

AVALIAÇÕES NUMÉRICAS DE GRÁFICOS DE CONTROLE DE REGRESSÃO BETA COM DIFERENTES RESÍDUOS

Cristine Rauber Oliveira^{1*}, Fábio Mariano Bayer²

1*. Estudante de IC da Universidade Federal de Santa Maria, UFSM

2. Orientador

Resumo:

Gráficos de controle de regressão são úteis no monitoramento de processos na presença de covariáveis controle. Nesses casos, um modelo de regressão é ajustado e os resíduos do modelo são monitorados por meio de gráficos de controle usuais, como os de Shewhart. Quando o objetivo é monitorar variáveis do tipo fração ou proporção, gráficos de controle de regressão beta podem ser considerados. Contudo, a literatura evidencia diversos resíduos possíveis para o modelo de regressão beta. O presente trabalho tem o objetivo de avaliar gráficos de controle de regressão beta usando diferentes resíduos. Os números médios de corrida (ARL) dos gráficos de controle baseados em cada resíduo são avaliados numericamente via simulações de Monte Carlo. Os resultados mostraram que o resíduo deviance é o mais adequado para ser utilizado no gráfico de controle de regressão beta.

Palavras-chave: Gráficos de controle; Regressão beta; resíduos.

Apoio financeiro: Os autores agradecem ao CNPq pelo auxílio financeiro recebido.

Trabalho selecionado para a JNIC pela instituição: UFSM.

Introdução:

A variabilidade está presente na maioria dos processos de produção, por mais que eles sejam cuidadosamente monitorados. Porém, fontes de variações podem ocorrer, como falhas de operadores, matéria-prima defeituosa ou máquinas operadas inadequadamente. Quando isso acontece dizemos que o processo de produção está fora de controle (Montgomery, 2007). Uma das ferramentas capazes de detectar alguma falha que possa ter ocorrido é o gráfico de controle.

Shewhart (1931) apresenta gráficos de controle para variáveis aleatórias independentes. No entanto, quando duas ou mais variáveis controle são relacionadas, o uso dos gráficos de controle de Shewhart pode não ser adequado, pois esses gráficos não consideram

toda a sequência de observações, somente o último dado da amostra (Montgomery, 2007). Para contornar esse problema, gráficos de controle de regressão foram propostos por Mandel (1969). Essa técnica consiste em relacionar variáveis de controle e a característica de qualidade do processo através de um modelo de regressão linear.

Contudo, quando a característica de qualidade se limita ao intervalo (0,1), utilizar os gráficos de controle de regressão usuais pode não ser adequado. Sant'Anna (2009) propõe um gráfico de controle de regressão beta capaz de monitorar os resíduos deviance de um processo em que a característica de qualidade é do tipo taxas ou proporções. O presente trabalho tem por objetivo realizar avaliações numéricas de gráficos de controle de regressão beta baseados nos diferentes resíduos apresentados por Espinheira (2008).

Metodologia:

A análise de regressão visa explicar a relação entre uma variável resposta, ou independente, e uma ou mais variáveis explicativas, ou dependentes. Para a modelagem de dados do tipo fração e proporção, Ferrari e Cribari-Neto (2004) propuseram um modelo de regressão beta. Seja y_1, y_2, \dots, y_n um conjunto de variáveis aleatórias independentes. Cada y_t , onde $t = 1, 2, \dots, n$, com média μ_t e precisão ϕ desconhecidas, possui densidade beta dada por:

$$f(y_t; \mu_t, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu_t \phi) \Gamma((1-\mu_t)\phi)} y_t^{\mu_t \phi - 1} (1-y_t)^{(1-\mu_t)\phi - 1},$$

com $0 < y_t < 1$, em que $0 < \mu_t < 1$ e $\phi > 0$. A média e a variância de y_t são dadas, respectivamente, por:

$$E(y_t) = \mu_t,$$

$$\text{Var}(y_t) = \frac{\mu_t(1-\mu_t)}{1+\phi}. \quad (1)$$

No modelo de regressão beta, a média μ_t é modelada através da seguinte estrutura:

$$g(\mu_t) = \sum_{i=1}^k x_{ti} \beta_i = \eta_t,$$

em que $\beta = (\beta_1, \dots, \beta_k)^T$ é um vetor de parâmetros desconhecidos e x_{t1}, \dots, x_{tk} são observações de k variáveis controles fixas e conheci-

das. Por fim, $g(\cdot)$ é uma função de ligação estritamente monótona e duas vezes diferenciável, conhecida como função de ligação. Neste trabalho assumimos função logit para $g(\cdot)$.

A função de log-verossimilhança é dada por

$$\ell(\beta, \phi) = \sum_{t=1}^n \ell_t(\mu_t, \phi),$$

em que

$$\begin{aligned} \ell_t(\mu_t, \phi) = & \log\Gamma(\phi) - \log\Gamma(\mu_t\phi) \\ & - \log\Gamma((1 - \mu_t)\phi) + (\mu_t\phi - 1)\log(y_t) \\ & + \{(1 - \mu_t)\phi - 1\}\log(1 - y_t). \end{aligned}$$

Os estimadores de máxima verossimilhança dos parâmetros do modelo são obtidos maximizando numericamente a função de log-verossimilhança usando um algoritmo de otimização não linear. Detalhes sobre inferências no modelo de regressão beta podem ser acessados em Ferrari e Cribari-Neto (2004) e Cribari-Neto e Zeileis (2009).

Baseado nesse modelo, Sant'Anna (2009) propõe o gráfico de controle de regressão beta, com os seguintes limites de controle:

$$\begin{aligned} \text{LCS} &= \bar{r}_d + w\sqrt{s^2\bar{r}_d}, \\ \text{LC} &= \bar{r}_d, \\ \text{LCI} &= \bar{r}_d - w\sqrt{s^2\bar{r}_d}, \end{aligned} \quad (2)$$

em que LCS é o limite de controle superior, LC o limite central, LCI o limite de controle inferior, \bar{r}_d e $s^2(\bar{r}_d)$ são a média e variância do resíduo deviance, respectivamente, e w é uma constante que determina o número médio de amostras até o primeiro alarme falso (ARL_0). Pela aproximação da distribuição normal, usualmente utiliza-se $w=3$, resultando em um $ARL_0=370$.

O resíduo deviance é definido por (Ferrari e Cribari-Neto, 2004):

$$r_t^d = \text{sign}(y_t - \hat{\mu}_t) \{2(\ell_t(y_t, \hat{\phi}) - \ell_t(\hat{\mu}_t, \hat{\phi}))\}^{1/2}, \quad (3)$$

em que $\text{sign}(\cdot)$ é a função sinal, $\ell_t(y_t, \hat{\phi})$ é a contribuição da t -ésima observação da função de log-verossimilhança, $\hat{\mu}_t$ é a estimativa de máxima verossimilhança de μ_t no modelo ajustado, y_t é a estimativa de μ_t no modelo saturado e $\hat{\phi}$ é a precisão estimada.

A análise de diagnóstico é parte importante na modelagem de regressão. A literatura estatística é povoada por extensivos trabalhos sobre análise de diagnóstico e resíduo. Em especial, destaca-se o trabalho de Espinheira et al. (2008) por propor e avaliar diversos resíduos para regressão beta. Neste trabalho, assumimos que cada resíduo, quando utilizado para determinação dos limites de controle em (2), resulta em um gráfico de controle de regressão beta diferente. O objetivo se torna avaliar comparativamente esses gráficos de controle concorrentes.

Adicionalmente ao resíduo deviance, considerado na proposta original de Sant'Anna (2009), consideramos o resíduo ponderado padronizado 2 (r_t^{ww}), o resíduo ponderado padronizado 1 (r_t^w), o resíduo ponderado (r_t^*) e o resíduo padronizado (r_t) (Espinheira et al., 2008). Esses resíduos são dados, respectivamente, por:

$$r_t^{ww} = \frac{y_t - \hat{\mu}_t}{\sqrt{\hat{v}_t(1 - \hat{h}_{tt})}}, \quad (4)$$

$$r_t^w = \frac{y_t - \hat{\mu}_t}{\sqrt{\hat{v}_t}}, \quad (5)$$

$$r_t^* = \frac{y_t - \hat{\mu}_t}{\sqrt{\hat{\phi}\hat{v}_t}}, \quad (6)$$

$$r_t = \frac{y_t - \hat{\mu}_t}{\sqrt{\widehat{\text{var}}(y_t)}}, \quad (7)$$

em que $\widehat{\text{var}}(y_t) = \frac{\hat{\mu}_t(1 - \hat{\mu}_t)}{(1 + \hat{\phi})}$, $\hat{y}_t^* = \log \frac{\hat{y}_t}{1 - \hat{y}_t}$,

$\hat{\mu}_t^* = \psi(\hat{\mu}_t\hat{\phi}) - \psi((1 - \hat{\mu}_t)\hat{\phi})$, $\hat{v}_t = \{\psi'(\hat{\mu}_t\hat{\phi}) + \psi'((1 - \hat{\mu}_t)\hat{\phi})\}$, onde $\psi(\cdot)$ é a função digama e \hat{h}_{tt} é o t -ésimo elemento da diagonal da matriz chapéu H^* , definida em Ferrari et al. (2011).

A análise de resíduos visa identificar alguma observação atípica ou um modelo mal ajustado. Para isso, os resíduos ordinários são comumente utilizados. Porém, dependendo da distribuição dos dados a serem analisados, alguns resíduos podem gerar resultados distorcidos. No caso da regressão beta, o resíduo ponderado padronizado 2, Equação (4), proposto por Espinheira (2008), é considerado o mais adequado. De acordo com Pierce e Schafer (1986), o resíduo deviance, Equação (3), é o mais indicado para análise de diagnóstico em diferentes modelos de regressão. O resíduo padronizado, Equação (7), definido para regressão beta por Ferrari (2004), é o mais usado em análises gráficas.

Resultados e Discussão:

Sendo ARL_0 o número médio de pontos que devem ser plotados até que ocorra o primeiro alarme falso, quando o processo está sob controle, o ARL_1 é o número médio de observações até a detecção de alguma alteração quando o processo está realmente fora de controle (Montgomery, 2007). Neste sentido, uma análise de sensibilidade de gráficos de controle é feita pela comparação dos ARL_1 quando fixados os mesmos ARL_0 (Pedrini, 2009).

Na avaliação numérica, todos os gráficos de controle foram ajustados para um ARL_0 igual a 200. No processo de calibração dos gráficos de controle, para cada resíduo usamos um conjunto de limites e uma sequência da constante w , registrando o ARL_0 . Após a calibração dos gráficos de controle, para avaliação do ARL_1 , simulamos uma amostra fora de controle, em que um valor de δ foi aditiva-

mente incluído na estrutura da média, da seguinte forma:

$$g(\mu_i) = \delta + \beta_0 + \beta_1 X_{1i},$$

em que $\beta_0 = -3,2$, $\beta_1 = 2$ e X_1 são ocorrências da distribuição uniforme (0,1) que mantiveram-se fixas durante todo o experimento de Monte Carlo. Com essas instâncias, o valor médio de μ , sob controle ($\delta = 0$), foi aproximadamente 0,1. O parâmetro de precisão foi fixado em $\phi = 120$. Todo esse processo foi replicado 5.000 vezes e então os ARL's empíricos de cada gráfico de controle foram definidos. As implementações computacionais foram realizadas em linguagem R (2013).

A Figura 1 mostra o ARL de cada resíduo. Nota-se que todos os gráficos de controle estão calibrados para um mesmo $ARL_0 = 200$, quando $\delta = 0$. Quanto menor for o valor de ARL_1 , maior a sensibilidade do gráfico de controle na detecção de alguma falha. Para valores negativos de δ , destaca-se os resíduos (4), (5), (6), pois detectam mais rapidamente alguma alteração ocorrida na média. Para valores positivos de δ , o resíduo (7) se sobressai. O resíduo (3) possui um desempenho regular tanto para valores positivos quanto negativos de δ .

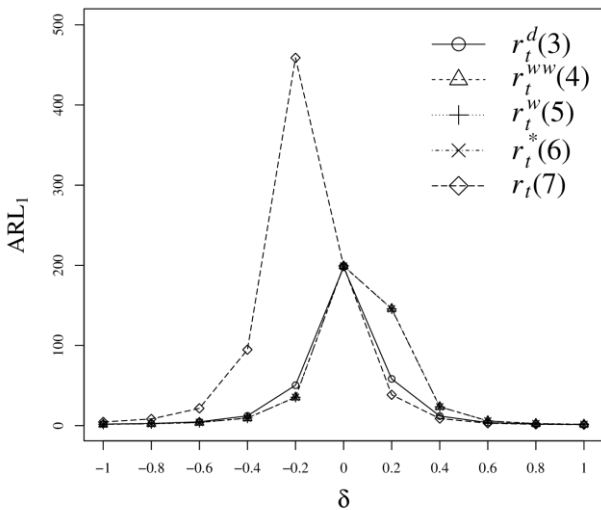


Figura 1: Gráfico ARL

Nota-se que o comportamento do resíduo padronizado (7) é errôneo, pois para um determinado intervalo de δ ele apresenta um ARL_1 maior que 200. Para justificar esse comportamento, utilizando uma amostra simulada, as Figuras 2, 3 e 4 mostram gráficos de controle de regressão beta para 3 resíduos avaliados, fixando $\delta = -0.2$. A medida que o valor de δ diminui, a média e a variância de y diminuem, conforme (1) e (7). Dessa forma, nota-se que os dados são menos dispersos, diminuindo a probabilidade de observar um valor fora dos limites, mesmo estando fora de controle. Consequentemente, tem-se um ARL_1 maior que ARL_0 , pois os resíduos tendem a permanecer

mais tempo em controle, como pode ser visto na Figura 3. Os gráficos de controle de regressão beta definidos pelos resíduos (3) e (4) são mais sensíveis. Quando $\delta = -0.2$, as Figuras 2 e 3 evidenciam que o resíduo ponderado padronizado 2 detecta alguma alteração na média mais rapidamente que o resíduo deviance. Porém, quando $\delta = 0.2$, ocorre o contrário, como pode ser verificado na Figura 1.

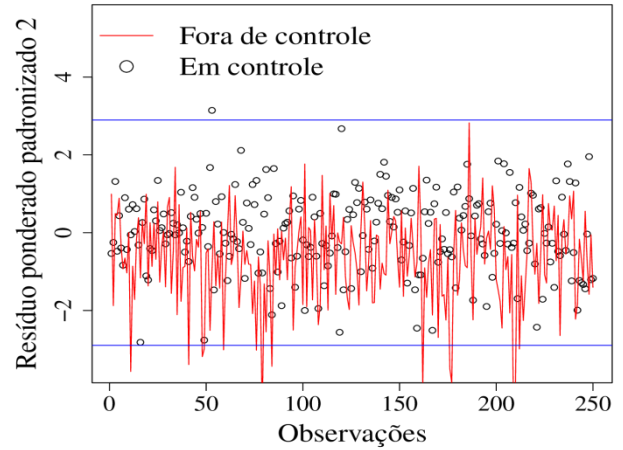


Figura 2: Gráfico de controle de regressão beta para uma amostra simulada utilizando o resíduo (4).

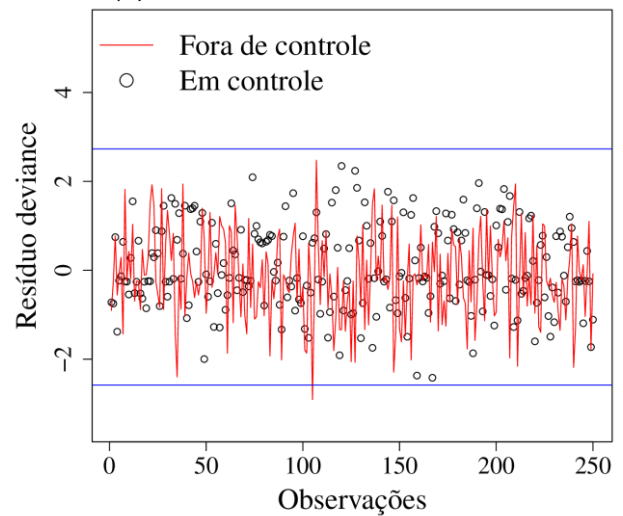


Figura 3: Gráfico de controle de regressão beta para uma amostra simulada utilizando o resíduo (3).

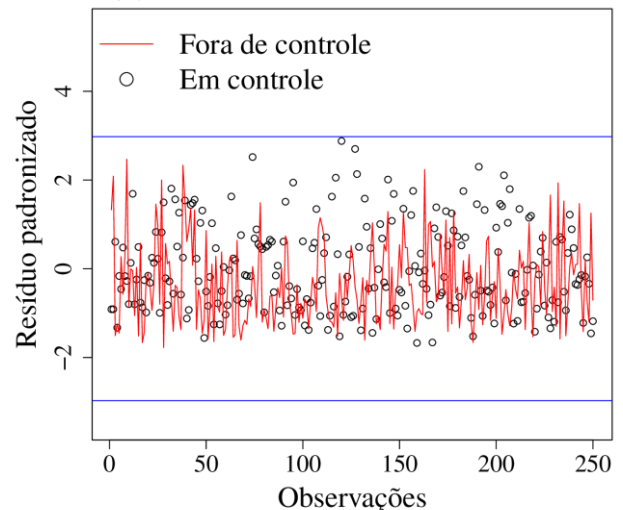


Figura 3: Gráfico de controle de regressão beta para uma amostra simulada utilizando o resíduo (7).

De forma geral, nota-se, pelas Figuras 1, 2, 3 e 4, que dependendo do valor de δ tem-se um resíduo que se comporta melhor que o outro. Todavia, em aplicações reais, não se pode prever a magnitude da alteração que ocorrerá na média. Por isso, o resíduo deviance se torna o mais adequado. Pois, independentemente do valor do δ , possui bom comportamento comparado aos outros resíduos analisados neste trabalho.

Conclusões:

Ao se monitorar variáveis do tipo taxas ou proporções relacionadas a variáveis controladas, recomenda-se utilizar gráficos de controle de regressão beta. Entretanto, na literatura existem vários resíduos que podem ser adotados. Com isso, foi realizada uma avaliação numérica de gráficos de controle de regressão beta, fundamentada em diferentes resíduos. Através de simulações de Monte Carlo foram encontrados os limites de cada gráfico e avaliados os seus respectivos ARL's. Os resultados indicaram que o resíduo deviance é o mais apropriado na aplicação em gráficos de controle de regressão beta.

Referências bibliográficas

Cribari-Neto, F., Zeileis, A., 2009. Beta regression in R.

Espinheira, P. L., Ferrari, S. L., Cribari-Neto, F., 2008. On beta regression residuals. *Journal of Applied Statistics*, 35 (4), 407-419.

Ferrari, S., Cribari-Neto, F., 2004. Beta regression for modelling rates and proportions. *Journal of Applied Statistics*, 31 (7), 799-815.

Ferrari, S. L., Espinheira, P. L., Cribari-Neto, F., 2011. Diagnostic tools in beta regression with varying dispersion. *Statistica Neerlandica* 65 (3), 337-351.

Mandel, B., 1969. The regression control chart. *Journal of Quality Technology* 1 (1), 1-9.

Montgomery, D. C., 2007. *Introduction to statistical quality control*. John Wiley & Sons.

Pedrini, D. C., 2009. Proposta de um método para aplicação de gráficos de controle de regressão no monitoramento de processos.

Pierce, D. A., Schafer, D. W., 1986. Residuals in generalized linear models. *Journal of the*

American Statistical Association, 81 (396), 977-986.

R Core Team, 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org/>

Sant'Anna, Â. M. O., 2009. Ferramentas para modelagem e monitoramento de características de qualidade do tipo fração.

Shewhart, W. A., 1931. *Economic control of quality of manufactured product*. ASQ Quality Press.