UNIVERSIDADE FEDERAL DE MINAS GERAIS

Programa de Pós-Graduação em Estatística Aplicada

Monografia

Análise de regressão múltipla aplicada à predição da medida do raio de canto médio de perfis quadrados e retangulares após perfilação à frio

Autora: Mirelle Oliveira Spindola

Orientadora: Profª Ilka Afonso Reis

Belo Horizonte

Dezembro/2017

Mirelle Oliveira Spindola

Análise de regressão múltipla aplicada à predição da medida do raio de canto médio de perfis quadrados e retangulares após perfilação à frio

Monografia apresentada ao Curso de Pós-Graduação em Estatística Aplicada do Instituto de Ciências Exatas (ICEx) da Universidade Federal de Minas Gerais (UFMG) como requisito para conclusão do curso.

Orientadora: Profª Ilka Afonso Reis

Belo Horizonte

Dezembro/2017

SUMÁRIO

1. INTRODUÇÃO 1
2. OBJETIVO 2
3. MATERIAIS E MÉTODO 3
3.1 Banco de dados 3
3.2 Metodologia 4
3.2.1 Análise de regressão 5
3.2.2 Análise de predição do modelo 8
4. RESULTADOS 9
4.1 Análise descritiva 9
4.2 Análise de regressão 12
4.3 Análise da qualidade de predição do modelo 19
5. CONCLUSÃO 20
6. CONSIDERAÇÕES FINAIS 21
Referências 22
Apêndice 23

LISTA DE ILUSTRAÇÕES

Figura 1 – Esboço das características dimensionais de perfis: a) lado externo (A,B) e espessura de parede (WT); b) concavidade (C); c) perpendicularidade (P); d) torção (T); e) empeno (E); f) raio de canto externo (R) [(2)]..... 1
Figura 2 – Matriz de dispersão da variável resposta R* e as possíveis variáveis explicativas associadas às dimensões do perfil e matéria-prima (n = 22.241)..... 9
Figura 3 – Matriz de dispersão da variável resposta R* e as novas variáveis criadas (Delta e Razão) (n = 22.241). 10
Figura 4 – Dispersão marginal do raio de canto médio *versus* a diferença entre o perímetro da matéria-prima e do perfil (n = 22.241). 11
Figura 5 – Dispersão marginal do raio de canto médio *versus* espessura de parede nominal do perfil (n = 22.241). 11
Figura 6 – Gráfico de dispersão do raio de canto médio *versus* diferença entre o perímetro da matéria-prima e do perfil (a) indicando os pontos considerados como *outliers* ou influentes (total de 1094 pontos). (b) Apresentando somente os dados utilizados no modelo (total de 13.713 pontos).. 13
Figura 7 – Gráfico de dispersão do raio de canto médio *versus* espessura nominal do perfil (a) indicando os pontos considerados como *outliers* ou influentes (total de 1094 pontos). (b) Apresentando somente os dados utilizados no modelo (total de 13.713 pontos). 14
Figura 8 – Avaliação da suposição da normalidade dos erros via: histograma dos resíduos (a) e gráfico de probabilidade normal dos resíduos com teste de Anderson-Darling (b)..... 16
Figura 9 – Gráfico de dispersão entre resíduos e variáveis do modelo: Delta (a) e WT (b). 17
Figura 10 – Gráfico de dispersão entre resíduos e valores ajustados. 18
Figura 11 – Gráfico de dispersão valores ajustados *versus* valores observados no conjunto de teste (n = 7.413). 19

TABELAS

Tabela 1 – Representação da tabela ANOVA para testar a significância do modelo de regressão múltipla [4]. 6
Tabela 2 – Representação da tabela ANOVA para testar a significância do modelo de regressão múltipla quando existem duplicatas. 7
Tabela 3 – Análise de variância para testar a significância da regressão (modelo ajustado sem observações influentes e *outliers*). 15
Tabela 4 – Análise para testar a significância dos coeficientes da regressão (modelo ajustado sem observações influentes e *outliers*). 15

1. INTRODUÇÃO

Tubos quadrados e retangulares, também denominados como perfis, são amplamente utilizados em aplicações de engenharia, tais como elementos estruturais (colunas, vigas, treliças) em pontes, passarelas, máquinas agrícolas, entre outros. Esses perfis quadrados e retangulares podem ser produzidos a partir da deformação à frio de tubos circulares entre rolos.

A norma ASTM A500 [(1)] especifica as tolerâncias dimensionais admissíveis para as características lado, espessura de parede, concavidade, perpendicularidade, torção, retilineidade e raio de canto. A **Figura 1** exemplifica em detalhe as características especificadas em norma.

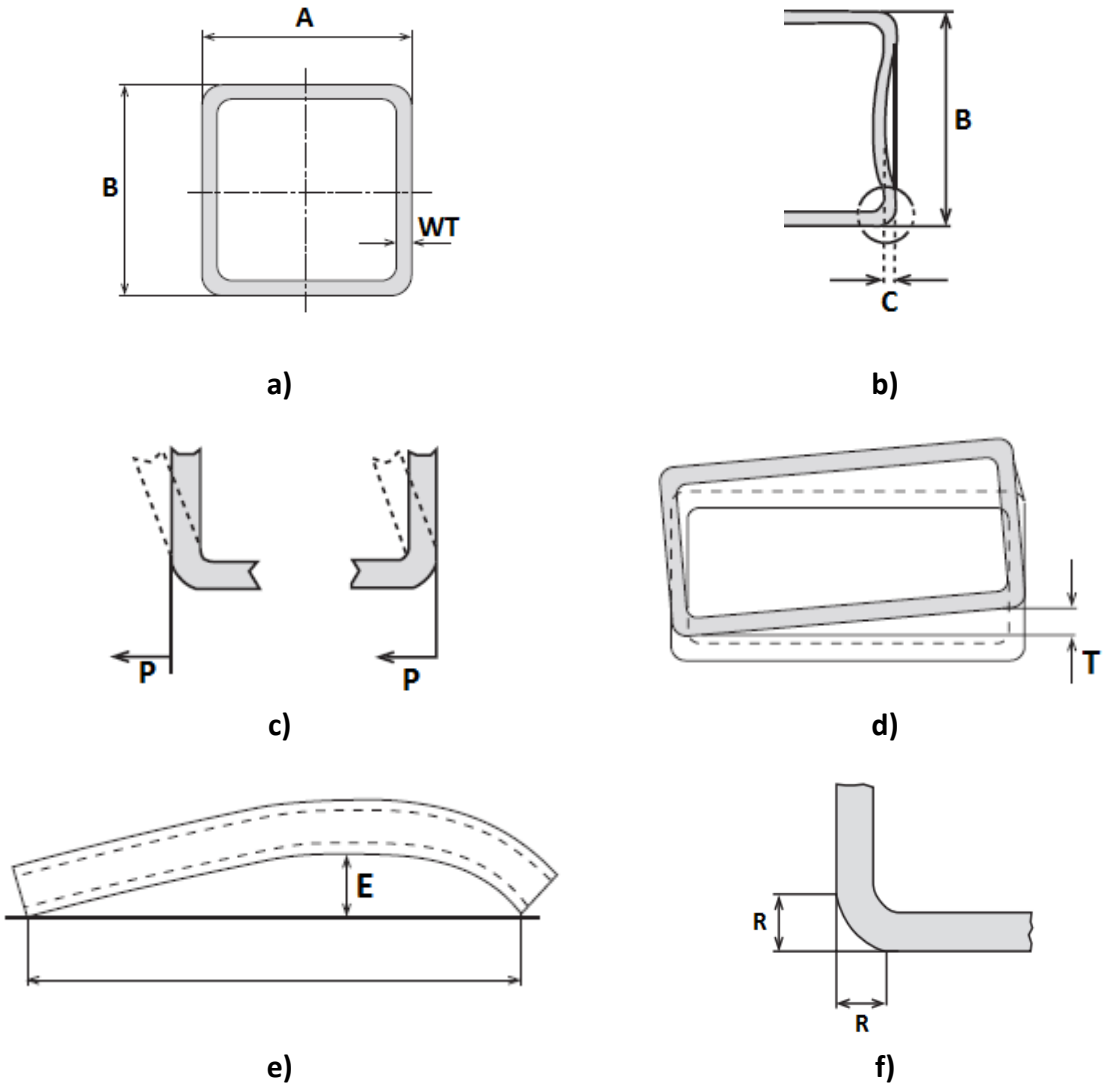


Figura 1 – Esboço das características dimensionais de perfis: **a)** lado externo (A,B) e espessura de parede (WT); **b)** concavidade (C); **c)** perpendicularidade (P); **d**) torção (T); **e)** empeno (E); **f)** raio de canto externo (R) [(2)].

Fonte: Vallourec & Mannesmann Tubes

Para a característica raio de canto externo (R), a norma especifica um valor máximo de três vezes o valor nominal de espessura de parede (WT). Entretanto, em algumas aplicações, para máquinas agrícolas principalmente, um menor valor de raio de canto é solicitado pelos clientes, com o intuito de facilitar a montagem de componentes que possuam encaixes ou com o objetivo de reduzir a quantidade de solda na região. Em outras situações, como em construção civil, o valor de raio de canto conforme ASTM A500 [(1)] não é exigido, mas o cliente deseja saber qual o valor previsto para realizar os cálculos estruturais, como momento de inércia, por exemplo.

Sabe-se que o raio de canto está diretamente relacionado com a matéria-prima utilizada, ou seja, com o tubo circular a ser deformado. Quanto maior o perímetro da matéria-prima em relação ao perímetro do perfil, menor será o raio de canto. Entretanto, um raio de canto muito pequeno não é desejado, pois pode gerar trincas na superfície interna do raio. Em contrapartida, se o perímetro da matéria-prima for muito menor que o perímetro do perfil, o raio de canto não atenderá os requisitos máximos estabelecidos em norma. Por fim, nesses casos pode ocorrer, ainda, o raio de canto assimétrico, ou seja, os quatro raios apresentam valores consideravelmente diferentes e isso pode gerar problemas estéticos na aplicação [(3)].

Vale ressaltar que as empresas de tubos circulares, em geral, produzem tubos com diâmetros pré-estabelecidos para fabricação, principalmente para reduzir custo com ferramental, visto que os clientes compram dimensões padronizadas de tubos circulares. Assim, dependendo da dimensão do perfil, é provável que a dimensão mais adequada para atendimento do raio de canto não seja produzida e, assim, faz-se necessário utilizar outra com dimensão mais próxima.

Diante desse contexto, quando novos perfis são demandados para as empresas que produzem tubos perfilados, em geral, são realizadas produções piloto em pequena escala, com o intuito de verificar se o perfil atende às especificações do cliente. Diante desse contexto, é importante estimar com precisão o valor de raio de canto que será obtido para perfis ainda não produzidos, baseando-se no perímetro e espessura de parede do perfil e diâmetros de matérias-primas disponíveis, com o objetivo de reduzir o tempo de resposta para o cliente e o custo com as produções piloto.

2. OBJETIVO

Construir um modelo de regressão múltipla para predizer o valor esperado para a característica raio de canto de perfis quadrados e retangulares em função de dimensões do perfil e do tubo circular utilizado como matéria-prima.

3. MATERIAIS E MÉTODO

3.1 Banco de dados

Nesse trabalho será utilizado o banco de dados de uma empresa produtora de tubos perfilados quadrados e retangulares. Nesse banco de dados estão registradas as medições de 26.537 peças, no período de amostragem corresponde às produções realizadas de janeiro/2015 até agosto/2017.

Nesse banco de dados são registrados o lote de produção, que, por meio de rastreabilidade interna, permite identificar qual matéria-prima foi utilizada, e os valores nominais e especificações do produto para as características lado do perfil, espessura de parede (WT) e raio de canto. Para cada lote de produção, em geral, são registrados os valores reais de pelo menos três peças, sendo duas medidas para o lado maior (A), duas medidas para o lado menor (B), quatro medidas de raio de canto (R), duas medidas de concavidade (C), perpendicularidade (P) e a medida de torção (T).

Na limpeza da planilha de registro de dados, foram excluídos os valores que não apresentavam sentido físico e/ou que tinham erros de digitação. Optou-se por utilizar somente os dados das peças com amplitude entre as medidas de um mesmo lado inferior a 5,0 mm e com amplitude entre as medidas dos raios de cantos inferior a 10,0 mm, visto que registros com maiores amplitudes também poderiam ser relacionados com erros de digitação ou peças com desvio de qualidade: peças assimétricas e/ou com lados fora do especificado pela ASTM A500.

A matéria-prima utilizada pode apresentar variação de ± 1% em relação ao diâmetro externo nominal e variação de -12,5% a +15% na espessura de parede. Como a dimensão da matéria-prima tem impacto direto na dimensão do perfil, tal variação ao longo do tubo, somada aos parâmetros de máquina da perfilação, pode gerar a não homogeneidade do raio de canto, ou seja, os quatro raios apresentarem medidas consideravelmente discrepantes na mesma seção. A norma ASTM A500 [(1)] não define um valor máximo de amplitude do raio canto na seção; porém, para algumas aplicações, tal variabilidade pode representar problemas estéticos e/ou de montagem no cliente. Diante disso, serão consideradas, na elaboração do modelo de regressão múltipla, somente as peças que apresentarem coeficiente de variação inferior a 0,15 para as quatro medidas da variável raio de canto.

Após limpeza do banco de dados, serão utilizadas na análise de regressão múltipla os dados de 22.241 peças.

Como na planilha de dados são registrados quatro valores para a característica raio de canto, uma em cada canto da peça, optou-se por utilizar como variável resposta a média desses resultados, que será denominada por **R***.

Além das informações registradas na planilha de dados, criou-se, ainda, as seguintes variáveis para avaliar uma possível correlação com a variável resposta **R***:

- perímetro da matéria-prima (P-MP) = πD, considerando o diâmetro externo da matéria-prima D;

- perímetro do perfil (P-P) = 2A + 2B, ou seja, a soma dos 4 valores de lado registrados na planilha de dados;

- Razão = P-MP / P-P, ou seja, a razão entre o perímetro da matéria-prima e perímetro do perfil;

- Delta = P-MP – P-P, ou seja, diferença entre o perímetro da matéria-prima e perímetro do perfil.

3.2 Metodologia

Acredita-se que a variável resposta raio de canto (R*) esteja relacionada com as seguintes variáveis:

- Espessura de parede do perfil (WT);

- Diâmetro da matéria-prima (D) ou perímetro da matéria-prima (P-MP);

- Perímetro do perfil (P-P);

- Razão entre perímetro da matéria-prima e do perfil (Razão);

- Diferença entre perímetro da matéria-prima e do perfil (Delta).

Avaliou-se a relação linear entre a variável resposta e as possíveis variáveis explicativas via matriz de gráficos de dispersão. Calculou-se, ainda, as estimativas para o coeficiente de correlação de Pearson (ρ_{XY}) para avaliar o grau de correlação linear entre as variáveis, o qual é dado por:

$$\rho_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \overline{y)}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

onde $\bar{x} = \frac{\sum_i^n x_i}{n}$ e $\bar{y} = \frac{\sum_i^n y_i}{n}$.

Quando mais próximo de 1 for o módulo do coeficiente de correlação de Pearson, mais forte é a correlação linear entre as variáveis. Se o coeficiente for positivo, significa que há uma correlação positiva, ou seja, quanto maior o valor de X maior será o valor de Y e vice-versa. E, se o coeficiente for negativo, existe uma correlação negativa entre as variáveis, ou seja, quanto maior o valor de X menor será o valor de Y e vice-versa.

Tal estudo descritivo da relação linear entre as variáveis auxiliou na definição das variáveis explicativas que seriam utilizadas no modelo de regressão. Não foram utilizadas as variáveis explicativas com alta correlação entre si para evitar multicolinearidade, além de se descartar aquelas variáveis explicativas que tinham pouca correlação com a variável resposta.

Segundo Montgomery e Runger (2012) [4], multicolinearidade ocorre quando a dependência entre as variáveis explicativas é forte. A multicolinearidade pode representar um problema na regressão, pois gera estimativas imprecisas dos coeficientes de regressão e, consequentemente, pode prejudicar a aplicabilidade do modelo estimado. Além disso, ela pode inflar as estimativas dos erros-padrões dos coeficientes de regressão, aumentando assim a probabilidade de erro do tipo II nos testes de significância desses coeficientes.

De modo a construir o modelo que permitisse prever com precisão o raio de canto médio, foram realizadas as análises apresentadas na Seção 4, utilizado os *softwares* Microsoft Excel 2016 e Minitab 16.

3.2.1 Análise de regressão

O modelo de regressão linear múltiplo descreve a relação entre a variável resposta, Y, e k variáveis independentes, conforme apresentado abaixo:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon,$$

onde β_j, j = 0, 1, ..., k, são os coeficientes de regressão e ϵ é um termo de erro aleatório [4].

O método dos mínimos quadrados é utilizado para estimar os coeficientes de regressão, de modo a minimizar a soma dos quadrados dos erros, conforme apresentado abaixo.

$$\sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})]^2,$$

onde n corresponde ao número total de observações. Já os resíduos são dados pela diferença do valor observado, y_i, e o valor ajustado, $\hat{y}_i$, dado por: $\hat{y}_i = \hat{\beta}_o + \hat{\beta}_1 x_{1i} + ... + \hat{\beta}_k x_{ki}$, onde $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_k$ são os estimadores de mínimos quadrados dos coeficientes de regressão.

As seguintes suposições são estabelecidas para que inferências estatísticas possam ser feitas no modelo de regressão linear [5]:

1- O valor esperado para os erros é igual a zero ($E(\epsilon_i) = 0$ para todo i).
2- A variância dos erros é constante ($V(\epsilon_i) = \sigma^2$ para todo i). Essa suposição de variância constante é chamada de homocedasticidade dos erros. A violação dessa suposição prejudica a estimativa da variância dos erros, o que pode influenciar o poder dos testes de hipóteses feitos sob o modelo de regressão.

3- Os erros são não correlacionados entre si.
4- Os erros são normalmente distribuídos. Se essa suposição não for válida, os testes de hipóteses e intervalos de confiança e predição podem estar equivocados.

Avaliou-se o teste F de Fisher para a significância da regressão a partir da tabela de análise de variância (ANOVA), representada na **Tabela 1**, na qual se apresenta a decomposição da variabilidade total da resposta (SQT) em duas fontes de variação: a variabilidade explicada pelo modelo de regressão (SQR) e a variabilidade devido ao erro (SQE) [(4)].

Tabela 1 – Representação da tabela ANOVA para testar a significância do modelo de regressão múltipla [(4)].

Fonte de Variação	Soma dos quadrados	Graus de liberdade	Média quadrática	F_0
Regressão	SQR	k	MQR	MQR/MQE
Erro	SQE	n – k – 1	MQE	
Total	SQT	n – 1		

As hipóteses para o teste de significância do modelo são:

H_0: $\beta_1 = \beta_2 = ... = \beta_k = 0$
H_1: $\beta_j \neq 0$ para no mínimo um j, j =1, 2, ..., k.

A hipótese nula de que os coeficientes do modelo de regressão são todos nulos é rejeitada se o valor da estatística de teste F_0 for maior que $f_{\alpha, k, n-k-1}$, o percentil (1-α) da distribuição F com k graus de liberdade no numerador e (n-k-1) graus de liberdade no denominador, sendo α o nível de significância do teste.

Posteriormente, caso a hipótese nula do teste F da tabela ANOVA for rejeitada, avalia-se a significância individual de cada um dos coeficientes de regressão, inclusive para o intercepto, através de um teste t-Student, considerando as seguintes hipóteses:

H_0: $\beta_j = 0$
H_1: $\beta_j \neq 0$, j=0, 1, 2, ..., k

A estatística de teste para essa hipótese é

$$T_0 = \frac{\hat{\beta}_j}{EP(\hat{\beta}_j)},$$

onde $EP(\hat{\beta}_j)$ é o erro padrão do estimador $\hat{\beta}_j$.

A hipótese nula de que o coeficiente β_j não é significativo (nulo) será rejeitada se o módulo do valor da estatística de teste T_0 for maior que $t_{\alpha/2, n-k-1}$, o percentil (1-α/2) da distribuição t-Student com (n-k-1) graus de liberdade, sendo α o nível de significância do teste.

Quando existem réplicas no banco de dados, ou seja, há valores repetidos nas k variáveis explicativas, é possível estimar a parte da variabilidade da resposta devida ao erro puro, que é a variabilidade que permanece nos valores de Y quando o valor de X é fixado. Sendo assim, a soma de quadrados do erro (SQE) pode ser decomposta em soma de quadrados devido ao erro puro (SQEP) e soma de quadrados de falta de ajuste (SQFA). Utiliza-se também um teste F para avaliar a hipótese nula de que não há falta de ajuste de um modelo linear aos dados. Nesse caso, a tabela ANOVA é apresentada conforme **Tabela 2**, onde m é o número de níveis distintos da variável resposta (x_1, x_2, ... , x_k).

Tabela 2 – Representação da tabela ANOVA para testar a significância do modelo de regressão múltipla quando existem duplicatas.

Fonte de Variação	Soma dos quadrados	Graus de liberdade	Média quadrática	F_0
Regressão	SQR	k	MQR	MQR/MQE
Erro	SQE	n – k – 1	MQE	
Falta de ajuste	SQFA	m – k – 1	MQFA	MQFA/MQEP
Erro puro	SQEP	n – m	MQEP	
Total	SQT	n – 1		

As hipóteses para o teste F da falta de ajuste são:

H_0: o modelo linear é adequado (não há falta de ajuste)
H_1: o modelo linear não é adequado (há falta de ajuste)

A hipótese nula de que o modelo linear é adequado é rejeitada se o valor da estatística de teste F_0 for maior que $f_{\alpha, m-k-1, n-m}$, o percentil (1-α) da distribuição F com (m-k-1) graus de liberdade no numerador e (n-m) graus de liberdade no denominador, sendo α o nível de significância do teste. Vale ressaltar que, nessa situação, não se deseja rejeitar a hipótese nula.

Para verificar a qualidade de predição ou capacidade de explicação do modelo, pode se usar o coeficiente de determinação, R^2, que expressa o percentual da variabilidade total da variável resposta que é explicada pelas variáveis explicativas no modelo de regressão. R^2 é dado por:

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT}.$$

A presença de multicolinearidade é verificada a partir do fator de inflação da variância (FIV) para cada variável explicativa. Qualquer FIV maior do que 1 indica algum nível de multicolinearidade; entretanto, em geral, assume-se que o FIV não deve ultrapassar 5 [(4)].

A verificação da adequação das suposições feitas para os erros do modelo de regressão é feita pela análise dos resíduos. Avaliou-se a suposição de que os erros seguem a distribuição

normal com variância constante utilizando os seguintes gráficos: gráfico de probabilidade normal dos resíduos, gráficos de dispersão entre resíduos e variáveis do modelo e gráfico de dispersão entre resíduos e valores preditos pelo modelo. Este último gráfico também foi utilizado para verificar a suposição da linearidade entre a variável resposta e as variáveis explicativas.

Os pontos de alavanca são considerados *outliers* em uma das variáveis, mas não na outra e os pontos influentes são considerados *outliers* nas variáveis resposta e explicativas. A existência de pontos de alavanca e/ou influentes pode exercer muita influência na determinação de R^2, nas estimativas dos coeficientes de regressão e na magnitude de média quadrática dos erros (MQE). No caso do MQE, que é o estimador para a variância do erro, a inflação de seus valores pode afetar as inferências estatísticas feitas para coeficientes e valores preditos pelo modelo. Para detecção de observações influentes, avaliou-se os resíduos studentizados, valores de alavanca (h) e a medida da distância de Cook (D). As observações que possuem valores de resíduos studentizados fora do intervalo [-2, 2], $h > 3(k+1)/n$ e/ou $D > 1$ podem ser possíveis pontos de alavanca e/ou influentes e, por esse motivo, devem ser eliminados para a realização de nova análise de regressão, avaliação da qualidade do ajuste do modelo de regressão e análise de resíduos [5,6].

O modelo linear mais adequado para explicar a variável resposta raio de canto foi um modelo de regressão múltipla. Os coeficientes de regressão foram estimados pelo método dos mínimos quadrados, com o auxílio do *software* Minitab.

3.2.2 Análise de predição do modelo

Como o banco de dados é relativamente grande (n = 22.241), optou-se por utilizar parte dos dados para validar a capacidade de predição do modelo. Por isso, após definição das variáveis explicativas a serem utilizadas no modelo de regressão, avaliou-se os gráficos de dispersão marginal entre a variável resposta e as variáveis explicativas com o objetivo de eliminar os dados que fossem discrepantes nas variáveis explicativas e resposta, para que esses não interferissem na validação do modelo.

Posteriormente, separou-se o conjunto de dados em duas partes: 2/3 dos dados para ser usado como conjunto de ajuste e o restante dos dados para servir de conjunto de teste.

Com os coeficientes estimados pelo conjunto de ajuste, estimou-se os valores da resposta do conjunto de teste e o intervalo de previsão de 95% para a observação individual. Vale mencionar que o intervalo de previsão em um ponto X é sempre mais largo que o intervalo de confiança para a média, pois o intervalo de previsão depende tanto do erro do modelo ajustado como do erro associado às futuras observações [4].

A qualidade de predição do modelo foi medida pelo coeficiente de correlação entre os valores preditos pelo modelo no conjunto de teste e os valores reais da resposta no conjunto de teste.

Avaliou-se, ainda, o Erro Quadrático Médio, que é definido como a média dos resíduos ao quadrado, sendo que o resíduo de uma observação é a diferença entre o valor predito para a resposta pelo modelo e o valor real da resposta.

4. RESULTADOS

4.1 Análise descritiva

A **Figura 2** apresenta a matriz de dispersão da variável resposta raio de canto médio (R*) e as possíveis variáveis explicativas. Os coeficientes de correlação de Pearson entre as variáveis também são apresentados.

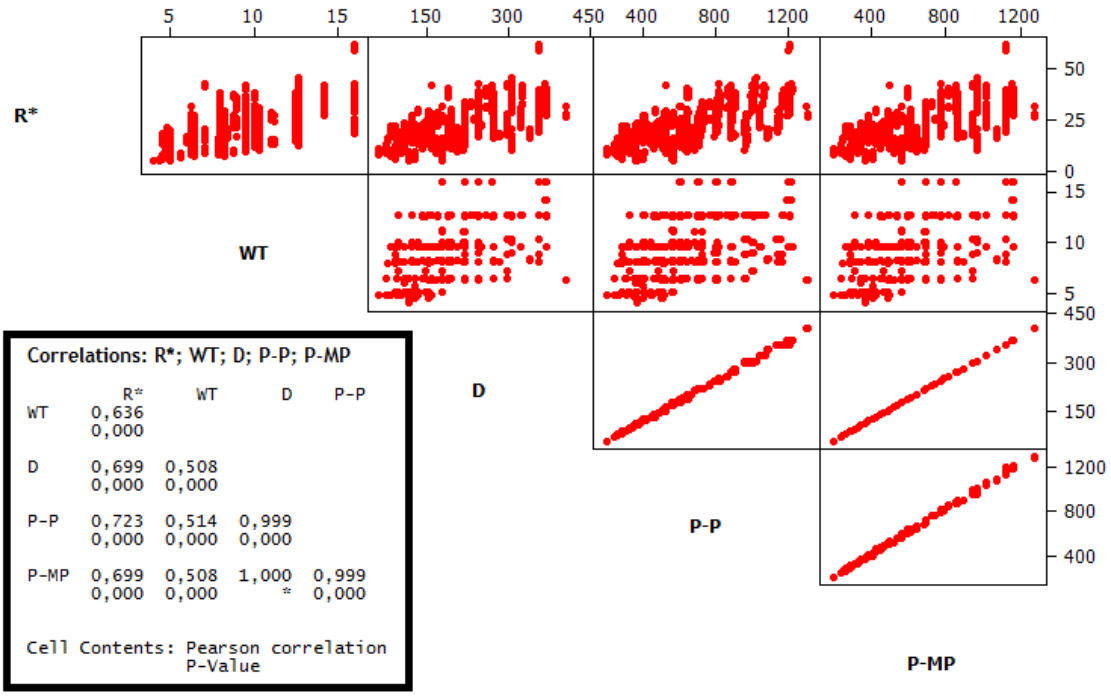


Figura 2 – Matriz de dispersão da variável resposta R* e as possíveis variáveis explicativas associadas às dimensões do perfil e matéria-prima (n = 22.241).

Observa-se que a variável resposta raio de canto médio, R*, apresenta correlação positiva e moderada com todas as outras variáveis analisadas. A partir da avaliação dos coeficientes de correlação, pode-se afirmar que o raio de canto médio (R*) possui maior correlação com as variáveis diâmetro da matéria-prima (D), perímetro da matéria-prima (P-MP), perímetro do perfil (P-P) e espessura de parede (WT).

Apesar da correlação moderada entre R* e as variáveis D, P-MP e P-P, observa-se que essas variáveis explicativas apresentam forte correlação linear entre si. Assim, para evitar o

problema de multicolinearidade, não são indicadas para serem utilizadas conjuntamente no ajuste de regressão e somente uma delas seria adequada para se utilizar no modelo.

Entretanto, para a aplicação desejada, não convém considerar somente a informação do perfil ou da matéria-prima isoladamente. Como já mencionado anteriormente, o raio de canto médio possui uma relação inversamente proporcional com o perímetro da matéria-prima em relação ao perímetro do perfil. Diante disso, foram criadas as variáveis Delta e Razão para expressar a relação entre o perímetro da matéria-prima (P-MP) e o perímetro do perfil (P-P). Sendo assim, a variável Razão foi definida como P-MP / P-P e a variável Delta foi definida como P-MP – P-P.

A **Figura 3** apresenta a matriz de dispersão da variável resposta raio de canto médio (R*) e essas novas variáveis criadas, mantendo na análise a variável espessura de parede (WT).

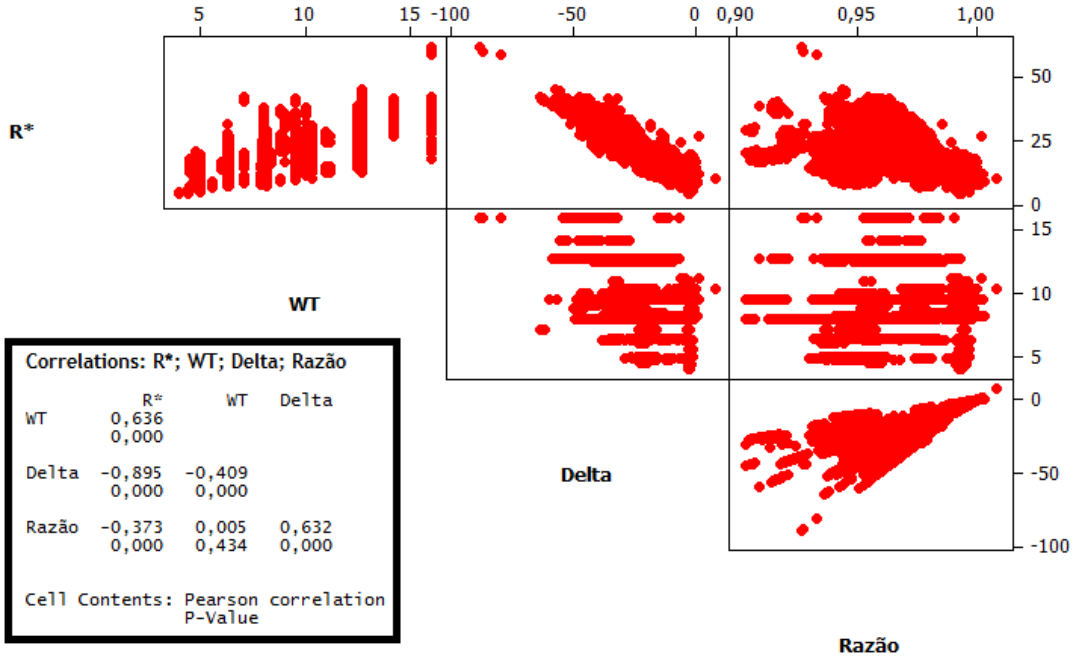


Figura 3 – Matriz de dispersão da variável resposta R* e as novas variáveis criadas (Delta e Razão) (n = 22.241).

Observa-se que a variável resposta raio de canto médio, R*, apresenta forte correlação negativa com a variável Delta criada. Já a variável Razão apresenta baixa correlação negativa com a variável resposta. As variáveis Razão e Delta, por sua vez, apresentam correlação linear moderada.

Diante disso, optou-se por utilizar no modelo de regressão as variáveis explicativas espessura de parede (WT) e diferença entre os perímetros da matéria-prima e perfil (Delta). Acredita-se que o problema de multicolinearidade não interferirá na análise, visto que a correlação entre essas variáveis explicativas é fraca.

Conforme mencionado anteriormente, utilizou-se 2/3 dos dados para realizar o ajuste de regressão (conjunto de ajuste) e o restante para validar a capacidade de predição do modelo (conjunto de teste). Utilizando todo o conjunto de dados observados (conjunto de ajuste mais conjunto de teste), avaliou-se os gráficos de dispersão marginal entre a variável resposta e as variáveis explicativas definidas com o objetivo de eliminar os dados que fossem discrepantes nas variáveis explicativas e resposta, para que isso não interferisse na validação do modelo.

A **Figura 4** e a **Figura 5** apresentam os gráficos de dispersão marginal do raio de canto médio *versus* a diferença entre o perímetro da matéria-prima e do perfil (Delta) e a espessura de parede nominal do perfil (WT), respectivamente.

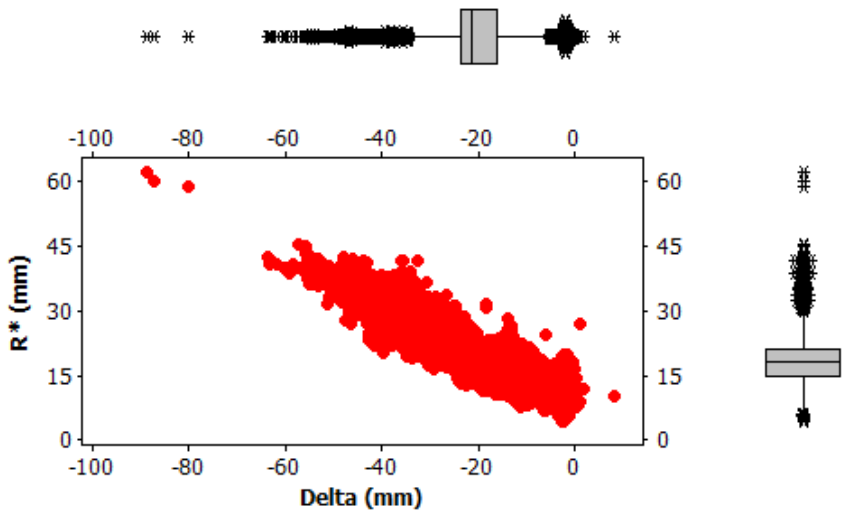


Figura 4 – Dispersão marginal do raio de canto médio *versus* a diferença entre o perímetro da matéria-prima e do perfil (n = 22.241).

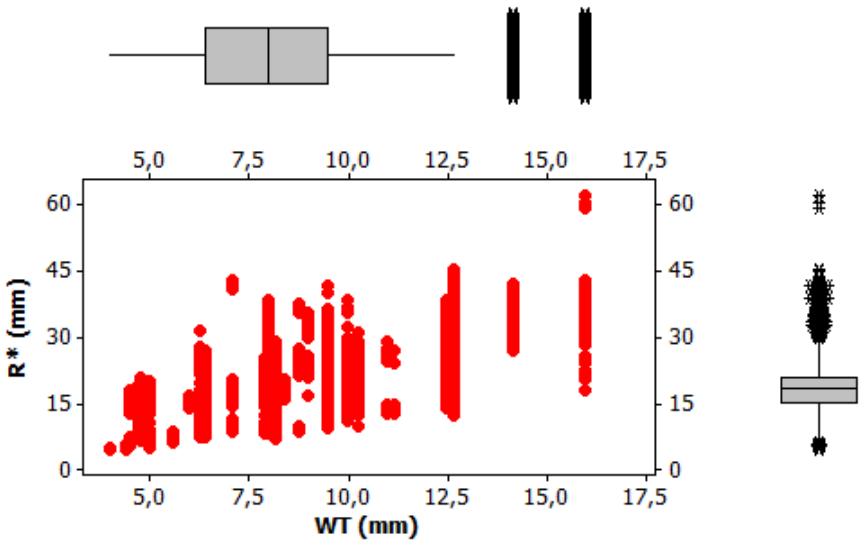


Figura 5 – Dispersão marginal do raio de canto médio *versus* espessura de parede nominal do perfil (n = 22.241).

Observa-se que existem muitos *outliers* tanto na variável resposta quanto explicativas. Entretanto, optou-se por excluir somente 3 pontos que apresentaram raio de canto médio discrepante em relação aos demais, cujos valores eram acima de 50 mm. Valores de raio de canto médio nessa ordem de grandeza não são usuais, sendo que esses 3 pontos excluídos são resultados de um lote de teste em que se utilizou uma matéria-prima com diâmetro externo muito inferior ao necessário para produzir o perfil.

Diante disso, da base inicial com 22.241 elementos removeu-se 3 deles e utilizou-se 14.825 elementos para realizar o ajuste do modelo (conjunto de ajuste) e 7.413 elementos para avaliar a qualidade de predição do modelo ajustado (conjunto de teste).

4.2 Análise de regressão

Ajustou-se o modelo de regressão múltipla com os 14.825 elementos do conjunto de ajuste. Entretanto, a análise dos resíduos indicou que os erros não seguiam a distribuição normal e apresentavam heterocedasticidade, conforme apresentado no Apêndice.

A partir da avaliação dos resíduos studentizados, valores de alavancas (h) e medida de distância de Cook, foi feita a identificação e eliminação de *outliers* e pontos de alavanca/influência. A partir dessa análise, foram retirados 829 elementos que apresentaram valores de resíduos studentizados fora do intervalo [-2, 2], 319 elementos que apresentaram $h > 0{,}0061$ ($3(k+1)/n = 9/14825$), sendo que 54 desses já tinham sido identificados fora do intervalo de aceite para os resíduos studentizados. Não foi identificado nenhum dado com medida de distância de Cook superior a 1.

A **Figura 6** e a **Figura 7** apresentam os 1094 pontos que foram considerados como *outliers* ou influentes na avaliação realizada anteriormente.

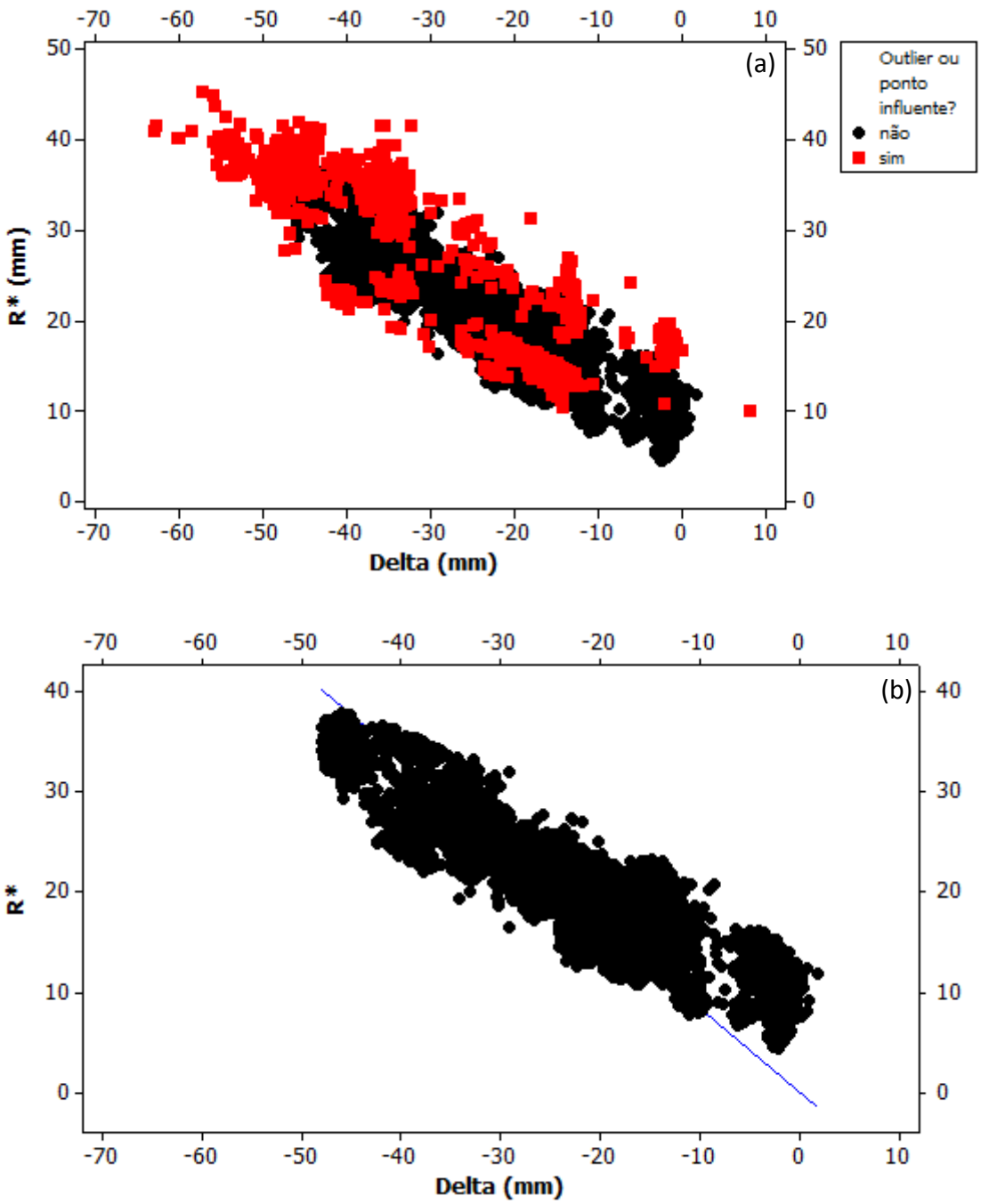


Figura 6 – Gráfico de dispersão do raio de canto médio *versus* diferença entre o perímetro da matéria-prima e do perfil (a) indicando os pontos considerados como *outliers* ou influentes (total de 1094 pontos). (b) Apresentando somente os dados utilizados no modelo (total de 13.713 pontos).

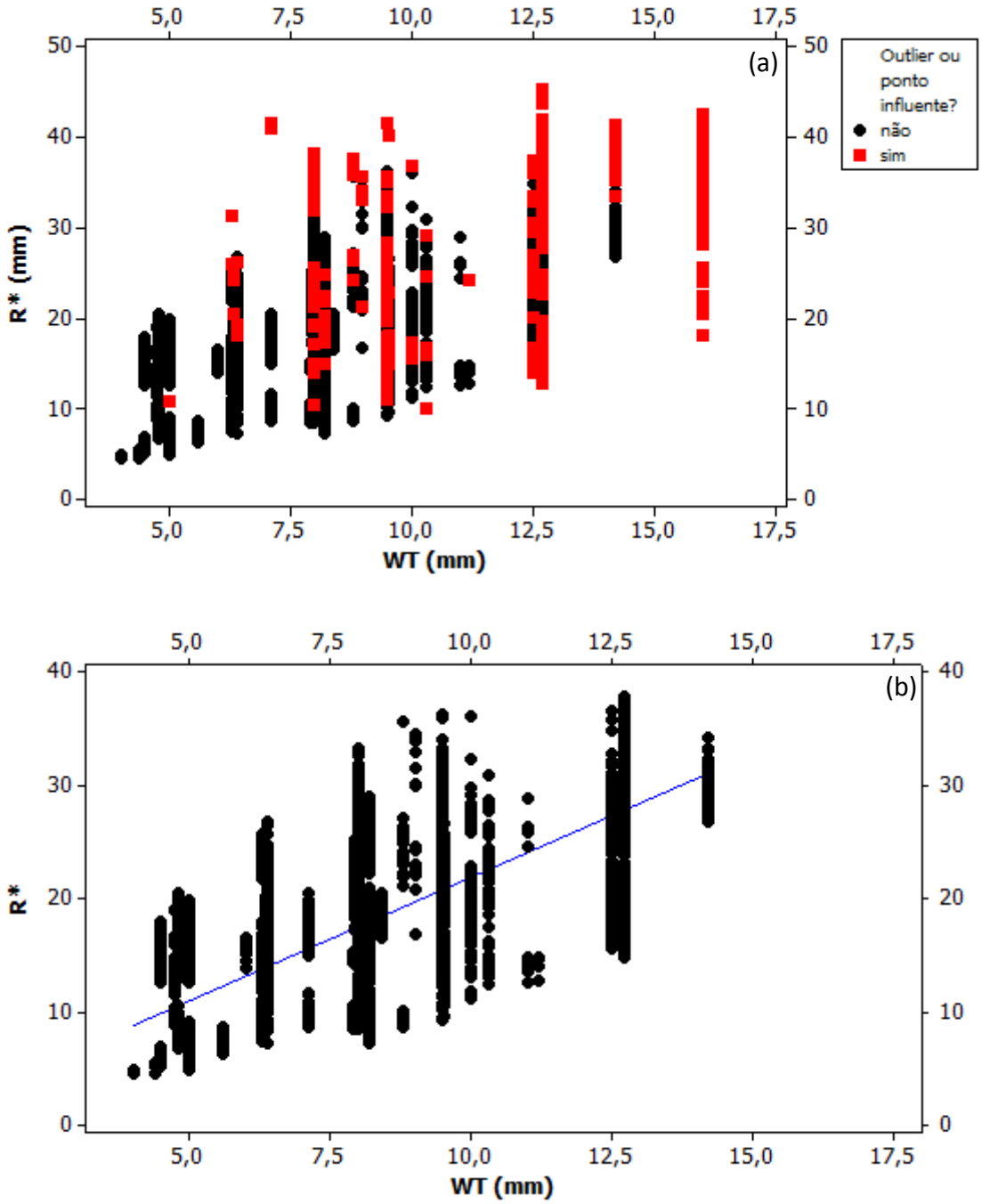


Figura 7 – Gráfico de dispersão do raio de canto médio *versus* espessura nominal do perfil (a) indicando os pontos considerados como *outliers* ou influentes (total de 1094 pontos). (b) Apresentando somente os dados utilizados no modelo (total de 13.713 pontos).

Um novo ajuste do modelo de regressão foi realizado após a eliminação dessas observações. A tabela de análise de variância é apresentada na **Tabela 3** e a análise dos coeficientes é apresentada na **Tabela 4**.

Tabela 3 – Análise de variância para testar a significância da regressão (modelo ajustado sem observações influentes e *outliers*).

Fonte de Variação	Soma dos quadrados	Graus de liberdade	Média quadrática	F_0	P-valor
Regressão	403.415	2	201.708	80.222,18	0,000
Erro	34.517	13.728	3		
Falta de ajuste	29.346	8.480	3	3,51	0,000
Erro puro	5.171	5.248	1		
Total	437.933	13.730			

Tabela 4 – Análise para testar a significância dos coeficientes da regressão (modelo ajustado sem observações influentes e *outliers*).

Termo	Coeficiente	EP Coeficiente	IC 95% Coeficiente	T	P-valor	FIV
Constante	0,60700	0,05606	0,497 ; 0,717	10,83	0,000	
Delta	-0,48735	0,00159	-0,490 ; -0,484	305,35	0,000	1,146
WT	0,92919	0,00698	0,916 ; 0,943	133,14	0,000	1,146

A equação do modelo ajustado é:

$$\hat{Y}_i = 0{,}607 - 0{,}487 Delta_i + 0{,}929 WT_i.$$

O modelo de regressão múltipla apresentou R^2 igual a 92,1%, o que indica que as variáveis utilizadas explicam 92,1% da variabilidade da variável resposta.

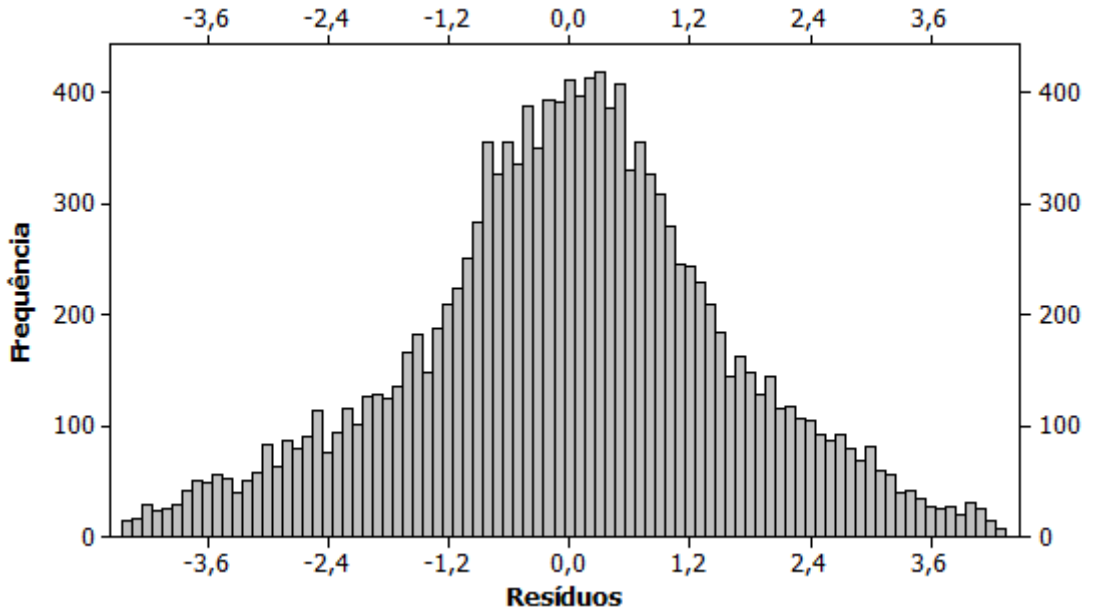
A partir da avaliação da **Tabela 3,** pode-se observar que pelo menos um dos coeficientes de inclinação é estatisticamente significativo ao nível de 5% de significância. Ao avaliar a **Tabela 4,** constata-se que todos os coeficientes são estatisticamente significativos, inclusive o de intercepto.

Não foi identificado problema de multicolinearidade entre as variáveis respostas, pois conforme apresentado na **Tabela 4**, o fator de inflação da variância é menor que 2.

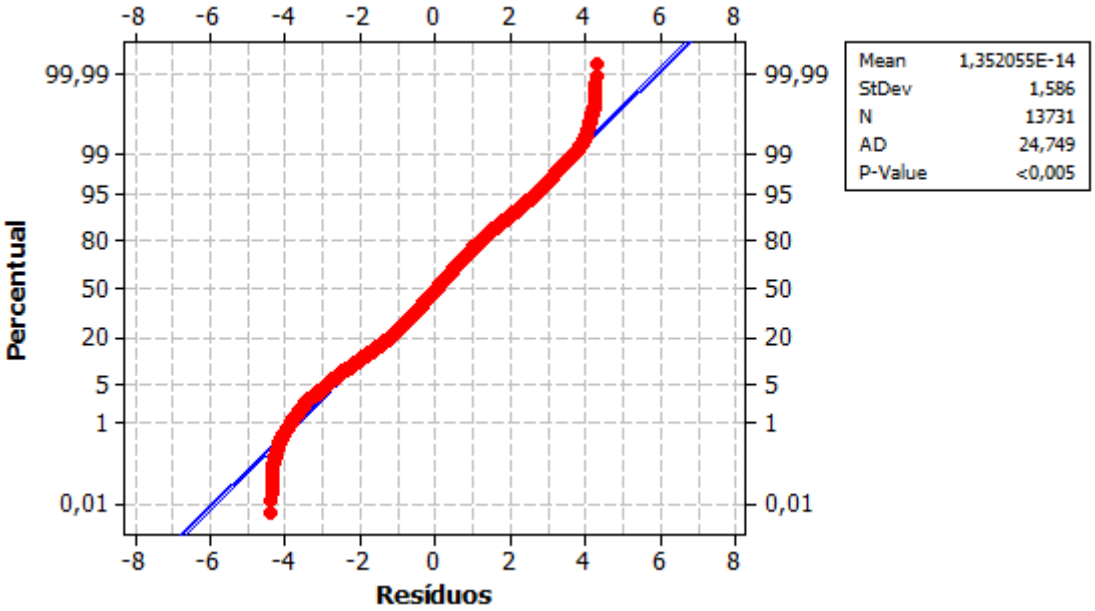
Já o teste para a falta de ajuste apresentou p-valor igual a zero, o que indica que há falta de ajuste, ou seja, o modelo linear não é adequado. Entretanto, como a amostra é muito grande, pode-se supor que qualquer desvio da linearidade, por menor que seja, seria detectado nesse teste.

Em continuidade a análise do modelo de regressão, avaliou-se os resíduos para avaliar a validade das suposições feitas para os erros do modelo (seguir uma distribuição normal com média zero e variância constante; e serem independentes).

Diante disso, foram avaliados o histograma dos resíduos e feito o teste de normalidade dos erros (teste de Anderson-Darling), com a confecção também do papel de probabilidade normal (**Figura 8**). Foram avaliados ainda o gráfico de dispersão dos resíduos *versus* variáveis do modelo (**Figura 9**) e o gráfico de dispersão dos resíduos *versus* valores preditos (**Figura 10**).

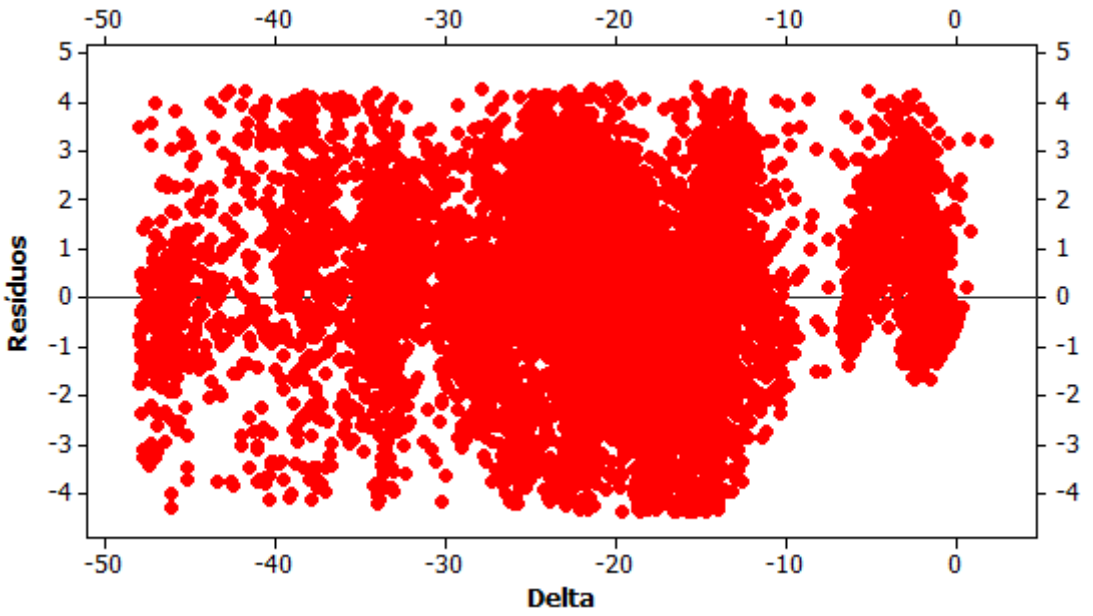


(a)

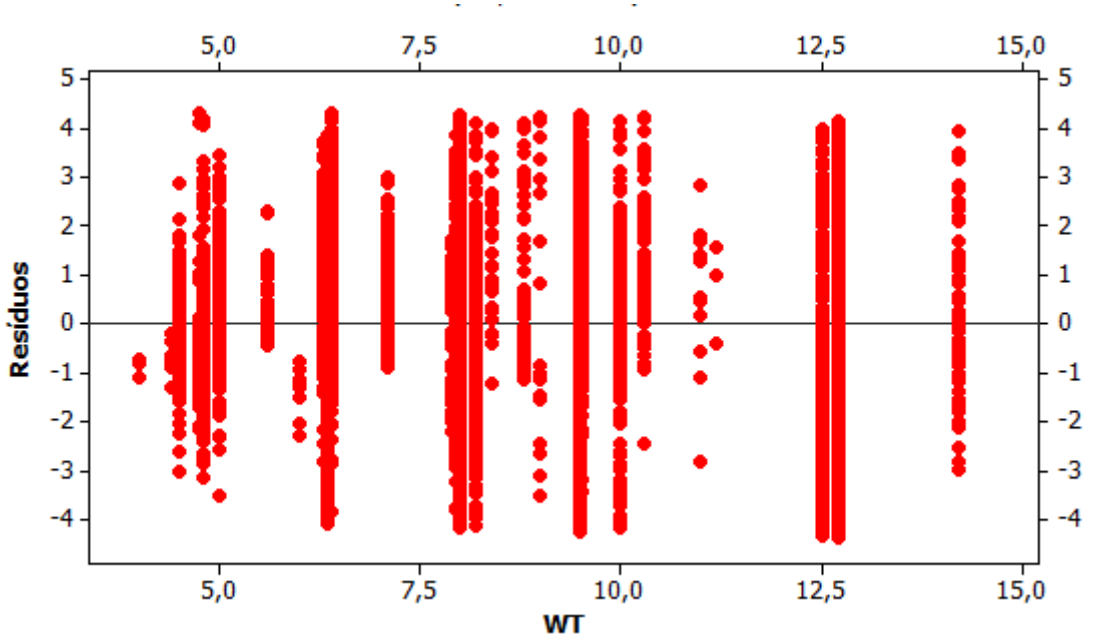


(b)

Figura 8 – Avaliação da suposição da normalidade dos erros via: histograma dos resíduos (a) e gráfico de probabilidade normal dos resíduos com teste de Anderson-Darling (b).



(a)



(b)

Figura 9 – Gráfico de dispersão entre resíduos e variáveis do modelo: Delta (a) e WT (b).

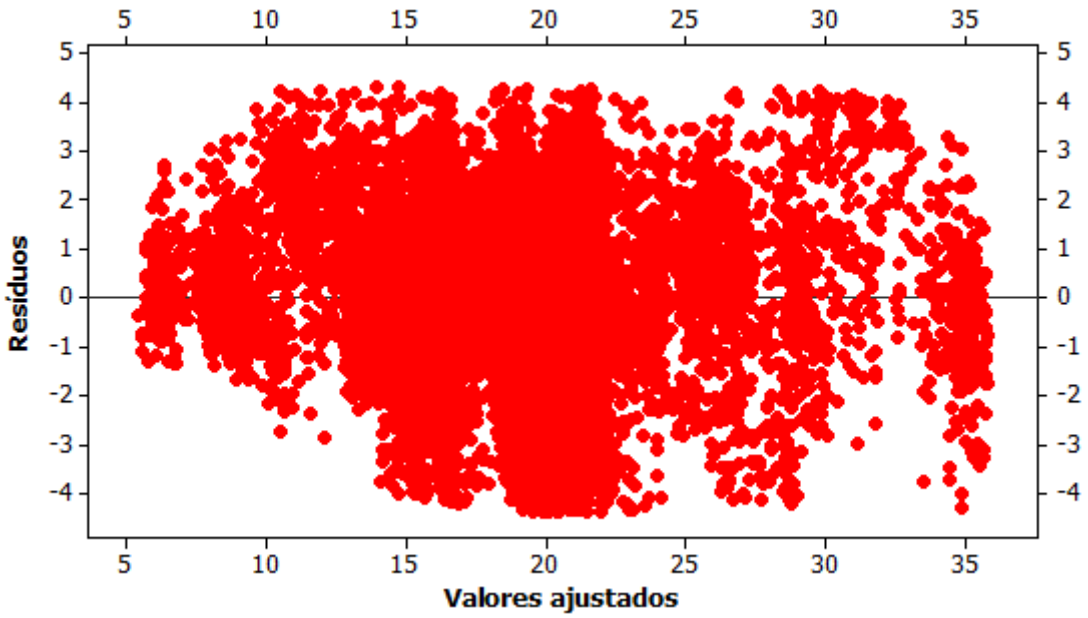


Figura 10 – Gráfico de dispersão entre resíduos e valores ajustados.

O teste de normalidade dos erros apresenta evidências de que os erros não seguem a distribuição normal ao nível de significância de 5%, conforme apresentado na **Figura 8** (p-valor < 0,005). O exame da **Figura 8** mostra que esse desvio da normalidade se dá provavelmente porque a distribuição dos erros deve ter caudas mais pesadas do que as da Distribuição Normal, mas ainda é simétrica. Sendo assim, dado que o teste F é robusto a pequenos desvios da normalidade e como a amostra é grande, acredita-se que isto não tenha interferido nas inferências estatísticas feitas pelos testes F e t-Student.

Pela análise do gráfico de dispersão dos resíduos contra variáveis do modelo (**Figura 9**), observa-se que os dados estão uniformemente distribuídos em torno de zero, o que indica que a suposição de linearidade do modelo pode ser considerada válida. Na **Figura 9(a)**, pode-se perceber que a variância dos resíduos parece ser menor para valores de Delta mais próximos de zero. Isto poderia causar uma subestimação no valor da variância do erro. No entanto, por se tratar de uma pequena parte do conjunto de dados e visto que a parte da variabilidade da resposta que é explicada pelo modelo de regressão é muito maior do que a parte devido ao erro, acredita-se que essa pequena violação na suposição de variância constante não interferiu de modo importante nas inferências feitas pelo modelo.

Por fim, o gráfico de dispersão dos resíduos *versus* valores preditos (**Figura 10**) mostra que os resíduos estão uniformemente distribuídos em torno de zero e ,apesar de apresentarem uma tendência de menores valores para menores valores ajustados, a suposição de linearidade do modelo pode ser considerada válida por ser uma região pequena se comparada com todo o conjunto de dados e em função do elevado valor de R^2 do modelo.

Interpretando os coeficientes do modelo resultante, pode-se afirmar que, para um perfil com espessura de parede nominal fixa, o aumento de 1 mm na diferença entre o perímetro da

matéria-prima e o perímetro do perfil (Delta) reduz, em média, 0,49 mm o raio de canto médio. Pode-se afirmar ainda que, para um perfil com diferença entre o perímetro de matéria-prima e perfil (Delta) fixo, o aumento de 1 mm na espessura de parede do perfil aumenta, em média, 0,93 mm o raio de canto médio.

Na próxima seção, o modelo estimado nesta seção será aplicado aos elementos do conjunto de teste para estimar o raio de canto médio desses elementos. Desse modo, será avaliada a qualidade de predição do modelo.

4.3 Análise da qualidade de predição do modelo

Com os coeficientes estimados pelo conjunto de ajuste, estimaram-se os valores da resposta do conjunto de teste e o intervalo de predição de 95% para a observação individual. De um total de 7.413 elementos do conjunto de teste, observou-se que, para 6513 elementos, o intervalo de predição de 95% para a observação individual continha o valor real da resposta para o elemento, o que equivale a uma taxa de acerto 87,86% para esse conjunto de teste.

A **Figura 11** apresenta o gráfico de dispersão dos valores ajustados *versus* os valores observados no conjunto de teste.

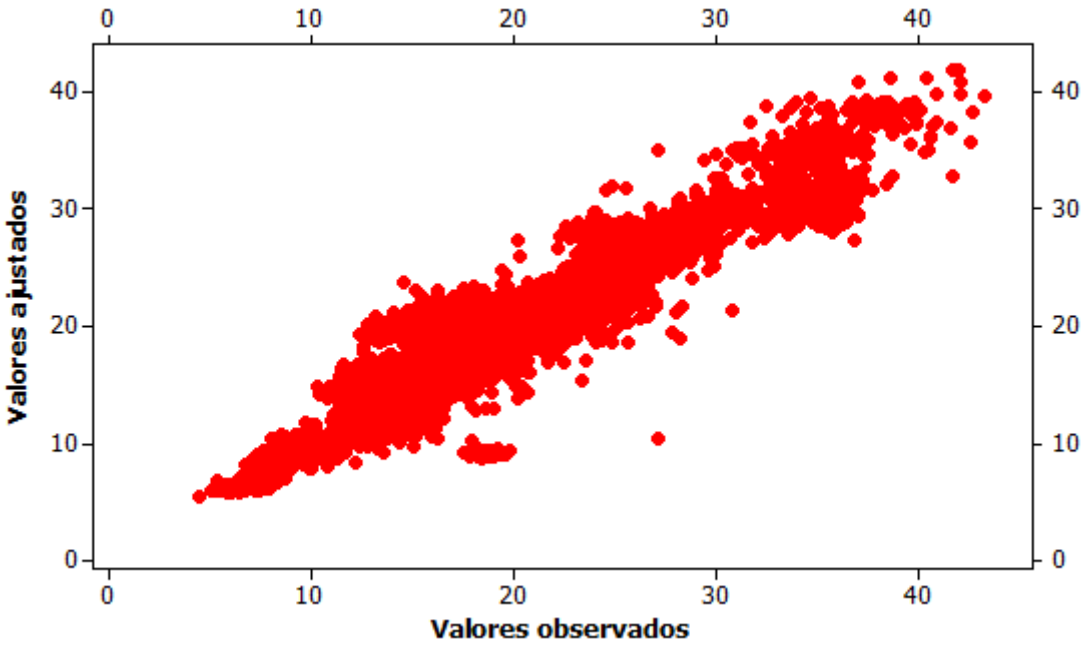


Figura 11 – Gráfico de dispersão valores ajustados *versus* valores observados no conjunto de teste (n = 7.413).

Na **Figura 11,** observa-se que os valores preditos e os observados no conjunto de teste apresentam forte correlação positiva. A qualidade de predição do modelo foi medida pelo coeficiente de correlação entre os valores preditos pelo modelo no conjunto de teste e os

valores reais observados da resposta no conjunto de teste. O coeficiente de correlação é igual a 0,941, o que indica que o modelo apresenta alta capacidade para prever a variável resposta raio de canto médio.

Além disso, o Erro Quadrático Médio (média dos erros ao quadrado) é igual a 4,65, sendo, portanto, um valor baixo, o que corrobora para a suposição de que o modelo de regressão múltipla definido no item 4.2 é adequado para prever com precisão o raio de canto médio de perfis quadrados e retangulares.

O modelo foi elaborado para perfis com espessura de parede de 4 mm a 14,2 mm e Delta variando de -47,95 mm a 1,77 mm. Assim, após eliminação dos dados considerados como extrapolação, com os 7.257 dados desse novo conjunto de teste, avaliou-se novamente o coeficiente de correlação que apresentou valor igual a 0,932 e Erro Quadrático Médio igual a 4,57. Além disso, o intervalo de predição de 95% para a média do raio de canto apresentou uma taxa de acerto igual a 88,15%.

5. CONCLUSÃO

Esse trabalho teve como objetivo construir um modelo de regressão múltipla para predizer o valor esperado para a característica média dos quatro raios de canto de perfis quadrados e retangulares em função de dimensões do perfil e do tubo circular utilizado como matéria-prima. Diante dos resultados apresentados, pode-se concluir que:

- A diferença do perímetro da matéria-prima e do perfil (Delta) e a espessura de parede (WT) foram as variáveis adequadas para ajustar um modelo de regressão múltipla para prever o raio de canto médio de perfis.
- As variáveis utilizadas no modelo de regressão múltipla explicam 92,1% da variabilidade da variável resposta.
- O modelo ajustado indicou que, para um perfil com espessura de parede nominal fixa, o aumento de 1 mm na diferença entre o perímetro da matéria-prima e o perímetro do perfil (Delta) reduz, em média, 0,49 mm o raio de canto médio. E, para um perfil com diferença entre o perímetro de matéria-prima e perfil (Delta) fixo, o aumento de 1 mm na espessura de parede do perfil aumenta, em média, 0,93 mm o raio de canto médio.
- O teste para a falta de ajuste indicou que o modelo linear não é adequado. Entretanto, como a amostra é muito grande, pode-se supor que qualquer desvio da linearidade, por menor que seja, seria detectado nesse teste. Além disso, o gráfico de resíduos *versus* predito pelo modelo mostrou que a suposição de linearidade do modelo pode ser considerada válida.
- A elevada correlação positiva (0,941) entre os valores preditos pelo modelo no conjunto de teste e os valores reais observados e o baixo valor de Erro Quadrático

Médio (4,65) indicam que o modelo de regressão múltipla é adequado para prever o raio de canto médio de perfis quadrados e retangulares.

Uma parcela do erro do modelo de regressão múltiplo ajustado pode ser associada com o uso de valores nominais para o diâmetro da matéria-prima e espessura de parede do perfil, visto que não existia no banco de dados os valores reais para essas características. Além disso, os erros na medição do raio de canto e comprimento dos lados, associados com o erro do instrumento de medição, fator humano, condições ambientais, erros de digitação, entre outros, são fatores que colaboram para a parcela da variabilidade de Y não explicada pelas variáveis explicativas do modelo.

Para evitar problemas de extrapolação, o modelo é indicado para perfis com espessura de parede de 4 mm a 14,2 mm e Delta variando de -47,95 mm a 1,77 mm.

6. CONSIDERAÇÕES FINAIS

Além do modelo apresentado nesse trabalho, avaliou-se também o modelo quadrático para a relação com a espessura de parede (WT) e outro para adicionar o termo de interação ao modelo. Essas tentativas não melhoraram a qualidade do ajuste, além de terem deixado o modelo mais completo.

Para a elaboração de trabalhos futuros, sugere-se:

- utilizar modelos mais sofisticados na análise de regressão para melhor se adequar aos dados, já que os erros mostraram ter distribuição com caudas mais espessas do que as da a distribuição normal.
- modelar a variabilidade da característica raio de canto em conjunto com a média dessa característica, ou seja, criar um modelo de regressão com resposta multivariada.

Referências

1 – ASTM A500 / A500M-13, Standard Specification for Cold-Formed Welded and Seamless Carbon Steel Structural Tubing in Rounds and Shapes, ASTM International, West Conshohocken, PA, 2013.

2 – VALLOUREC & MANNESMANN TUBES. Perfis MSH de seções circulares, quadradas e retangulares. Disponível em: <https://edisciplinas.usp.br/pluginfile.php/224159/mod_resource/content/1/dimens%C3%B5es-valores-est%C3%A1ticos.pdf>. Acessado em: 30 de novembro de 2017.

3 – MORDINI, F. Structural Hollow Sections from an Engineer's perspective: What size should I use? Disponível em: <http://steelconstruction.org.za/structural-hollow-sections-from-an-engineers-perspective-what-size-should-i-use/>. Acessado em: 01 de dezembro de 2017.

4 - MONTGOMERY, D. C.; RUNGER, G. C. **Estatística aplicada e probabilidade para engenheiros**. 5 ed. LTC, 2012.

5 – CHATTERJEE, S.; SIMONOFF, J.S. **Handbook of Regression Analysis**. John Wiley & Sons, INC., 2013.

6 – DM&P. **Apostila de Métodos estatísticos para P&D**, 2012.

Apêndice

1. Saídas do Minitab para a análise de regressão múltipla antes da avaliação de pontos influentes e/ou de alavanca.

Regression Analysis: R* versus Delta; WT

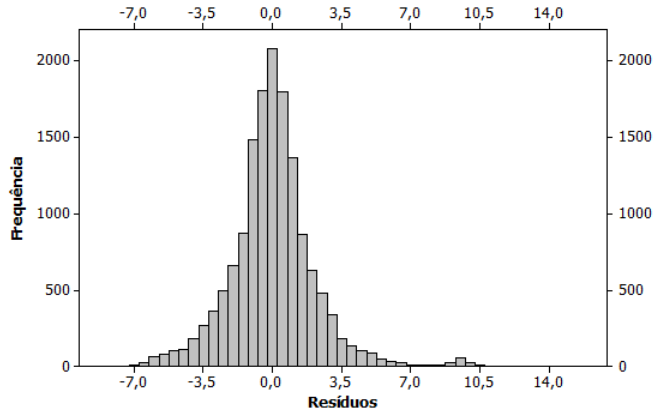
The regression equation is
R* = 0,985 - 0,488 Delta + 0,888 WT

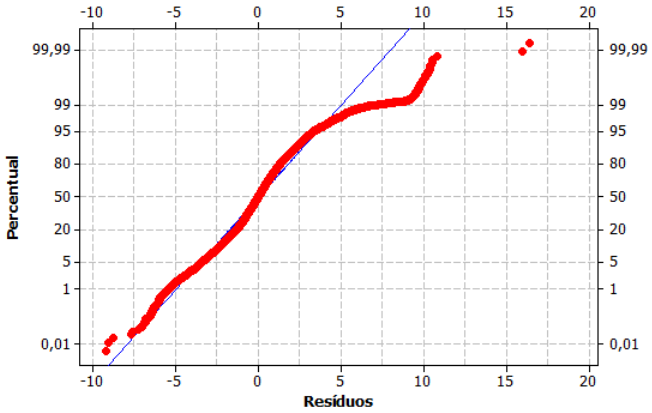
Predictor	Coef	SE Coef	T	P	VIF
Constant	0,98485	0,06637	14,84	0,000	
Delta	-0,487919	0,001929	-252,96	0,000	1,201
WT	0,888302	0,008181	108,58	0,000	1,201

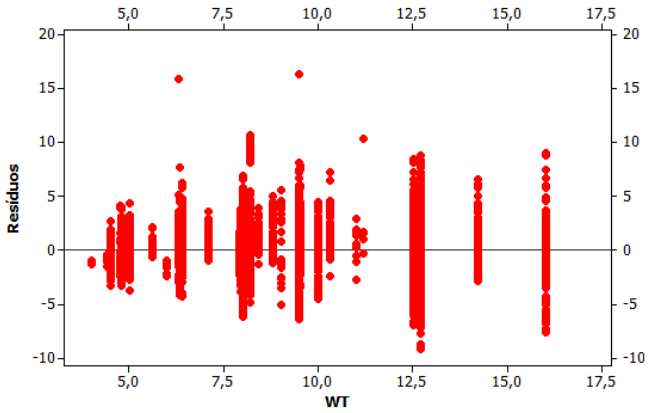
S = 2,13342 R-Sq = 88,8% R-Sq(adj) = 88,8%

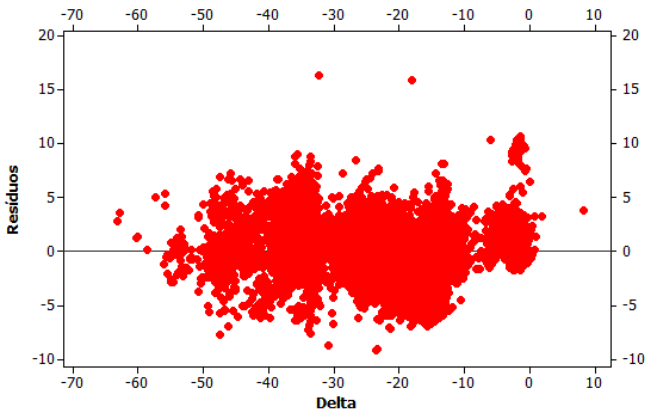
Analysis of Variance

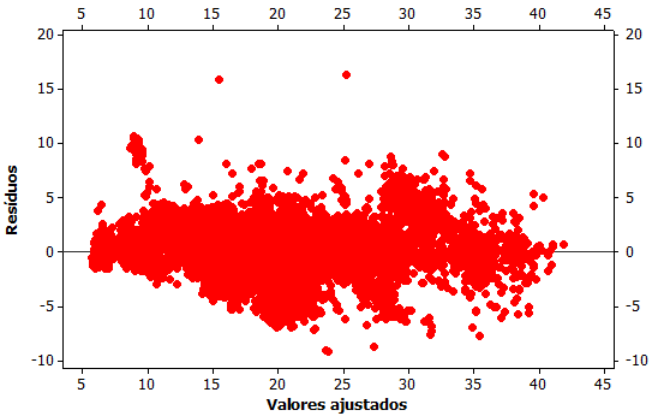
Source	DF	SS	MS	F	P
Regression	2	537115	268557	59004,61	0,000
Residual Error	14822	67462	5		
Lack of Fit	9279	54906	6	2,61	0,000
Pure Error	5543	12555	2		
Total	14824	604576			











2. Histograma do erro quadrático médio considerando todos os dados do conjunto de teste.