PLANEJAMENTO DE GRÁFICOS DE CONTROLE DE REGRESSÃO VIA SIMULAÇÃO

Ana Carolina Campana Nascimento[1], José Ivo Ribeiro Júnior[1], Moysés Nascimento[1]

1. Professor da Universidade Federal de Viçosa, Avenida Peter Henry Rolfs, s/n Campus Universitário 36570-000, Viçosa – MG – Brasil. (ana.campana@ufv.br)

Recebido em: 04/05/2012 – Aprovado em: 15/06/2012 – Publicado em: 30/06/2012

RESUMO

O gráfico de controle de regressão é útil para monitorar um processo onde existe a atuação conjunta de variáveis correlacionadas. Neste caso, o objetivo é controlar a relação entre elas. O relacionamento linear entre as variáveis é representado pela reta de regressão e em torno dela, são estabelecidos os limites de controle estimados a partir de dados históricos. Porém, se existir uma grande variabilidade nos dados, o limite será alargado e, portanto, irão conter todos ou quase todos os dados, dando uma falsa impressão de que o processo está sob controle. Assim, o estabelecimento dos limites de controle por meio de dados simulados, que representem uma meta de interesse baseada no coeficiente de correlação entre as variáveis em estudo, é bastante apropriado em viabilizar o controle da relação entre estas variáveis.

PALAVRAS CHAVE: coeficiente de correlação, simulação de dados, controle estatístico de processos

PLANNING GRAPHICS CONTROL OF REGRESSION VIA SIMULATION

ABSTRACT

The graphic control of the regression is useful to track a process where there is joint action of variables correlated. In this case, the objective is to control the relationship between them. The linear relationship between the variables is represented by the line of regression and around it, are established boundaries of control estimated from historical data. But if there is great variability in the data, the limit will be extended and, therefore, will contain all or nearly all the data, giving a false impression that the process is under control. Thus, the establishment of the limits of control through simulated data, representing a target of interest based on the correlation coefficient between the variables under study, is quite appropriate to facilitate the control of the relationship between these variables.

KEYWORDS: correlation coefficient, simulation data, statistical control processes

1 INTRODUÇÃO

O gráfico de controle foi originalmente proposto por W. A. SHEWHART[1] em 1924 com a intenção de eliminar variações anormais em um determinado processo produtivo. Portanto, para controlar a qualidade de um produto, são necessárias: a medição e a identificação de variações ocorridas no processo de produção. Geralmente, os gráficos de controle são utilizados para avaliar o estado de controle estatístico de um processo, pois servem para diferenciar se as variações que ocorrem são devidas a causas especiais ou aleatórias (MONTGOMERY,1997).

Existem vários estudos sobre o controle de variáveis em processos produtivos, mas na maioria das vezes, são em relação ao controle de uma única variável. Porém, em muitos casos, o processo sofre a interferência conjunta de mais de uma e então surge a necessidade de um controle simultâneo dessas variáveis, que podem ser correlacionadas. Uma forma bastante eficiente de representar relações entre variáveis é através do gráfico de controle de regressão, apresentado inicialmente por DIPAOLA (1945), que descreve o controle simultâneo de duas variáveis que possuem uma relação de causa e efeito nos processos produtivos.

Ultimamente o gráfico de controle de regressão tem sido utilizado na resolução de diversos problemas, como no trabalho de VILALLOBOS (2003), que teve como objetivo fazer previsões e monitorar a qualidade do ar numa interseção controlada por semáforo, através das variáveis: “atrasos” medidos em segundos e a “poluição” medida em concentrações de monóxido de carbono (CO). JACOBI *et al.,* (2002) exemplificaram o uso desta metodologia com dados coletados no setor de Engenharia de Saneamento e Meio Ambiente.

A literatura apresenta três tipos possíveis de limites de controle e todos são definidos basicamente em função da variação aleatória ou residual dos dados históricos. Segundo PEDRINI & TEN CATEN (2011) a presença de valores extremos (*outliers*) nos dados históricos compromete o desempenho de um gráfico de controle, visto que, os mesmos inflacionam a variância residual (QME) alargando o limite de controle e consequentemente reduzindo o poder do gráfico.

Desta forma, para que se tenha um gráfico sensível à detecção de pontos fora dos limites, e consequentemente, do processo fora de controle estatístico, é necessário que se retirem os pontos atípicos antes de se calcular o QME, para que o mesmo seja o mais aleatório possível. Outra maneira de contornar o problema é estabelecer os limites de controle a partir de uma meta de interesse, que depende da necessidade e da correlação entre as variáveis estudadas.

Diante do exposto este trabalho tem por objetivo estabelecer uma estratégia prática e rápida para o planejamento e implantação dos gráficos de controle de regressão por meio da correlação entre as variáveis estudadas e simulação de dados.

2 METODOLOGIA

2.1 MODELO DE REGRESSÃO LINEAR SIMPLES

A base para a construção do gráfico de controle de regressão é dada pela teoria de regressão linear simples. A análise de regressão linear constitui no estabelecimento de modelos que interpretem a relação funcional entre variáveis. O

[1] Em memorando datado de 16 de maio de 1924, o Dr. Shewhart propôs o seu gráfico de controle para análise de dados resultantes de inspeção, fazendo com que os procedimento baseado na detecção e correção de produtos defeituosos começassem a ser substituídos por estudo e prevenção dos problemas relacionados à qualidade, de modo a impedir que os produtos defeituosos fossem produzidos.

modelo de regressão linear simples é dado por $y = \beta_0 + \beta_1 x + \varepsilon$, em que β_0, o intercepto, β_1, o coeficiente angular, são constantes desconhecidas e ε é o componente do erro aleatório. Assumem-se erros não correlacionados com média zero e variância desconhecida σ^2.

A estimação de β_0 e β_1 é feita através do método dos mínimos quadrados ordinários, o qual consiste em minimizar a soma dos quadrados dos erros. Os estimadores obtidos por este método são dados por

- β_0: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$;
- β_1: $\hat{\beta}_1 = \dfrac{\sum_i y_i x_i - \dfrac{\left(\sum_i y_i\right)\left(\sum_i x_i\right)}{n}}{\sum_i x_i^2 - \dfrac{\left(\sum_i x_i\right)^2}{n}}$.

Em que $\bar{x}$ e $\bar{y}$ são as médias das variáveis X e Y respectivamente. Assim o modelo ajustado, dá uma estimativa pontual da média de Y para um particular valor de X, é escrito da seguinte forma $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

A diferença entre o valor observado y_i e o valor correspondente ajustado $\hat{y}_i$ é o resíduo. Matematicamente $\hat{e}_i = y_i - \hat{y}_i \qquad i = 1, \ldots, n.$

Para que o modelo ajustado seja validado, é necessária a observação de alguns pressupostos, nos quais os resíduos são de extrema importância para a verificação dos mesmos. Os quatro principais pressupostos são: normalidade, homocedasticidade, independência e linearidade (MONTGOMERY & PECK, 1982).

Uma importante aplicação do modelo de regressão é a previsão de uma nova observação y correspondente a um especificado valor da variável independente x. Se x_0 é o valor da variável independente de interesse, então $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$, é o ponto estimado para o novo valor resposta y_0.

A construção do intervalo de previsão para uma observação y_0 é feita da seguinte maneira. Note que $\psi = y_0 - \hat{y}_0$, é normalmente distribuída com média zero e variância

$$V(\psi) = V(y_0 - \hat{y}_0) = \sigma^2 \times \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right], \quad \text{e, logo}$$

$$y_0 - \hat{y}_0 \approx N\left(0, \sigma^2 \times \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right),$$

Para cada estimativa particular da posição da reta e para cada valor particular x_0, os limites do intervalo de previsão a $100(1-\alpha)\%$ são dados por (DRAPER & SMITH, 1966):

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{QME \times \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)} \le y_0 \le \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{QME \times \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}.$$

Antes, porém, de começar a fazer previsões, deve-se atentar para algumas advertências importantes de DOWNING & CLARK (2000) citados por JACOBI (2002):

- Qualquer previsão baseada em um modelo de regressão é uma previsão condicional, pois a previsão da variável dependente está sujeita ao valor da variável independente;
- a reta de regressão é estimada utilizando-se dados passados, que não poderá predizer dados futuros se a relação entre X e Y se modificar;
- muitas previsões de regressão que procuram prever valores de Y em situações em que o valor de X está fora do intervalo estudado, conhecidas como extrapolações, são muito menos confiáveis do que previsões baseadas em valores da variável independente contidos no intervalo de valores previamente observados;
- o simples fato de existir uma forte associação entre duas variáveis não significa que haja entre elas uma relação de causa e efeito.

2.2 GRÁFICO DE CONTROLE DE REGRESSÃO

Conhecendo o processo e, portanto, as variáveis X e Y correlacionadas que o influenciam, pode-se então utilizar o gráfico de controle de regressão para fazer o monitoramento do mesmo.

O gráfico de controle de regressão, apresentado inicialmente por DIPAOLA (1945), é uma forma bastante eficiente de representar relações entre variáveis, pois o mesmo descreve o controle simultâneo das duas variáveis correlacionadas. Este tipo de gráfico exige do gerente de produção que este saiba relacionar as variáveis que possuem uma relação de causa e efeito. O estabelecimento errado das variáveis dependente e independente, certamente levará os gestores a tomarem decisões erradas, que poderão comprometer toda a estabilidade do processo (ALMEIDA, 2003).

Para se estabelecer o gráfico de controle de regressão, é necessário, primeiramente, que se analisem os dados históricos em um diagrama de dispersão, com o objetivo de confirmar se existe ou não a relação linear entre as variáveis, e ainda, de verificar a existência de pontos que fogem ao comportamento geral dos dados. Neste tipo de gráfico, os *outliers* podem indicar observações onde, por alguma razão, o relacionamento comum entre as duas variáveis de interesse não exista.

A partir daí, os pontos que saem do padrão linear mostrado pelo diagrama de dispersão devem ser investigados. Se esses pontos forem devidos a causas especiais eles não deverão ser utilizados para estabelecer o gráfico de controle.

Desse modo, o gráfico de controle de regressão, construído apenas com as observações que não fogem ao comportamento geral dos dados, deve ser usado como padrão do processo de forma que os novos dados sejam sobrepostos contra os limites calculados a priori.

Quanto às linhas dos limites de controle, estas são totalmente ou quase paralelas à reta de regressão e não ao eixo horizontal, como é o caso do gráfico de controle convencional e devem ser calculadas de acordo com um dos três métodos apresentados a seguir:

1. Limites Simples ($k\sigma$): $\hat{y} \pm k \times \sqrt{QME}$;

2. Limites de Predição: $\hat{y} \pm k \times \sqrt{QME \times \left[1 + \frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right]}$;

3. Limites de Confiança: $\hat{y} \pm k \times \sqrt{QME \times \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right]}$,

onde $QME = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$ e $S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}$.

Para tanto, necessita-se decidir qual o valor de *k* será utilizado, ou seja, quão perto dos limites (superior e inferior) se permite que o processo varie. Por exemplo, o uso de k=2 garante um limite mais estreito, com maior risco de alarmes falsos. O de k=3 garante limites mais largos, com menor probabilidade de alarmes falsos. Esta decisão, além de estatística deve ser gerencial, baseada na economia e na experiência do processo em questão.

Devido ao fato da construção dos limites de controle ser baseada em dados históricos, uma grande variabilidade nos dados pode fazer com que o limite seja alargado, fazendo com que o mesmo englobe todos ou quase todos os dados, o que pode dar uma falsa impressão de processo sob controle. Para contornar esse possível problema objetivou-se estabelecer uma estratégia prática e rápida para o planejamento e implantação dos gráficos de controle de regressão por meio da correlação entre as variáveis estudadas e simulação de dados.

2.3 PROPOSTA

A recomendação se baseia na estimação dos limites de controle a partir de uma meta de interesse, que vai depender da necessidade e da correlação entre as características estudadas, ao invés da utilização de dados históricos. Com isso busca-se monitorar o processo dentro de limites “ideais”, isto é, limites baseados na correlação desejada e, portanto, na maximização técnica da relação entre as variáveis estudadas. Isso deve ser feito, de forma que, ao estimar a correlação entre as variáveis X e Y sejam impostas metas progressivas, isto é, constrói-se o gráfico de controle baseado nos limites estimados em função de uma correlação, em valor absoluto, maior que aquela apresentada pelos dados.

Para construção do gráfico de controle de regressão a partir da meta de interesse, a qual se baseia no coeficiente de correlação, têm-se os seguintes passos:

1. plotar os dados históricos num diagrama de dispersão;
2. estimar o coeficiente de correlação linear (r), excluindo os pontos que fogem ao padrão dos dados;
3. simular os dados com o coeficiente de correlação desejado, maior em valor absoluto, daquele estimado;
4. definir e calcular os limites de controle a partir dos dados simulados;
5. coletar os dados e construir o gráfico de controle de regressão.

Desta forma são estabelecidas três regiões no gráfico de controle de regressão (figura 1). A primeira se refere aos pontos sob controle. A segunda e a terceira, aos pontos fora de controle estatístico. No entanto, as regiões 2 e 3 podem ser consideradas como melhores ou piores que o controle, de acordo com a aplicação.

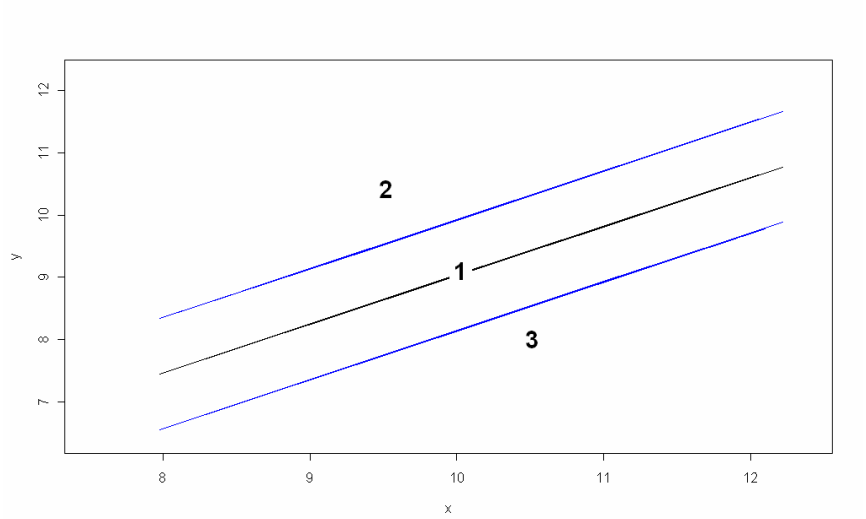

FIGURA 1. Regiões do gráfico de controle de regressão.

A recomendação das estimativas dos limites de controle baseadas num coeficiente de correlação maior, em valor absoluto, daquele observado nos dados, permite uma grande flexibilidade para a implantação do monitoramento pelo gráfico de controle de regressão. Além de possibilitar aumentos absolutos e progressivos da relação entre X e Y, ele estabelece o máximo possível dessa relação.

Para exemplificar o método proposto, foram simulados quatro conjuntos de 50 pares de dados (x_i, y_i). Em cada conjunto, as variáveis foram simuladas de modo que os coeficientes de correlação entre elas fossem iguais à: $r_1 = 0{,}9$, $r_2 = 0{,}7$, $r_3 = 0{,}5$ e $r_4 = 0{,}3$, no primeiro, segundo, terceiro e quarto conjuntos de dados, respectivamente.

Em cada caso, foram estabelecidos os gráficos de controle de regressão contendo os três tipos de limite (simples, predição e confiança), utilizando-se k=2. De acordo com a distribuição de t, estima-se um intervalo com aproximadamente $100(1-0{,}05)\%$ dos dados contidos.

Admitindo-se que a meta requerida esteja acima do que é observado no processo, foram considerados os seguintes casos:

a. Dados com $r_2 = 0{,}7$ foram sobrepostos aos limites de controle estimados para $r_1 = 0{,}9$;
b. Dados com $r_3 = 0{,}5$ foram sobrepostos aos limites de controle estimados para $r_2 = 0{,}7$.
c. Dados com $r_4 = 0{,}3$ foram sobrepostos aos limites de controle estimados para $r_3 = 0{,}5$.

Além disso, compararam-se os três métodos existentes para a estimação dos limites de controle, quanto à sensibilidade de detecção de pontos fora de controle.

As simulações, estimativas dos limites de controle e as construções dos gráficos de controle de regressão, foram obtidas utilizando a linguagem de programação do software livre R. Os algoritmos utilizados para simulação das variáveis correlacionadas e para a construção do gráfico de controle de regressão com os três tipos de limites de controle podem ser encontrados em: www.det.ufv.br/~anacarolina.

RESULTADOS E DISCUSSÃO

Considerando os quatro conjuntos de 50 pares de dados, como sendo os dados históricos, construiu-se os diagramas de dispersão apresentados nas figuras 2, 3, 4 e 5.

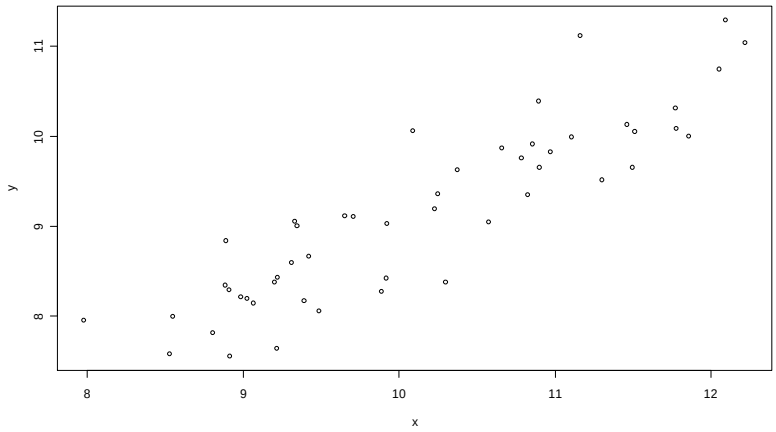

FIGURA 2. Diagrama de dispersão de Y em função de X para r_1=0,9.

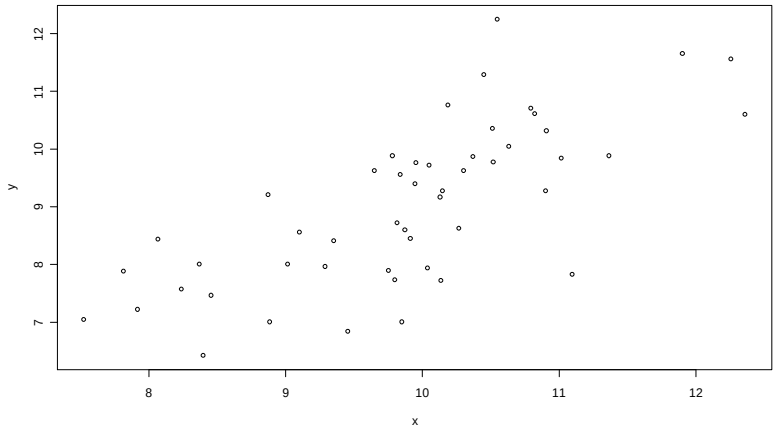

FIGURA 3. Diagrama de dispersão de Y em função de X para r_2=0,7.

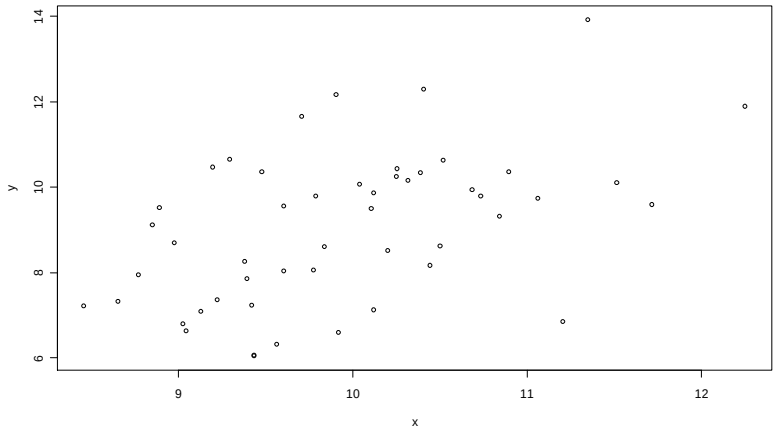

FIGURA 4. Diagrama de dispersão de Y em função de X para r_3=0,5.

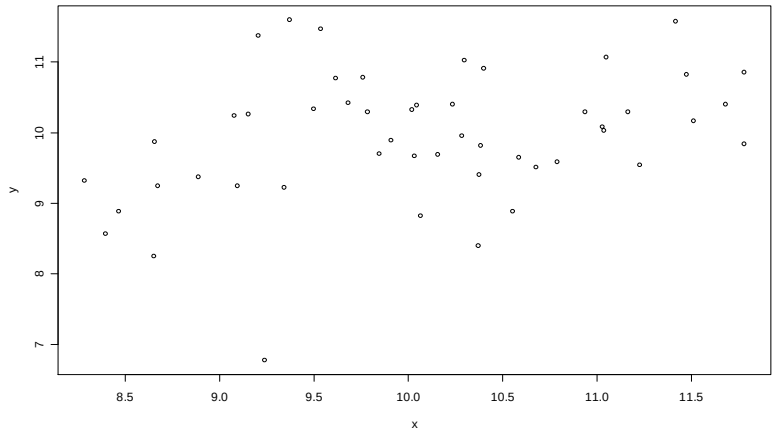

FIGURA 5. Diagrama de dispersão de Y em função de X para r_4=0,3.

Verifica-se nos quatro casos que não há pontos que fogem ao padrão dos dados e então, utilizando todos os dados simulados, foram estimados os limites de controle (figuras 6, 7, 8 e 9), com base nos três métodos descritos: predição, confiança e simples.

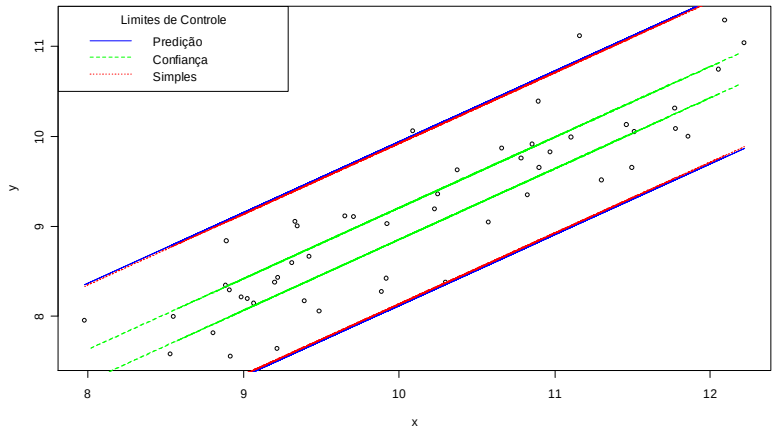

FIGURA 6. Estimativas dos limites de controle de Y em função de X para r_1=0,9.

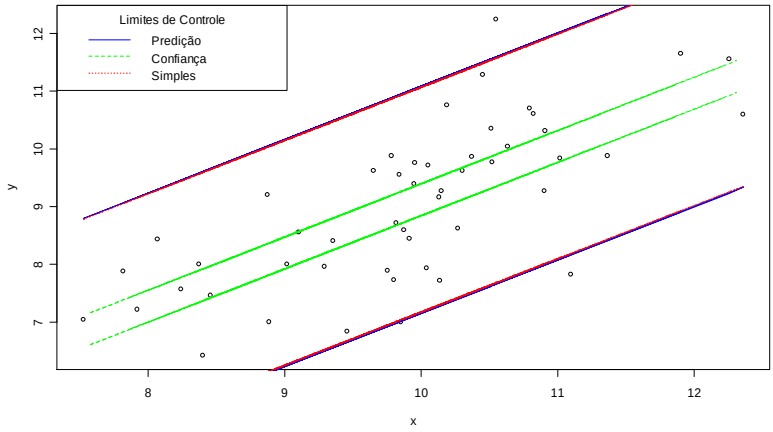

FIGURA 7. Estimativas dos limites de controle de Y em função de X para r_2=0,7.

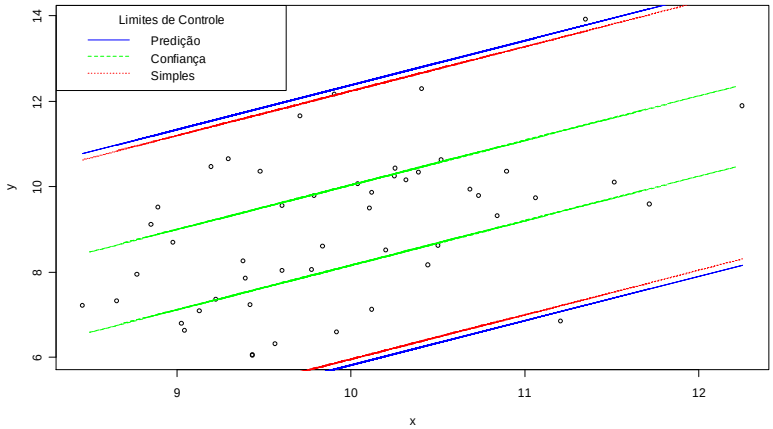

FIGURA 8. Estimativas dos limites de controle de Y em função de X para r_3=0,5.

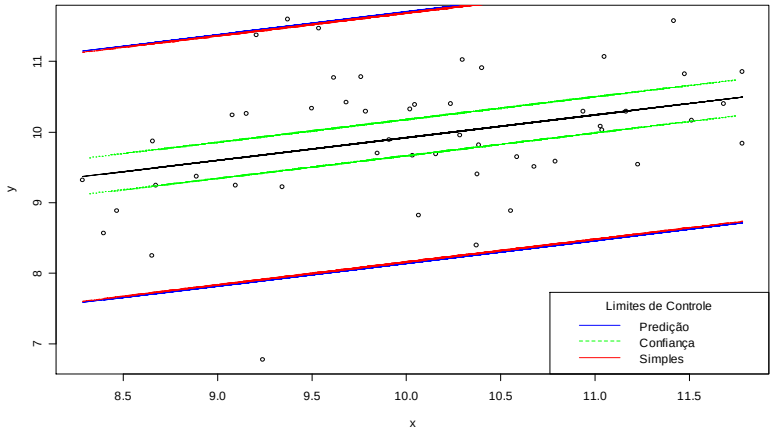

FIGURA 9. Estimativas dos limites de controle de Y em função de X para r_4=0,3.

As figuras 7, 8 e 9 mostram que o uso de dados históricos pode causar uma falsa impressão de que o processo esteja sob controle estatístico, uma vez que se existir uma grande variabilidade nos dados históricos, o limite será alargado fazendo com que o mesmo englobe todos ou quase todos dados. Por esta razão, propõe-se o estabelecimento dos limites de controle a partir de uma meta de interesse baseada na estimação do coeficiente de correlação.

Além disso, observou-se que os limites de controle simples e baseados no intervalo de predição têm estimativas similares. Há uma pequena diferença apenas para r_3=0,5, em que o de predição apresenta um intervalo mais largo. No entanto, o mais importante é que esses dois métodos de estimação proporcionam limites que compreendem aproximadamente 95% dos pares de dados (x_i, y_i). Como a simulação foi aleatória, há um indicativo de que o processo está sob controle estatístico para a relação entre as variáveis X e Y, no que se refere a todos os dados.

Por outro lado, as estimativas dos limites de controle baseados no intervalo de confiança, não proporcionaram um intervalo que compreendesse pelo menos 95% dos pares de dado, mas sim, para a média Y. Portanto, pode-se concluir que esse método não é eficiente em estimar um intervalo adequado à variação aleatória da relação entre as variáveis estudadas.

OLIN (1998) comparou os três métodos descritos acima e concluiu que os limites simples são os recomendados, pela facilidade de compreensão, simplicidade e similaridade aos gráficos de Shewhart.

Com base nos resultados optou-se pela construção do gráfico de controle de regressão a partir dos limites de controle simples ou de predição. A partir daí, passa-se para a segunda fase da implantação do gráfico de controle de regressão, que é a do monitoramento propriamente dita.

A recomendação nessa fase é de impor metas progressivas a partir do coeficiente de correlação estimado entre as variáveis X e Y. Se a estimativa for entre 0,3 e 0,5, constrói-se o gráfico de controle baseado nos limites de controle estimados em função de r_3=0,5 e sobrepõe-se aos dados coletados (figura 10).

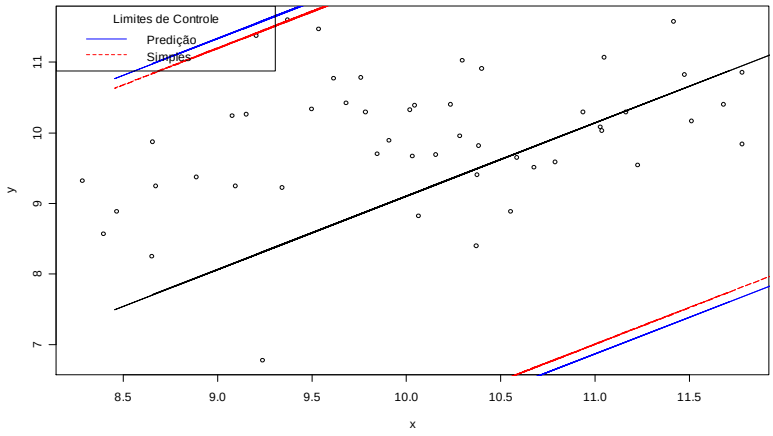

FIGURA 10. Gráfico de controle de regressão construído para os dados de r_4=0,3.

Se os dados apresentarem estimativas da correlação entre 0,5 e 0,7, sobrepõe-se aos limites de controle estimados em função de r_2=0,7 (figura 11).

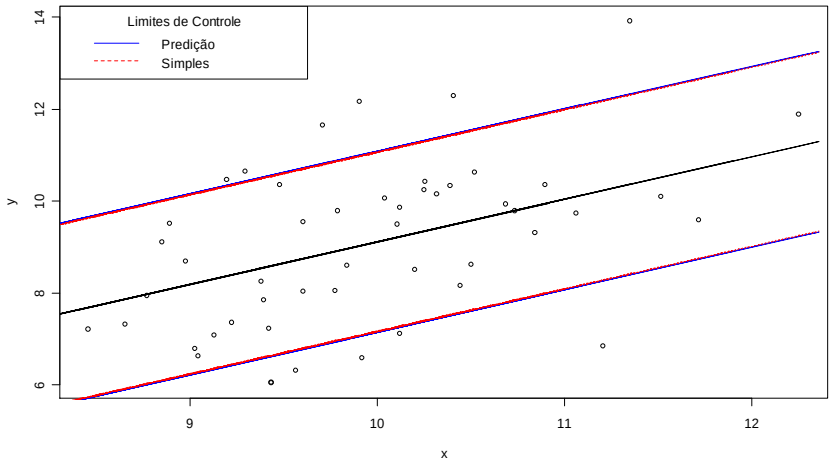

FIGURA 11. Gráfico de controle de regressão construído para os dados de r_3=0,5.

Se os dados apresentarem estimativas da correlação entre 0,7 e 0,9, sobrepõe-se aos limites de controle estimados em função de r_1=0,9 (figura 12).

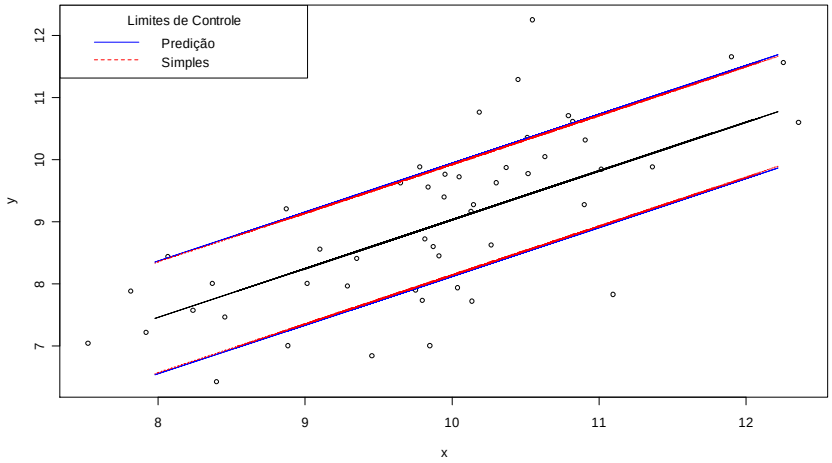

FIGURA 12. Gráfico de controle de regressão construído para os dados de r_2=0,7.

Pode-se observar que o gráfico de controle de regressão teve um bom desempenho em acusar pontos que tendem a comprometer a relação de interesse estabelecida para as variáveis nos casos em que a correlação amostral foi aproximadamente igual ou superior a r=0,5. Nos casos em que o processo apresenta pouca correlação entre as variáveis estudadas, como na figura 10 ($r \cong 0,3$), a utilização deste método não é recomendada uma vez que se exige a necessidade de uma relação que possa ser controlada, o que não se confirmou para a menor correlação testada. Neste caso, o limite inferior de controle compreendeu um grande intervalo dos valores de X no gráfico de controle de regressão.

No entanto, pode-se construir gráficos de controle mais ou menos sensíveis em detectar pontos fora de controle. Por exemplo, para os dados com r_3=0,5,

poderia-se construir um gráfico menos ou mais sensível baseado nos limites de controle estimados a partir de r=0,6 ou r=0,8, respectivamente.

Pode-se observar, a partir das figuras 11 e 12, que os limites de controle simulados a partir da meta de interesse são mais estreitos, o que permite viabilizar a busca para o verdadeiro estado de controle do processo. Deve-se tentar levar os pontos que estão na região “fora de controle”, no caso das regiões 2 e 3 (figura 1), para a região 1 (sob controle). Posteriormente, buscar aumentos absolutos e progressivos da relação entre X e Y, que possibilitem estabelecer o máximo possível dessa relação.

CONCLUSÕES

- A recomendação de estimar os limites de controle a partir de dados simulados em função de um coeficiente de correlação definido é adequada ao monitoramento realizado pelo gráfico de controle de regressão.
- As estimativas dos limites de controle baseadas no limite simples e no intervalo de predição são adequadas.
- No gráfico de controle de regressão, são geradas três regiões: melhor, sob e pior que o controle.
- O método não é indicado nos casos em que as variáveis estudadas apresentam fraca correlação.

REFERÊNCIAS

ALMEIDA, S. S. **Desenvolvimento de Gráficos de Controle Aplicados ao Modelo Funcional de Regressão.** 2003. 166f. Tese **(**Doutorado em Engenharia de Produção) - UFSC, Florianópolis.

DIPAOLA, P. P. Use of correlation in quality control. **Industrial Quality Control**, v.2, n.1, p.10-14, 1945.

DOWNING, D.; CLARK, J. **Estatística Aplicada: Um modo fácil de dominar os conceitos básicos**. 2 ed. São Paulo: Saraiva, 2006. 455p.

DRAPER, N.R.; SMITH, H. **Applied regression analysis**. New York: John Wiley, 1966. 407p.

JACOBI, L. F.; SOUZA, A. M.; PEREIRA, J. E. S. Gráfico de controle de regressão aplicado na monitoração de processos. **Revista Produção**, v. 12, n. 1, p. 46-59, 2002.

MONTGOMERY, D. C.; PECK, E. A. **Introduction to linear regression analysis**. New York: John Wiley, 1982. 500 p.

MONTGOMERY, D.C. **Introduction to statistical quality control**. 3rd ed. New York: Jonh Wiley, 1997. 677 p.

OLIN, Bryan D. **Regression control charts revisited: methodology and case studies**. In: Annual Fall Technical Conference, 42º, New York, 1998. Proceedings, New York, p. 1-17, 1998.

PREDINI, D.C.; TEN CATEN, C.S. Método para aplicação de gráficos de controle de regressão no monitoramento de processos. **Produção**, v.21, p.106-117, 2011.

VILALLOBOS, L. D. C. Uso dos gráficos de controle da regressão no processo de poluição em uma interseção sinalizada. In: XXIII **Encontro Nac. de Eng. de Produção**, 21, Ouro Preto, 2003.