma honesta a taxa de erro. A amostra de teste consiste em dividir o conjunto de dados em

duas amostras independentes. Usa-se uma dessas amostras para obter a função de classificação e a amostra seguinte servirá como teste para estimação da taxa de erro. Sugere-se usar $\frac{2}{3}$ dos dados para treinamento e $\frac{1}{3}$ para teste. Assim, verifica-se a taxa de má-classificação através do uso de um conjunto de teste e um conjunto de treinamento para a formação do modelo, onde $\frac{1}{3}$ corresponde ao conjunto de teste. As amostras referentes ao conjunto de treinamento apresentam 303 observações correspondendo ao grupo não reservatório e 774 observações para o grupo reservatório. A tabela 3 descreve as classificações obtidas através do conjunto com 538 amostras de teste para o modelo proposto. A partir desse conjunto de teste obtivemos uma taxa de má-classificação em torno de 16.73% para o modelo logístico.

Modelo de Análise Discriminante

Utilizando o mesmo conjunto de teste obtido para a análise do modelo logístico, com probabilidade a priori para o grupos de forma proporcional, ou seja, grupo 1 com probabilidade aproximada 0.7176 e grupo zero com probabilidade aproximada 0.2823. A matriz de confusão referente ao modelo proposto (4) apresenta uma taxa de erro de 14.87%, ver tabela 4.

Tabela 4: Matriz de confusão para o modelo de análise discriminante baseado no modelo (4).

População Verdadeira	Classificação Realizada	
	0	1
0	114	39
1	41	344

Conclusões

Após tomarmos conhecimento das características das variáveis utilizadas para definição do modelo, através da análise do desvio, foi definido um melhor modelo contendo as variáveis *GR*, $\log(ILD)$, *RHOB*, *Poço* e *Zona*. O modelo proposto (4) foi abordado também na análise discriminante com o intuito de compararmos as taxas de má-classificação entre os dois modelos utilizados. A utilização da curva ROC objetivou avaliar o modelo de regressão logística. Dessa forma, observamos que ao utilizarmos a curva ROC, o modelo logístico apresentou boa capacidade preditiva pois a área sob a curva foi estimada em 0.91 pela regra dos trapézios. O uso da curva ROC vem propor então a utilização do ponto de corte definido por esta para utilizá-lo no modelo logístico ao invés de utilizar o corte de 50% da regessão logística para a classificação da variável fácies.

Para verificar a eficiência entre os métodos, observamos qual apresentava menor taxa de má-classificação das observações. Sendo assim, após obtermos as matrizes de confusão para regressão logística (tabela 3), cuja taxa de má-classificação corresponde a 16.73%, e análise discriminante (tabela 4), cuja taxa de má-classificação é de 14.87%, observamos que a análise discriminante apresenta menor taxa de má-classificação.

Segundo as considerações de Cox e Snell (1989), poderíamos então decidir que o melhor método para aplicação do conjunto de dados seria usar o modelo logístico devido a não necessidade de obtermos sub-populações. Mas dado os resultados obtidos, podemos concluir que os dois métodos se mostram eficientes quando modelado o conjunto de dados.

Referências

[1] B. M. Hill, "The three-parameter lognormal distribution and Bayesian analysis of a point-source-epidemic." *Journal of the American Statistical Association*, 58, n. 301 (1963) 72-84.

[2] D. R. Wingo "Fitting three-parameter lognormal models by numerical global optimization." *Computational Statistics & Data Analysis*, 2, n. 1 (1984) 13-15.

[3] F. Louzada-Neto e E. Z. Martinez, "Metodologia Estatística para testes diagnósticos e laboratoriais com respostas dicotomizadas." *Revista de Matemática e Estatística*, 18 (2000) 83-101.

[4] H. Hirose, "Maximum likelihood parameter estimation in the three-parameter log-normal distribution using the continuation method", *Computational Statistics & Data Analysis*, 24 (1997) 139-152.

[5] R. A. Fisher, "The use of multiple measurements in taxonomic problems", *Annals of Eugenics*, 7 (1936) 179-188.

[6] J. A. Anderson, "Logistic discrimination". (In: P. R. Krishnaiah and L. N. Kanal, eds), v.2, pp. 169-161, Handbook of Statistics, North Holland, Amsterdam, 1982.

[7] D. R. Cox e E. J. Snell, "Analysis of Binary Data", 2.ed, Chapman and Hall, UK, 1989.

[8] E. Vittinghoff, D. V. Glidden, S. C. Shiboski, C. E. McCulloch, "Regression Methods in Biostatistics: linear, logistic, survival, and repeated measures models", Springer Science+Business Media, Inc., New York, 2005.

[9] J. D. W. Hosmer e S. Lemeshow, "Applied Logistic Regression", John Wiley & Sons, New York, 1989.

[10] J. E. Thomas, A. A. Triggia, C. A. Correia, C. Verotti-Filho, J. A. D. XAvier, J. C. V. Machado, J. E. Sousa-Filho, J. L. de Paula, N. C. M. de Rossini, N. E. S. Pitombo, P. C. V. de M. Gouvea, R. de S. Carvalho, R. V. Barragam, "Fundamentos de Engenharia de Petróleo", Editora Interciência, Rio de Janeiro, 2001.

[11] R. A. Johnson e D.W. Wichern, "Applied multivariate statistical analysis", 3.ed, Englewood Cliffs: Prentice-Hall, pp. 642, 1992.