Regressão Beta para modelagem de taxas e proporções

Beatriz Ariadna da Silva Ciríaco[1], Wesley Almeida Cruz[1], and Eliardo Guimarães da Costa[1]

[1] *Departamento de Estatística, Universidade Federal do Rio Grande do Norte*

Resumo

Esse trabalho discute o modelo de regressão Beta para modelar variáveis númericas que pertençam ao intervalo $(0, 1)$. Comentamos sobre a necessidade de usar esse modelo em comparação à uma regressão linear com variáveis transformadas, bem como mencionamos as propriedades da regressão Beta, e os cálculos necessários para conduzir a regressão e sua análise de diagnóstico. Ao final, é apresentado uma aplicação desse modelo em dados reais.

: Distribuição Beta, Modelo de Regressão Beta, Método da Máxima verossimilhança, Algorítmos de otimização

1 Introdução

Em uma situação em que se deseja aplicar uma análise de regressão para uma variável resposta que está contida no intervalo $(0, 1)$, é comum utilizar uma transformação da variável original afim de contornar o problema da variável resposta não estar contida na reta Real ($\mathbb{R}$) e, assim tornando obsoleta a regressão linear simples, por exemplo. Uma transformação bastante comum é a função logito dada por $\tilde{y} = \log(y/(1-y))$.

Apesar de funções como a logito ter sucesso em mapear uma variável resposta na forma $(0, 1) \longrightarrow \mathbb{R}$, ainda assim possuem três defeitos. Primeiramente, a interpretação da regressão está em relação em média da variável transforamda e não da original, tornando difícil o entendimento das suas conclusões. Além disso, regressões desse tipo tem, geralmente, erros heterocedásticos, mostrando maior variação em torno da média e menos nas extremidades. Finalmente, a distribuição de taxas e proporções normalmente são assimétricas e assim a condição de normalidade de Gauss-Markov provavelmente será violada.

Dessa forma, uma alternativa menos danosa foi proposta por Ferrari e Cribari-Neto (2004) para a modelagem de variáveis aleatórias contínuas que assumem valores em um determinado intervalo (a, b) de tal forma que $a < b; a, b \in \mathbb{R}$, como por exemplo o de taxas e proporções. Nesse modelo, assumimos que a variável resposta segue uma distribuição Beta e esse modelo é conhecido como **Modelo de Regressão Beta**. A interpretação dos parâmetros no modelo proposto é em termos da média da variável resposta e o modelo é naturalmente heterocedástico e é robusto a dados assimétricos. Devido à flexibilidade natural da distribuição beta e dos diversos formatos que a sua densidade pode assumir essa distribuição de probabilidade é recomendada para a modelagem de proporções.

Nesse trabalho, será apresentado a distribuição de probabilidade Beta, o Modelo de Regressão Beta, suas propriedades, análise de diagnóstico e uma aplicação utilizando o pacote *betareg* do *Software RStudio*.

2 Modelo de probabilidade Beta

Seja X uma variável aleatória contínua que pertença a família de distribuição Beta com parâmetros de forma α e β ($X \sim \text{Beta}(\alpha, \beta)$), então a sua função densidade de probabilidade é dada

por:

$$f(x;\alpha,\beta) = \frac{1}{\mathrm{B}(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}\, I_{(0,1)}(x), \quad \alpha,\beta > 0, \tag{1}$$

em que $\mathrm{B}(\alpha,\beta) = \dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ e $\Gamma(\cdot)$ é a função gama.

Uma maneira de encontrar o k-ésimo momento da distribuição Beta é através da Função Geradora de Momentos que, nesse caso, é dado por

$$M_X(\alpha,\beta;t) = 1 + \sum_{k=1}^{\infty}\left(\prod_{r=0}^{k-1}\frac{\alpha+r}{\alpha+\beta+r}\right)\frac{t^k}{k!} \tag{2}$$

Dada esta função em (2), o k-ésimo momento pode ser obtido da seguinte forma

$$\mathbb{E}(X^k) = \left.\frac{\partial^k M_X(t)}{\partial t^k}\right|_{t=0} = \frac{\Gamma(\alpha+\beta)\Gamma(\alpha+k)}{\Gamma(\alpha+\beta+k)\Gamma(\alpha)}, \quad k = 1,2,3,\ldots. \tag{3}$$

A partir de (3), é trivial derivar às seguintes propriedades da distribuição Beta que serão úteis posteriormente. São elas:

$$\mathbb{E}(X) = \frac{\alpha}{\alpha+\beta} \tag{4}$$

e

$$\mathrm{Var}(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{5}$$

A família de distribuição Beta é bastante versátil e uma gama de situações podem ser modeladas por essa distribuição com sucesso. A flexibilidade natural deste modelo de probabilidade permite seu uso em diversas aplicações como em Johnson et al. (1995) e Bury (1999) . Entretanto, essas aplicações não incluem situações em que o pesquisador esteja interessado em conduzir uma análise de regressão para modelar o comportamento de uma variável resposta em relação a covariáveis exógenas.

3 Modelo de Regressão Beta

Ferrari e Cribari-Neto (2004) propuseram uma alternativa para conduzir uma análise de regressão em dados contínuos no intervalo $(0,1)$ utilizando a distribuição de probabilidade Beta. A modelagem e os procedimentos inferenciais propostos por Ferrari e Cribari-Neto (2004) são semelhantes aos modelos lineares generalizados (McCullagh Nelder, 1989), porém no caso descrito, a variável resposta não pertence a família exponencial. Uma alternativa ao modelo de regressão Beta é o modelo Simplex proposto por Jørgensen (1997) que possui quatro parâmetros. Em contra partida, modelo de regressão Beta possui somente dois parâmetros e é flexível o suficiente para abranger diversos cenários.

Nesse trabalho, o objetivo é aplicar uma análise de regressão em uma variável resposta Y que assume valores no intervalo $(0,1)$ com o modelo de regressão Beta. Porém, é possível generalizar para intervalos $[a,b]$, com $a < b$ conhecidos, utilizando transformações convenientes, como $\frac{Y-a}{b-a}$. Além disso, se a variável resposta assume valores nas extremidades 0 e 1, uma transformação útil é $\frac{(Y(n1)+0.5)}{n}$, em que n é o tamanho da amostra, como pode ser visto em Smithson e Verkuilen (2006) e Cribari-Neto e Zeileis (2010).

A função densidade de probabilidade da distribuição Beta é dada pela equação (1), indexada

por α e β. Contudo, em um contexto de regressão é útil modelar a média da variável resposta Y. Além disso, também é típico que o modelo de regressão possua um parâmetro de dispersão (ou precisão). Para obter essa estrutura na modelagem, deve-se reparametrizar a distribuição Beta de forma que os seus parâmetros se tornem parâmetros de média e dispersão. Considerando $Y \sim \text{Beta}(\alpha, \beta), \mu = \frac{\alpha}{\alpha+\beta}$ e $\phi = \alpha+\beta$, a reparametrização resulta na nova distribuição Beta$(\mu\phi, (1-\mu)\phi)$, em que μ é o parâmetro de média e ϕ o parâmetro de dispersão que serão utilizados para a modelagem de regressão. De acordo com a nova reparametrização, a densidade da variável aleatória Y é dada por:

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} I_{(0,1)}(y), \tag{6}$$

em que $0 < \mu < 1$ e $\phi > 0$. Segue das equações (4) e (5):

$$\mathbb{E}(Y) = \mu \tag{7}$$

e

$$\text{Var}(Y) = \frac{V(\mu)}{1+\phi}, \tag{8}$$

em que $V(\mu) = \mu(1-\mu)$.

Os gráficos abaixo fazem referência ao comportamento da função de densidade de probabilidade da distribuição Beta reparametrizada ao fixar o parâmetro de dispersão ϕ e variando o parâmetro média μ, nas situações em que $\phi = \{5, 100\}$ e $\mu = \{0.1, 0.25, 0.5, 0.75, 0.9\}$.

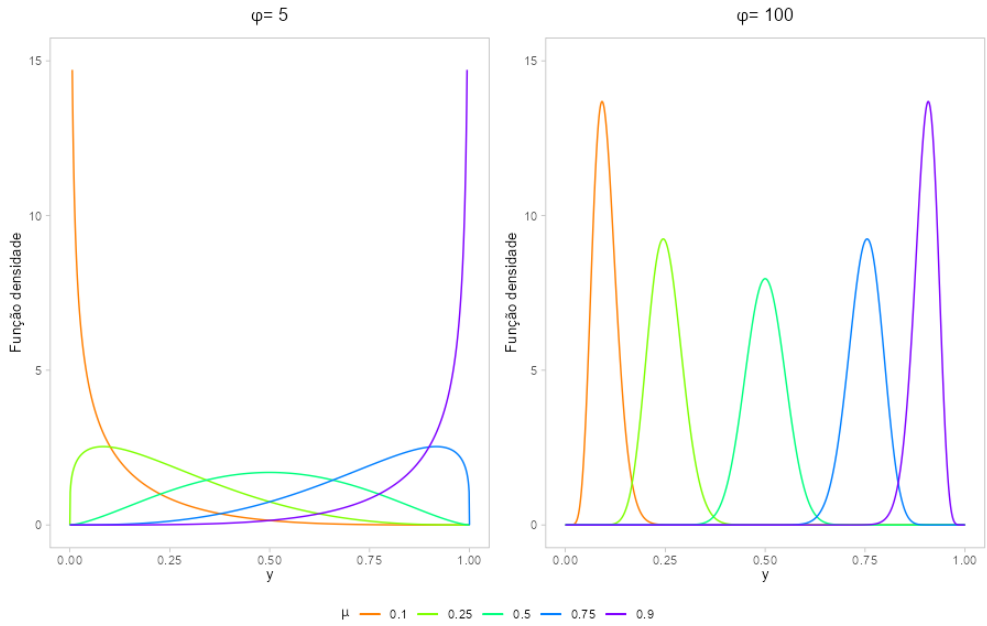

Figura 1. Comportamento da função densidade da Beta com diversas combinações de (μ, ϕ).

É possível notar na Figura 1 que uma variável aleatória Beta apresenta diversos formatos distintos conforme as combinações dos parâmetros, em particular quando $\mu = 0.5$ a distribuição se torna simétrica independente do valor de ϕ. Além do mais, com o parâmetro de média μ fixado é perceptível que a dispersão da distribuição diminui a medida que ϕ aumenta.

Seja $Y_1, \ldots, Y_n$ variáveis aleatórias independentes, em que Y_i, $i = 1, \ldots, n$, que seguem a

função densidade de probabilidade dada na equação (6) com média μ_i e parâmetro de precisão ϕ desconhecido. Assumindo que a seguinte função de ligação seja adequada para modelar a média de Y_i,

$$g(\mu_i) = \sum_{t=1}^{k} x_{it}\beta_t = \eta_i, \tag{9}$$

obtemos assim o modelo, em que $\beta = (\beta_1, \ldots, \beta_k)^T$ é o vetor dos parâmetros da regressão e $x_{i1}, \ldots, x_{ik}$ são as observações das k covariáveis $(k < n)$, que assumimos ser fixos e conhecidos.

A função de ligação $g(\cdot)$ que mapeia $(0, 1) \longrightarrow \mathbb{R}$ é estritamente monótona e diferenciável até a segunda ordem. Existem outras possíveis funções de ligação, como por exemplo a função log-log $g(\mu) = -\log[-\log(\mu)]$, a função complemento log-log $g(\mu) = \log[-\log(1-\mu)]$, a função probito $g(\mu) = \Phi^{-1}(\mu)$, em que $\Phi(\cdot)$ é a função de distribuição acumulada da variável aleatória Normal Padrão, entre outras. Para mais transformações, ver Atkinson (1985).

Uma função de ligação particularmente útil é a função logito $g(\mu) = \log[\mu(1-\mu)]$, em que é possível escrever μ_i da seguinte forma:

$$\mu_k = \frac{e^{x_k^T \beta_k}}{1 + e^{x_k^T \beta_k}}. \tag{10}$$

Dessa forma, a interpretação dos parâmetros da regressão pode ser entendida como, ao aumentar c unidades na k-ésima covariável sem que haja alterações nas demais covariáveis, e seja μ_τ a média de Y restrito as novas covariáveis, denote μ a média de Y com as covariáveis originais. É trival, notar que a razão de chances nessa situação é dada por:

$$e^{c\beta_k} = \frac{\mu_\tau(1-\mu_\tau)}{\mu(1-\mu)}. \tag{11}$$

A função de log-verossimilhança é dada pela seguinte expressão:

$$\ell(\beta, \phi) = \sum_{i=1}^{n} \ell_i(\mu_i, \phi) \tag{12}$$

em que,

$$\begin{aligned} \ell_i(\mu_i, \phi) &= \log\Gamma(\phi) - \log\Gamma(\mu_i\phi) - \log\Gamma[(1-\mu_i)\phi] + (\mu_i\phi - 1)\log(y_i) \\ &+ [(1-\mu_i)\phi - 1]\log(1-y_i). \end{aligned}$$

Com μ_i definido na equação (9). Seja $y_i^* = log[y_i/(1-y_i)]$ e $\mu_i^* = \psi(\mu_i\phi) - \psi[(1-\mu_i)\phi]$. Então a função escore pode ser obtida pela derivação da função de log-verossimilhança com respeito à parâmetros desconhecidos e são dadas por, respectivamente:

$$U_\beta(\beta, \phi) = \phi X^T T(y^* - \mu^*),$$

$$U_\phi(\beta, \phi) = \sum_{i=1}^{n} \{\mu_i(y_i^* - \mu_i^*) + \log(1-y_i) - \psi[(1-\mu_i)\phi] + \psi(\phi)\}.$$

Sendo X uma matriz $n \times k$, em que a i-ésima linha é x_i^T, $T = \text{diag}\{1/g'(\mu_1), \ldots, 1/g'(\mu_n)\}$ e y^* e μ^* são vetores de y_i^* e μ_i^*, definidos anteriormente

Para obter a expressão da informação de Fisher, vamos considerar $W = \text{diag}(w_1, \ldots, w_n)$, com

$$w_i = \frac{c_i}{(g'\mu_i)^2}.$$

em que $c_i = \phi\{\psi'(\mu_i\phi_i) + \psi'[(1-\mu_i)\phi]\}$ e $c = (c_1, \ldots, c_n)^T$, $\psi'(\cdot)$ é a função trigama. Seja $D = \text{diag}(d_1, \ldots, d_n)$, com $d_i = \psi'(\mu_i\phi)\mu_i^2 + \psi'[(1-\mu_i)\phi](1-\mu_i^2) - \psi'(\phi)$. De acordo com Ferrari e Cribari-Neto (2004), a matriz de informação de Fisher é dada por:

$$\mathbb{K} = \mathbb{K}(\beta, \phi) = \begin{pmatrix} \mathbb{K}_{\beta\beta} & \mathbb{K}_{\beta\phi} \\ \mathbb{K}_{\phi\beta} & \mathbb{K}_{\phi\phi} \end{pmatrix} \tag{13}$$

Sendo $\mathbb{K}_{\beta\beta} = \phi X^T W X$, $\mathbb{K}_{\beta\phi} = \mathbb{K}_{\phi\beta}^T = X^T T c$ e $\mathbb{K}_{\phi\phi} = \text{Tr}(D)$.

Os estimadores de máxima verossimilhança $\hat{\beta}$ e $\hat{\phi}$ para β e ϕ são obtidos quando $U_\beta(\beta, \phi) = 0$ e $U_\phi(\beta, \phi) = 0$, respectivamente. Esses estimadores não possuem forma fechada e devem ser encontrados por meio de algorítmos númericos iterativos, como por exemplo o Newton–Raphson ou Quasi-Newton mencionados em Nocedal e Wright (1999). No processo de otimização destes algorítmos é necessário um valor inicial para a estrutura iterativa, a literatura sugere o uso da estimava por mínimos quadrados ordinários para β obtido por uma regressão linear de variáveis transformadas $g(Y_1), \ldots, g(Y_n)$ como ponto inicial. Também é necessário um valor inicial para ϕ, usando como base a expressão em (8), o ϕ pode ser calculado em função de μ_i e y_i da seguinte forma, $\phi = \mu_i(1-\mu_i)/\text{Var}(Y_i) - 1$. Note que,

$$\text{Var}(g(Y_i)) \approx \text{Var}\{g(\mu_i) + (Y_i - \mu_i)g'(\mu_i)\} = \text{Var}(Y_i)[g'(\mu_i)]^2,$$

Assim, o valor inicial sugerido para ϕ é,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\mu_i^{**}(1-\mu_i^{**})}{\sigma_i^{**2}} - 1,$$

em que μ_i^{**} é obtido aplicando $g^{-1}(\cdot)$ na i-ésima predição da regressão linear das variáveis transformadas $g(Y_1), \ldots, g(Y_n)$ em X, $\mu_i^{**} = g^{-1}(x_i^T(X^TX)^{-1}X^Tz)$ e $\sigma_i^{**} = e^{**T}e^{**}/[(n-k)g'(\mu^{**})^2]$, em que $e^{**} = z - X(X^TX)^{-1}X^Tz$ é o vetor dos resíduos da regressão por mínimos quadrados ordinários das variáveis resposta transformadas.

4 Análise de diagnóstico

Após ajustar o modelo de regressão Beta, é importante realizar uma análise de diagnóstico para checar o ajuste. Uma medida de explicação da variabilidade pode ser obtido calculando o pseudo $-R^2(R_p^2)$ definido como $\text{Cor}(\hat{\eta}, g(y))^2$. A medida (R_p^2) varia entre 0 e 1 e a concordância perfeita ocorre quando $(R_p^2) = 1$.

A discrepância do ajuste ($\text{D}(y, \mu, \phi)$) pode ser medida utilizando a seguinte expressão:

$$\text{D}(y, \mu, \phi) = \sum_{i=1}^{n} (r_i^d)^2,$$

em que $r_i^d = \text{sign}(y_i - \hat{\mu}_i)[2\ell_i(\tilde{\mu}_i, \hat{\phi}_i)\ell_i(\hat{\mu}_i, \hat{\phi}_i)]^{1/2}$. Note que a i-ésima observação contribui para $(r_i^d)^2$ no desvio e assim observações com valores nominais altos podem ser vistos como discrepantes.

Também é possível definir os resíduos padrões da seguinte forma:

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\text{Var}}(y_i)}}$$

sendo $\hat{\mu}_i = g^{-1}(x_i^T\hat{\beta})$ e $\hat{\text{Var}}(y_i) = [\hat{\mu}_i(1-\hat{\mu}_i)]/(1+\hat{\phi})$. Um gráfico desses resíduos em relação ao indexador das observações não deve mostrar padrões, caso isso ocorra, deve-se verificar se a função de ligação utiliza é a mais adequada.

Como a distribuição dos resíduos é assumida desconhecida, um gráfico de envelope semi-normais simulados é uma ferramenta de diagnóstico bastante útil. Esse tipo de gráfico pode ser produzido utilizando o seguinte algorítmo:

i) Ajuste o modelo e gere uma amostra de observações pseudo-aleatória simulada de tamanho n independente usando os valores preditos como o modelo real;

ii) Ajuste o modelo para os dados gerados e compute os valores absolutos dos resíduos ordenados;

iii) Repita i) e ii) k vezes;

iv) Compute a média, mínimo e máximo nos n conjuntos de k estatísticas de ordem;

v) Faço o gráfico desses valores e dos resíduos ordenados da amostra original em relação aos escores semi-normais $\Phi^{-1}[(i+n-1/8)/(2n+1/2)]$.

Os valores mínimos e máximos das k estatísticas de ordem criam os envelopes. Observações correspondentes ao resíduo absoluto que se encontram além ou aquém ao envelope requer investigação. Além disso, se vários pontos não estão dentro do envelope, isso evidencia que o modelo ajustado não é o adequado.

Em seguida, iremos definir indicadores de observações influentes e análise residual. O método de *generalized leverage* proposto por Wei *et. al* (1998), pode ser generalizado pela seguinte expressão evaluada no estimador de máxima verossimilhança de um parâmetro desconhecido θ:

$$\text{GL}(\theta) = \text{D}_\theta\left(-\frac{\partial^2\ell}{\partial\theta\,\partial\theta^T}\right)^{-1}\frac{\partial^2\ell}{\partial\theta\,\partial^T}$$

em que ℓ é a função da log-verossimilhança de um prâmetro desconhecido θ e $\text{D}_\theta = \partial\mu/\partial\theta^T$. Para o caso da regressão Beta, estamos interessados em aplicar essa generalização para os parâmetros (β,ϕ) da seguinte maneira:

$$\text{GL}(\beta) = TX(X^TQX)^{-1}X^TTM$$

sendo $M = \text{diag}\{m_1,\ldots,m_n\}$ com $m_i = 1/[y_i(1-y_i)]$ e $Q = \text{diag}\{q_1,\ldots,q_n\}$ com

$$q_i = \left\{\phi[\psi'(\mu_i\phi) + \psi'((1-\mu_i)\phi)] + (y_i^* - \mu_i^*)\frac{g''(\mu_i)}{g'(\mu_i)}\right\}\frac{1}{g'(\mu_i)^2}.$$

Assim, podemos obter a seguinte expressão:

$$\text{GL}(\beta,\phi) = \text{GL}(\beta) + \frac{1}{\gamma\phi}TX(X^TQX)^{-1}X^TTf(f^TTX(X^TQX)^{-1}X^TTM - b^T)$$

em que $b = (b_1,\ldots,b_n)^T$ com $b_i = -(y_i-\mu_i)/[y_i(1-y_i)]$. Quando ϕ é suficientemente grande, $\text{GL}(\beta,\phi) \approx \text{GL}(\beta)$. Para mais detalhes, ver Ferrari e Cribari-Neto (2004).

Uma medida de influência de cada observação nas estimativas dos parâmetros da regressão é a distância de Cook dada por $k^{-1}(\hat{\beta}-\hat{\beta}_{(i)})^TX^TWX(\hat{\beta}-\hat{\beta}_{(i)})$ em que $\hat{\beta}_{(i)}$ é a estimativa do parâmetro sem a i-ésima observação. Para evitar ter que ajustar o modelo $n+1$ vezes, é recomendado utilizar

uma aproximação da distância de Cook, dada por

$$C_i = \frac{h_{ii} r_i^2}{k(1-h_{ii})^2}$$

em que h_{ii} é um elemento da matriz $H = W^{1/2}X(X^TWX)^{-1}X^TW^{1/2}$.

5 Aplicação

Para uma ilustração de como aplicar uma modelagem de regressão Beta, considere o conjunto de dados coletado por Prater (1956). O conjunto de dados possui 32 observações (sem dados faltantes) e contém as seguintes variáveis:

- **yield**: Proporção de óleo não-refinado convertido para gasolina depois de distilação e fracionamento;
- **gravity**: Quão pesado é óleo não-refinado em comparação com àgua (Gravidade API);
- **pressure**: Pressão de vapor do óleo não-refinado (lbs/in^2);
- **temp10**: Temperatura em que 10% do óleo foi vaporizado (F°);
- **temp**: Temperatura em que todo o óleo foi vaporizado (F°);
- **batch**: Fator indicando a condição dos lotes.

A variável resposta na análise será **yield** (numérica entre 0 e 1) e as covariáveis serão **temp** (numérica 0 à infinito) e **batch** (categórica com 10 níveis). O conjunto de dados foi analisado por Atkinson (1985), no trabalho foi utilizado um modelo de regressão linear e foi percebido que a distribuição do erro não era simétrica, gerando resíduos discrepantes.

Assim, ele aplicou uma transformação na variável resposta de forma que **yield*** pertença a reta Real e, então, aplicou o modelo de regressão linear na variável transformada. Para o conjunto de dados proposto, Ferrari e Cribari-Neto (2004), desenvolveram a regressão Beta e a utilizaram para modelar os dados. As estimativas se encontram na Tabela 1.

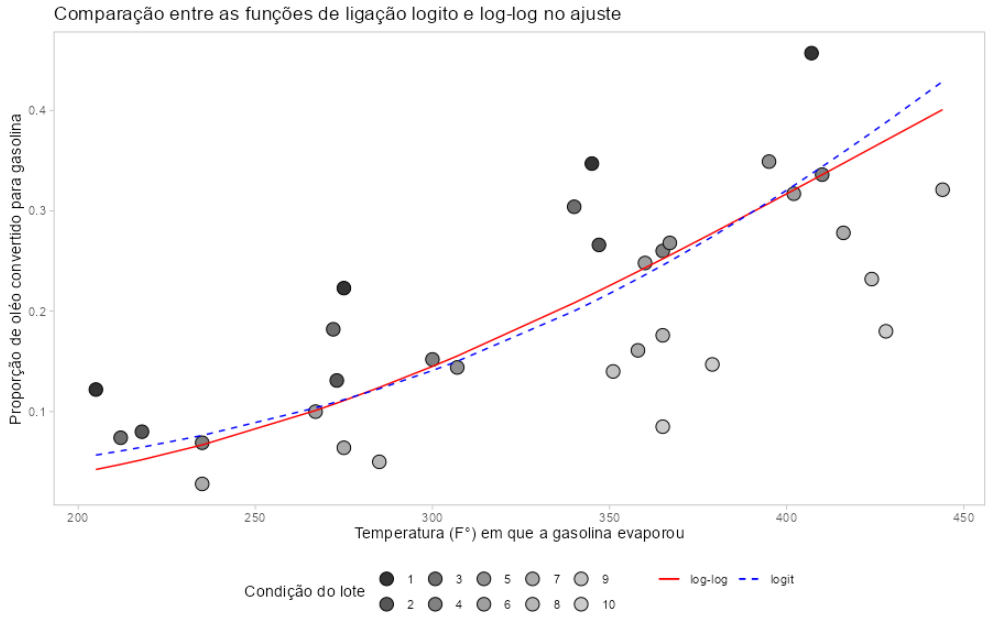

Figura 2. Comparação entre as funções de ligação logito e log-log na modelagem da regressão Beta

	Estimativa	Erro Padrão	z value	Pr(>|z|)	Signif.
(Intercepto)	-6.16	0.18	-33.78	<2e-16	***
temp	0.01	0	26.58	<2e-16	***
batch1	1.73	0.1	17.07	<2e-16	***
batch2	1.32	0.12	11.22	<2e-16	***
batch3	1.57	0.12	13.54	<2e-16	***
batch4	1.06	0.1	10.35	<2e-16	***
batch5	1.13	0.1	10.95	<2e-16	***
batch6	1.04	0.11	9.81	<2e-16	***
batch7	0.54	0.11	4.98	6.29e-07	***
batch8	0.5	0.11	4.55	5.3e-06	***
batch9	0.39	0.12	3.25	0.00114	**

Tabela 1. Coeficientes da regressão Beta com a função de ligação logito

O valor do pseudo-R^2 foi igual à 0.9617 em 51 iterações (BFGS com algorítmo quasi-Newton). Os gráficos de diagnóstico podem ser encontrados na Figura 4.

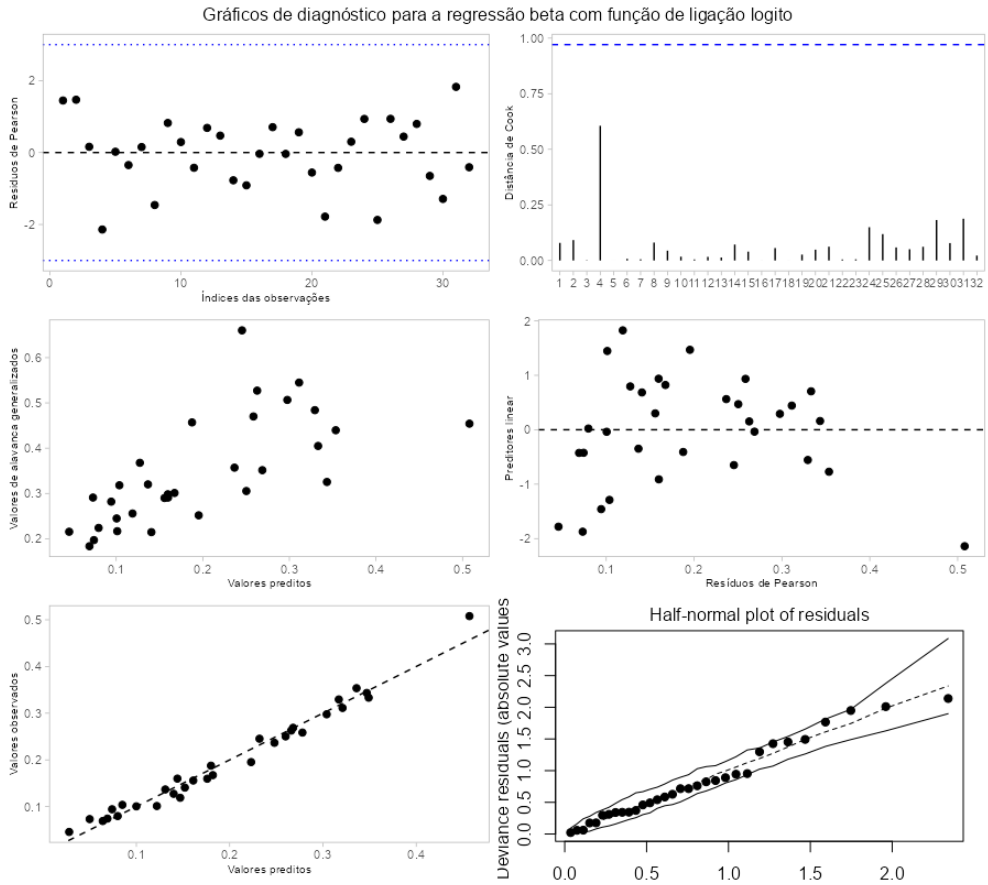

Figura 3. Gráficos de diagnóstico da modelagem com a função de ligação logito.

É notável que existe ao menos um ponto que aparenta ser discrepante, a observação 4, nos gráficos da distância de Cook, Resíduos por Índice, Alavanca generalizada, Valores Preditos por Observáveis e Envelope é o que possui maior destaque. A observação 29 também possui destaque

no gráfico de alavanca, mas ao investigar esse ponto não apresenta uma diferença muito larga em relação aos demais pontos, com exceção do 4.

Portanto, foi aplicado uma segunda modelagem *post-hoc* sem a quarta observação, dada que ela foi julgada como influente nos dados, como pode ser visto na Tabela 2.

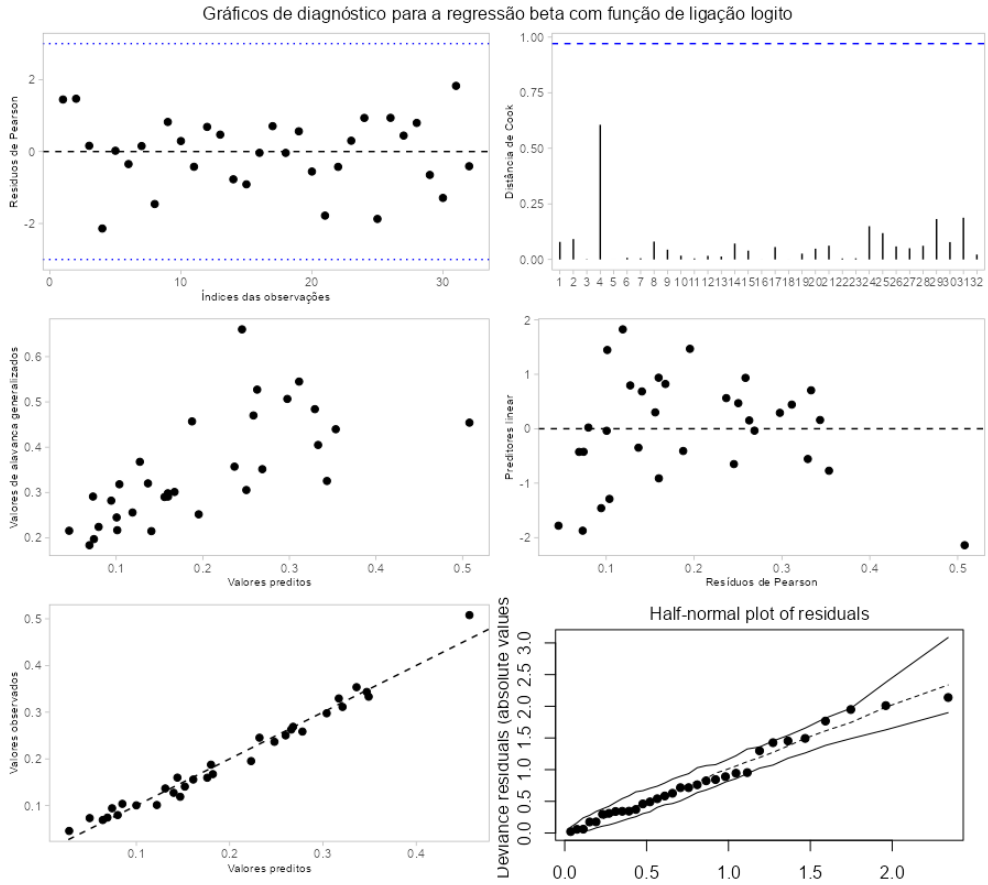

Figura 4. Gráficos de diagnóstico da modelagem com a função de ligação logito.

As estimativas pontuais entre os dois modelos não foram muito divergentes, mas o parâmetro de precisão ϕ aumentou de 440.3 para 577.8. Porém, a diferença entre os erros padrão assintóticos dos dois modelos foram negligenciáveis.

	Estimativa	Erro Padrão	z value	Pr(>|z|)	Signif.
(Intercepto)	-6.36	0.17	-37.04	<2e-16	***
temp	0.01	0	29.05	<2e-16	***
batch1	1.89	0.1	18.83	<2e-16	***
batch2	1.37	0.1	13.15	<2e-16	***
batch3	1.63	0.1	15.8	<2e-16	***
batch4	1.08	0.09	12.04	<2e-16	***
batch5	1.15	0.09	12.7	<2e-16	***
batch6	1.06	0.09	11.38	<2e-16	***
batch7	0.57	0.1	5.91	3.39e-09	***
batch8	0.5	0.1	5.25	5.25e-07	***
batch9	0.39	0.1	3.71	0.0002	***

Tabela 2. Coeficientes da regressão Beta com a função de ligação logito sem a 4° observação

6 Considerações Finais

Esse trabalho discute o modelo de regressão Beta, pensado para modelar variáveis contínuas mensuráveis no intervalo $(0, 1)$, uma situação bastante comum em diversos cenários, como quando o pesquisador deseja modelar proporções. A suposição que deve ser considerada é que a variável resposta seja uma Beta com parâmetros desconhecidos α e β. Dado a flexibilidade da distribuição da variável aleatória Beta, é possível modelar diversas situações diferentes, assimétricas ou não. Para conduzir a análise de regressão Beta é necessário reparametrizar a distribuição de Y de forma que os novos parâmetros representem uma medida de média e precisão.

Se a função de ligação utilizada for a logit, então os parâmetros podem ser interpretados em termos da razão de chance ao alterar c unidades em uma cováriavel. A estimação dos parâmetros é realizada através da máxima verossimilhança que possui funções escores fechadas, outros processos necessários também são discutidos na seção 4. Ao final, foi demonstrado a funcionalidade por meio de um exemplo utilizando o conjunto de dados coletado por Prater (1956), ajustando um modelo de regressão Beta com função de ligação logito e também foi conduzido uma análise de diagnóstico. Ao encontrar um ponto influente, foi realizado uma análise de regressão $post_hocsemopontodeinfluência$.

As análises foram replicadas utilizando o Software RStudio (Versão 4.1.1) com auxílio do pacote tidyverse para fins de transformação de variáveis e gráficos e o pacote betareg para realizar a regressão Beta. Para visualizar os códigos realizados acesse o seguinte repositório do GitHub: https://github.com/wesleyacruzzz/trabalho_regressao_mestrado

7 Bibliografia

- *Bury, K. (1999) Statistical Distributions in Engineering (New York: Cambridge University Press).*
- *Cribari-Neto F, Zeileis A (2010). "Beta Regression in R." Journal of Statistical Software, 34(2), 1–24. URL http://www.jstatsoft.org/v34/i02/.*
- *Ferrari SLP, Cribari-Neto F (2004). "Beta Regression for Modelling Rates and Proportions." Journal of Applied Statistics, 31(7), 799–815.*
- *Johnson, N. L., Kotz, S. Balakrishnan, N. (1995) Continuous Univariate Distributions, vol. 2, 2nd edn (New York: Wiley).*
- *Jørgesen, B. (1997) Proper dispersion models (with discussion), Brazilian Journal of Probability and Statistics, 11, pp. 89–140.*
- *McCullagh, P. Nelder, J. A. (1989) Generalized Linear Models, 2nd edn (London: Chapman and Hall).*
- *Nocedal, J. Wright, S. J. (1999) Numerical Optimization (New York: Springer-Verlag).*
- *RStudio Team. RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL. Disponível em: http://www.rstudio.com/.*
- *Wei, B.-C., Hu, Y.-Q. Fung, W.-K. (1998) Generalized leverage and its applications, Scandinavian Journal of Statistics, 25, pp. 25–37.*