USO DE MODELOS ESTATÍSTICOS NA ANÁLISE DE DADOS DE RESERVATÓRIOS DE PETRÓLEO

Rejane dos Santos BRITO[1]
Getúlio José Amorim do AMARAL[2]

▪ RESUMO: Segue a proposta de um modelo que solucione o problema abordado em estudos geológicos referente a poços de petróleo. A fim 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. Dessa forma, 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.

1 Introdução

A análise deste trabalho baseia-se num conjunto de dados obtidos de poços de petróleo tal que, por meio da regressão logística e da análise discriminante, propõe-se um modelo preditivo para identificar tipos de rochas (litologias) favoráveis à 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") mediante a perfuração de poços, sendo estas caracterizadas como "rochas

[1]Departamento de Estatística e Informática, Universidade Federal Rural de Pernambuco – UFRPE, CEP: 52171-900, Recife, PE, Brasil. E-mail: *janesbrito@gmail.com*
[2]Departamento de Estatística, CCEN, Universidade Federal de Pernambuco – UFPE, CEP: 50740-540, Recife, PE, Brasil. E-mail: *gjaa@feller.de.ufpe.br*

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 as 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 adequação 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, serão utilizadas técnicas de diagnóstico para o 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.

2 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 mediante o uso da transformação logística $g(\pi)$, conhecida por *logit*, cujo parâmetro canônico é dado por

$$\eta = g(\pi) = \log[\pi/(1-\pi)], \tag{1}$$

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

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

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.

3 Análise discriminante

Considere M populações ou grupos $\pi_1, \ldots, \pi_M$, $M \geq 2$. Suponha que cada população π_i tem densidade de probabilidade $f_i(x)$ em R^p. O objetivo da análise discriminante é alocar um indivíduo a um destes grupos com base em suas medidas x.

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 análise discriminante é aplicada a duas populações, temos que a função de classificação se resume a

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

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 fácies não-reservatório e o grupo π_2 o representante da fácies 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.

4 Análise de dados em reservatório de petróleo

4.1 Introdução

Uma variável resposta e sete variáveis explicativas serão utilizadas na modelagem estatística e serão explicadas detalhadamente nos parágrafos seguintes.

A variável resposta é a variável *fácies* que se caracteriza pela subdivisão em fácies reservatório e não-reservatório, onde a fácies reservatório indica a presença de petróleo.

A variável *Poço* representa os três poços analizados, sendo estes nomeados como $Poço_1$, $Poço_2$ e $Poço_3$. A variável *Zona* segue uma ordem que depende da profundidade e esta ainda define as possíveis litologias a ocorrerem na mesma. Mais especificamente, temos que para cada poço existe a presença dos diferentes tipos de zona, e ainda, em cada zona há os característicos tipos de litologia. De modo geral, existem 11 tipos de zonas que descrevem a estrutura do poço, mas estas foram reagrupadas e, por conseqüinte, os fatores a serem utilizados para a análise estatística são $Zona_{1A}$, $Zona_{1B}$, $Zona_2$, $Zona_{3A}$, $Zona_{3B}$ e $Zona_4$.

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 por meio 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 disto, através da densidade é possível o cálculo da porosidade e a identificação das zonas de gás.

Uma outra variável definida como sônico (DT) mede a diferença nos tempos de trânsito de uma onda mecânica por meio 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 mediante bombardeio dirigido de nêutrons rápidos. Os mais modernos medem a quantidade de nêutrons epitermais ou termais da rocha após bombardeio. São utilizados para estimativa de porosidade, litologia e detecção de hidrocarbonetos leves ou gás.

4.2 Análise descritiva das variáveis

Antes de ajustar os modelos, foi realizada uma análise da estrutura de correlação entre as variáveis. Na Tabela 1 temos a matriz de correlação entre as variáveis do modelo. A matriz indica a presença de uma forte correlação de algumas das variáveis explicativas entre si.

Na tabela 2 temos a análise descritiva das variáveis analisadas em cada poço e a análise conjunta destes. Ao particionar os poços a fim de analisá-los individualmente, observa-se que estes apresentam as mesmas características entre as variáveis.

Depois de alguns modelos iniciais, verificou-se a necessidade de aplicar uma transformação na variável *ILD*. A transformação segue por meio da aplicação do logaritmo devido ao fato de originalmente essa variável não realizar 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 mediante Hill (1963), Hirose (1997), Wingo (1984), entre outros.

Tabela 1 - Matriz de correlação entre as variáveis propostas para definição do modelo

Covariáveis	GR	DT	log (ILD)	RHOB	NPHI
GR	1.00				
DT	0.81	1.00			
log (ILD)	−0.74	−0.61	1.00		
RHOB	−0.39	−0.76	0.20	1.00	
NPHI	0.84	0.94	−0.62	−0.71	1.00

4.3 Modelo de regressão logística

Na tabela 3 temos a descrição da análise de desvio por meio do processo de definição de modelos significativos obtidos a cada adição, sendo o modelo final apresentado pela estrutura mostrada na equação (4). As estimativas dos parâmetros deste modelo adotado podem ser vistas na tabela 4.

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

Tabela 2 - Análise descritiva dos poços de maneira geral e individual

	Variável	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo	D. Padrão	Assimetria	Curtose
	GR	30.300	47.900	64.400	70.560	89.950	141.800	25.874	0.574	2.257
	DT	59.500	74.550	84.800	85.430	95.500	118.600	12.669	0.203	2.063
Geral	log (ILD)	0.000	1.131	1.758	1.871	2.510	5.011	0.957	0.442	2.834
	RHOB	2.130	2.380	2.430	2.428	2.480	2.630	0.078	−0.417	3.107
	NPHI	8.100	19.700	25.300	25.240	30.700	39.500	6.507	−0.033	2.091
	GR	34.600	52.600	70.000	72.880	88.900	135.800	23.909	0.477	2.361
	DT	59.500	79.600	91.100	89.870	100.000	118.600	12.619	−0.172	2.212
Poço 1	log (ILD)	0.000	1.030	1.629	1.846	2.532	5.011	1.102	0.720	3.201
	RHOB	2.130	2.333	2.390	2.392	2.450	2.600	0.082	−0.119	2.840
	NPHI	8.100	22.000	27.100	26.340	31.200	39.500	6.191	−0.409	2.601
	GR	31.400	45.500	62.600	70.450	94.150	141.800	27.693	0.545	2.051
	DT	59.700	72.250	79.300	81.090	90.100	104.600	10.552	0.203	1.868
Poço 2	log (ILD)	0.531	1.224	1.841	1.859	2.402	3.721	0.761	0.341	2.312
	RHOB	2.310	2.420	2.470	2.459	2.500	2.630	0.060	−0.233	2.685
	NPHI	9.900	19.200	23.700	24.330	29.400	38.800	6.206	0.135	2.050
	GR	30.300	47.900	60.700	68.690	87.920	140.200	25.111	0.716	2.498
	DT	61.200	75.030	85.550	86.840	98.400	115.200	13.405	0.191	1.864
Poço 3	log (ILD)	0.000	1.099	1.758	1.906	2.771	3.910	1.033	0.151	2.061
	RHOB	2.150	2.370	2.420	2.422	2.470	2.610	0.079	−0.291	3.126
	NPHI	10.200	19.520	25.250	25.410	31.380	39.200	6.965	0.037	1.938

Ao verificar as estimativas da razão de chances para cada efeito das variáveis na tabela 4, tem-se que os valores estimados das razões de chances $\hat{\Psi}(constante)$, $\hat{\Psi}(log(ILD))$, da associação conjunta de $\hat{\Psi}(Zona_{1A},\ Zona_{1B})$, $\hat{\Psi}(Zona_{1A},\ Zona_2)$, $\hat{\Psi}(Zona_{1A},\ Zona_{3A})$, $\hat{\Psi}(Zona_{1A},\ Zona_{3B})$, $\hat{\Psi}(Zona_{1A},\ Zona_4)$ e da associação conjunta das interações do modelo referentes a $\hat{\Psi}(Poço_2 \times Zona_2)$, $\hat{\Psi}(Poço_2 \times Zona_{3A})$, $\hat{\Psi}(Poço_2 \times Zona_{3B})$, $\hat{\Psi}(Poço_2 \times Zona_4)$ apresentam uma forte contribuição na relação desses efeitos com a variável resposta, aqui referida como fácies reservatório. Ou seja, indicam de forma particular a chance de se definir fácies reservatório quando analisada uma variável, em particular, e considerada as demais variáveis como constantes. É observado ainda uma fraca, mas ainda positiva, associação dos efeitos das variáveis por meio das seguintes razões de chances: $\hat{\Psi}(Poço_1, Poço_3)$, $\hat{\Psi}(Zona_{1A}, Zona_{3B})$ e $\hat{\Psi}(Poço_3 \times Zona_{3B})$. Quanto a estimativa da razão de chances para os efeitos das variáveis referentes às $\hat{\Psi}(RHOB)$, $\hat{\Psi}(Poço_3 \times Zona_{1B})$, $\hat{\Psi}(Poço_3 \times Zona_2)$, $\hat{\Psi}(Poço_3 \times Zona_{3A})$ e $\hat{\Psi}(Poço_3 \times Zona_4)$, tem-se que estas apresentam uma contribuição na variável resposta correspondente a um fator protetor, ou seja, uma associação negativa de contribuição da variável.

Finalmente, temos que para as razões de chances $\hat{\Psi}(GR)$, $\hat{\Psi}(Poço_1, Poço_2)$, $\hat{\Psi}(Poço_2 \times Zona_{1B})$, a associação com a variável resposta quase não existe, pois lembremos que, quando a razão de chances é um, pode-se considerar que a associação com a variável resposta seja nula.

Tabela 3 - Modelo obtido após realização da análise de desvio para definição das variáveis presentes no mesmo

Modelo	Desvio	Diferença	g.l.	p-valor	Testando
Constante	1922.37	-	-	-	-
+ GR	1239.42	682.95	1	$1.526\,\exp^{-150}$	GR
+ log(ILD)	1178.63	60.82	1	$6.332\,\exp^{-15}$	log (ILD)| GR
+ Zona	1110.52	68.11	5	$2.531\,\exp^{-13}$	Zona| GR + log (ILD)
+ Poço × Zona	1080.25	30.27	12	0.002545	Poço×Zona| GR + log (ILD) +Poço + Zona
+ RHOB	1071.39	8.86	1	0.002912	Poço × Zona | GR + log (ILD) + Poço +Zona + RHOB

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

Tabela 4 - Estimativas dos parâmetros e respectivos desvio padrão e razão de chances referentes ao modelo logístico com efeitos principais para explicar a ocorrência de fácies reservatório

Efeito	Estimativa	D. Padrão	Razão de Chances $(\hat{\Psi})$
Constante	11.443539(3.296)	3.472115	9.329663×10^4
GR	$-0.050913(-7.903)$	0.006442	0.950361
log *(ILD)*	1.102887(5.844)	0.188718	3.012853
RHOB	$-3.832972(-2.978)$	1.287028	2.164518×10^{-2}
$Poço_2$	$-0.016636(-0.042)$	0.400497	0.983501
$Poço_3$	0.500014(1.356)	0.368781	1.648745
$Zona_{1B}$	0.776408(1.780)	0.436123	2.173650
$Zona_2$	1.009750(2.359)	0.428093	2.744915
$Zona_{3A}$	1.179471(2.262)	0.521477	3.252653
$Zona_{3B}$	0.446934(1.023)	0.436856	1.563511
$Zona_4$	1.379191(2.843)	0.485044	3.971686
$Poço_2 \times Zona_{1B}$	0.070685(0.120)	0.591231	1.073243
$Poço_3 \times Zona_{1B}$	$-0.864786(-1.460)$	0.592236	0.421142
$Poço_2 \times Zona_2$	1.254208(2.126)	0.589878	3.505062
$Poço_3 \times Zona_2$	$-0.810829(-1.341)$	0.604504	0.444897
$Poço_2 \times Zona_{3A}$	0.898854(1.243)	0.723076	2.456778
$Poço_3 \times Zona_{3A}$	$-0.286676(-0.407)$	0.704204	0.750755
$Poço_2 \times Zona_{3B}$	1.285961(2.040)	0.630259	3.618142
$Poço_3 \times Zona_{3B}$	0.490592(0.796)	0.637715	1.633282
$Poço_2 \times Zona_4$	1.202030(1.866)	0.644263	3.326865
$Poço_3 \times Zona_4$	$-1.034729(-1.679)$	0.616433	0.355323

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, é preciso 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, estaria sendo utilizada a 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.

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

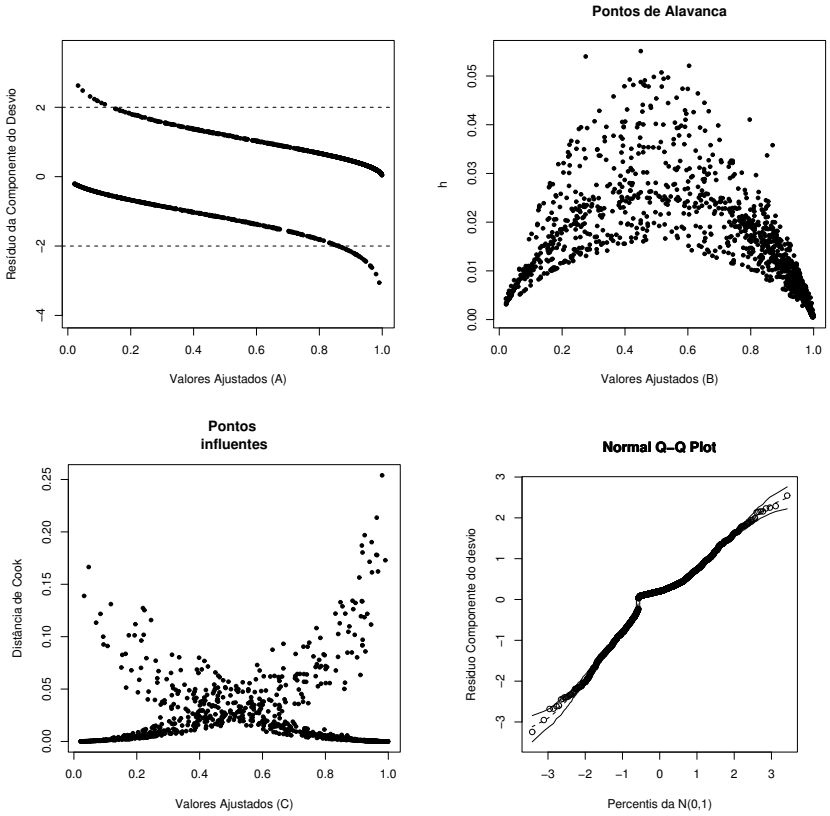

Figura 1 - Análise dos resíduos do modelo ajustado.

a análise de diagnóstico do modelo de regressão logística. Na figura 1(A) tem-se 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 verificar o gráfico da figura 1(B), referente aos pontos de alavanca versus valores ajustados, observa-se a aparente existência de dispersão de alguns pontos.

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.

Apesar das informações descritas pelos gráficos, ao verificar o desvio do modelo é observado 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.

Tabela 5 - 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	0.7 - 0.9	0.9 - 1.0
$R^*_{P_i}$ ou $R^*_{D_i}$	Menor/Maior	Moderado	Moderado a menor	Moderado	Menor/Maior
LD_i	Menor	Maior	Moderado	Maior	Menor
$\hat{h}_{ii}$	Menor	Maior	Moderado a menor	Maior	Menor

Na tabela 5 são apresentadas as medidas de diagnóstico: resíduo de Pearson padronizado($R^*_{P_i}$), resíduo do componente desvio padronizado ($R^*_{D_i}$), distância de Cook (LD_i) e alavancagem ($\hat{h_{ii}}$), relacionando estes com as probabilidades ajustadas. As situações consideradas na tabela 5 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 um grau de liberdade ao nível de significância de 95%. 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.

Quando utilizado um ponto de corte em 50% para o modelo logístico é observado uma taxa de erro em torno de 15.47%. Na tabela 6 consta a matriz de confusão para o modelo logístico.

Tabela 6 - 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, mediante o método simples de classificação logístico e também por meio 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

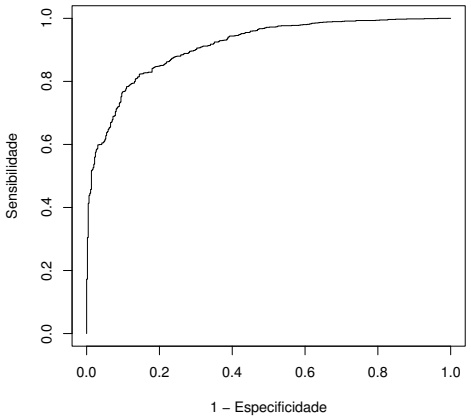

Figura 2 - Curva ROC do modelo logístico proposto.

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 por meio 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 por meio 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. Mediante a 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.

Por meio de uma amostra de teste é possível 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.

Tabela 7 - 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

Assim, verifica-se a taxa de má-classificação por meio 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 7 descreve as classificações obtidas por meio do conjunto com 538 amostras de teste para o modelo proposto. A partir desse conjunto de teste foi obtida uma taxa de má-classificação em torno de 16.73% para o modelo logístico.

4.4 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 8.

Tabela 8 - 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

A média das variáveis por grupo para o modelo proposto pela equação (4) é dada pela tabela 9. Ao verificar os coeficientes estimados para o modelo representado pela equação 4, observa-se que os coeficientes das variáveis $log(ILD)$, $Poço_3$, $Zona_2$, $Zona_4$, $Poço_2 \times Zona_{1B}$, $Poço_3 \times Zona_{1B}$, $Poço_2 \times Zona_2$, $Poço_2 \times Zona_{3A}$, $Poço_3 \times Zona_{3A}$, $Poço_2 \times Zona_{3B}$, $Poço_3 \times Zona_{3B}$ e $Poço_2 \times Zona_4$ apresentam maior ênfase em suas características, ou seja, auxiliam de forma positiva na classificação dos grupos no modelo de análise discriminante.

Conclusões

Após tomar conhecimento das características das variáveis utilizadas para definição do modelo, por meio da análise do desvio, foi definido como o melhor modelo o que continha 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.

Tabela 9 - Valores médios das variáveis por grupo no modelo discriminante linear e a estimativa dos coeficientes

Variável	Média dos grupos		Coeficiente
	Não Reservatório	Reservatório	
GR	96.234320	60.726490	−0.040467
$\log(ILD)$	1.056693	2.183419	0.349664
$RHOB$	2.411254	2.433346	−3.734885
$Poço_2$	0.339934	0.396641	−0.363749
$Poço_3$	0.349835	0.329457	0.170243
$Zona_{1B}$	0.290429	0.080103	−0.048403
$Zona_2$	0.161716	0.229974	0.232578
$Zona_{3A}$	0.085809	0.157623	0.070476
$Zona_{3B}$	0.125412	0.143411	−0.285982
$Zona_4$	0.108911	0.236434	0.173034
$Poço_2 \times Zona_{1B}$	0.105611	0.025840	0.406046
$Poço_3 \times Zona_{1B}$	0.105611	0.024548	−0.523776
$Poço_2 \times Zona_2$	0.049505	0.073643	0.977192
$Poço_3 \times Zona_2$	0.052805	0.085271	−0.407053
$Poço_2 \times Zona_{3A}$	0.019802	0.062015	1.030522
$Poço_3 \times Zona_{3A}$	0.029703	0.046512	0.211381
$Poço_2 \times Zona_{3B}$	0.033003	0.041344	1.255827
$Poço_3 \times Zona_{3B}$	0.039604	0.058139	0.191865
$Poço_2 \times Zona_4$	0.026403	0.152455	0.934092
$Poço_3 \times Zona_4$	0.049505	0.055555	−0.318819

A utilização da curva ROC objetivou avaliar o modelo de regressão logística. Dessa forma, observa-se que ao utilizar 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 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, observou-se qual apresentava menor taxa de má-classificação das observações. Sendo assim, após obter as matrizes de confusão para regressão logística (tabela 7), cuja taxa de má-classificação corresponde a 16.73%, e análise discriminante (tabela 8), cuja taxa de má-classificação é de 14.87%, observou-se que a análise discriminante apresenta menor taxa de má-classificação.

Segundo as considerações de Cox e Snell (1989), seria possível 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, conclui-se que os dois métodos se mostram eficientes quando modelado o conjunto de dados.

Agradecimentos

À FACEPE, ao CNPQ, ao professor Gauss Moutinho Cordeiro, a Marcelo Hardman, Vitor Hugo Simon e aos revisores.

BRITO, R. S.; AMARAL, G. J. A. Use of statistical models to analyse data from oil reservoir. *Rev. Bras. Biom.,* São Paulo, v.25, n.3, p.93-107, 2007.

- *ABSTRACT: The aim of this work, from the point of view of statistical modeling, is to analyze a dataset through the logistic regression and to propose a new model that solves a problem related to geological studies regarding oil wells. The trial for obtaining the data aims to realize the evaluation of formations, from which are defined the productive capacity and the appraisal for his backups of oil and gas. The nature of the problem lead us also to evaluate the power of classification for the proposed model with regard to the two definite groups: reservoir facies and not-reservoir facies.*

- *KEYWORDS: Residuals analysis; ROC curves; classification's models.*

Referências

ANDERSON, J. A. Logistic discrimination. In: KRISHNAIAH, P. R.; KANAL, L. N. (Ed.). *Handbook of statistics*. Amsterdam, 1982.

COX, D. R.; SNELL, E. J. *Analysis of binary data*. $2^{nd.}$ *ed.*, Chapman and Hall, 1989.

FISHER, R. A. The use of multiple measurements in taxonomic problems. *Ann. Eugen.*, London, v.7, p.179-188, 1936.

HILL, B. M. The three-parameter lognormal distribution and Bayesian analysis of a point-source-epidemic. *J. Am. Stat. Assoc.*, v.58, n.301, p.72-84, 1963.

HIROSE, H. Maximum likelihood parameter estimation in the three-parameter log-normal distribution using the continuation method. *Comput. Stat. Data Anal.*, Amsterdan, v.24, pp. 139-152, 1997.

HOSMER, J. D. W.; LEMESHOW, S. *Applied logistic regression*. New York: John Wiley & Sons, 1989.

JOHNSON, R. A.; WICHERN, D.W. Applied multivariate statistical analysis. 3.ed. Englewood Cliffs: Prentice-Hall, 1992. 642p.

LOUZADA-NETO, F.; MARTINEZ, E. Z. Metodologia Estatística para testes diagnósticos e laboratoriais com respostas dicotomizadas. *Rev. Mat. Estat.*, São Paulo, v.18, p.83-101, 2000.

THOMAS, J. E.et al. *Fundamentos de engenharia de petróleo*. Rio de Janeiro: Editora Interciência, 2001.

VITTINGHOFF, E.; GLIDDEN, D. V.; SHIBOSKI, S. C.; McCULLOCH, C. E. *Regression methods in biostatistics: linear, logistic, survival, and repeated measures models*. New York: Springer Science+Business Media, 2005.

WINGO, D. R. Fitting three-parameter lognormal models by numerical global optimization. Comput. *Stat. Data Anal.*, Amsterdam, v.2, n.1, p.13-15, 1984.

Recebido em 29.07.2007.

Aprovado após revisão em 21.12.2007.