Universidade Federal de São Carlos
Centro de Ciências Exatas e de tecnologia
Departamento de Estatística

SELEÇÃO DE VARIÁVEIS: UMA APLICAÇÃO A DADOS DE MOINHO DE CIMENTO

Lissa Kido Higashizawa

Trabalho de Conclusão de Curso

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Seleção de variáveis: uma aplicação a dados de moinho de cimento

Lissa Kido Higashizawa
Orientador: Profa. Dra. Daiane Aparecida Zuanetti (DEs - UFSCar)
Coorientador: Profa. Dra. Rosineide Fernando da Paz (UFC)

Trabalho de Conclusão de Curso a ser apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
20 de Dezembro de 2019

Agradecimentos

Gostaria de deixar o meu profundo agradecimento às professoras orientadoras Profa. Dra. Daiane Aparecida Zuanetti e Profa. Dra. Rosineide Fernando da Paz por todo o apoio neste Trabalho de Conclusão de Curso.

Gostaria de agradecer também a parceria de Rilmar Farias Mendes da Silva [1] e seu orientador Prof. Dr. Dmontier Pinheiro Aragão Júnior [2] por possibilitar a realização deste trabalho.

Agradeço a todos os funcionários da instituição de ensino UFSCar por todo apoio e por proporcionarem um ambiente propício para o desenvolvimento do meu Trabalho de Conclusão de Curso.

Por fim, agradeço a minha mãe, que apesar de todas as dificuldades, me ajudou na realização do meu sonho.

[1] Estudante de Engenharia de Produção na Universidade Federal do Ceará (UFC).
[2] Professor e pesquisador da UFC, Doutor em Engenharia de Produção.

Resumo

Tendo como objeto de estudo um determinado moinho que produz cimento, utilizamos duas técnicas de seleção de variáveis, LASSO e *stepwise*, para identificar variáveis que influenciam na potência do motor e que, consequentemente, impactam a produção de cimento. Também consideramos as covariáveis defasadas em 4 momentos diferentes e realizamos a análise de diagnóstico para os modelos estimados com identificação de pontos influentes.

Entre todos os modelos analisados, escolhemos a seleção que foi feita pelo *stepwise* sem pontos influentes e sem defasagens no tempo, que apresenta o menor valor para os critérios de escolha função de risco, AIC e BIC.

Palavras-chave: *LASSO, moinho de cimento, seleção de variáveis, stepwise.*

Sumário

Lista de Tabelas

Lista de Figuras

Capítulo 1

Introdução

O processo de cominuição (fragmentação) de cimento é realizado através dos moinhos verticais de rolos, com capacidade produtiva de até 500t/h (Jorgensen, 2004). Uma das empresas fabricantes de cimento que utiliza esse tipo de moinho é a empresa Cimento Apodi [1] (Apodi, 2011) e existe o interesse em saber como a produção de cimento pode ser melhorada e aumentada.

Um dos objetivos da empresa no momento é prever os valores de algumas variáveis de interesse, como por exemplo, a potência do motor. Essas variáveis, chamadas de variáveis respostas, podem ser explicadas ou previstas por um conjunto de outras variáveis, denominadas covariáveis ou variáveis explicativas. Essas principais covariáveis devem ser selecionadas entre várias condições e características de funcionamento do moinho, pois algumas delas possibilitam uma grande melhora no seu desempenho caso sejam feitas as alterações apropriadas em seus valores.

A empresa atualmente já realiza um processo que eles denominam de "stepteste" que consiste em selecionar algumas variáveis, entre todas as disponíveis, por meio da experiência dos envolvidos no processo e checar via algumas medidas de estatística exploratória se a variável realmente influencia ou não na variável resposta, potência do motor, por exemplo, que será considerada nesse trabalho.

Esse processo de seleção, além de demorado, pode não fornecer o conjunto "ideal" de variáveis para maximizar a eficiência da produção de cimento, pois não é possível testar todas as combinações de variáveis existentes e considerar o efeito conjunto delas. Assim, para esse problema em específico, é necessário especificar uma metodologia mais

[1] Fabricante de cimentos, concretos e argamassas, a empresa é uma *joint venture* multinacional que atualmente está presente nas regiões Norte e Nordeste do Brasil.

quantitativa e objetiva e, ao mesmo tempo, tornar o processo de seleção de variáveis mais eficiente. A empresa geralmente utiliza as variáveis selecionadas como informações em algoritmos de otimização (SVM, *support vector machine*, e rede neural MLP, *multilayer perceptron* - Izbicki e Santos 2018; Pal e Mitra 1992) para atingir um funcionamento mais eficiente do moinho.

Outra suspeita da empresa, e que será considerada nesse trabalho, é que as variáveis que influenciam na potência do motor possuem uma defasagem no tempo, ou seja, são as condições anteriores (de segundos ou minutos atrás) do funcionamento do moinho que determinam a potência do motor em cada momento observado.

Dessa forma, os objetivos principais desse estudo é selecionar covariáveis que influenciam a produção e a potência do motor e identificar a defasagem das covariáveis no tempo que mais influenciam as variáveis principais do processo. Ao final do projeto, é esperado que haja uma estratégia de seleção de variáveis que pode ser adotada pela empresa para melhorar suas previsões e tornar a produção de cimento mais eficiente, além de uma análise completa de influência de diversas covariáveis na variável resposta potência do motor.

Outras variáveis poderiam ter sido utilizadas como variável resposta, por exemplo vibração do moinho. Mas, como essa variável muda seu valor em um espaço de tempo muito pequeno, a coleta de dados deveria ser feita em intervalos de tempo ainda menor do que foi realizado para esse trabalho e a potência do motor é atualmente a variável de maior interesse da empresa.

As metodologias para seleção de variáveis consideradas e comparadas nesse trabalho são o *Least Absolute Shrinkage and Selection Operator* (LASSO, Tibshirani 1996) e *Stepwise* (James *et al.*, 2013). O método *stepwise* associado à regressão linear foi escolhido por ser um dos métodos mais tradicionais de seleção de variáveis e o LASSO por ter despontado na última década como uma alternativa eficiente ao primeiro. A comparação entre os métodos é realizada via Função de Risco (Izbicki e Santos, 2018), AIC (Akaike, 1974) e BIC (Schwarz *et al.*, 1978). O impacto de pontos atípicos na seleção das variáveis e consequente poder preditivo é analisado via análise de resíduos e medida de distância de Cook (Cook, 1977).

Esse relatório está organizado como segue: o Capítulo 2 aborda com mais detalhes as metodologias utilizadas ao longo do trabalho, o Capítulo 3 mostra a aplicação dessas metodologias no banco de dados da empresa e no Capítulo 4, apresentamos as conclusões

feitas em relação aos dados do moinho de cimento e possíveis estudos futuros. Os códigos utilizados neste trabalho estão registrados no Apêndice A.

Capítulo 2

Métodos de seleção e validação de modelos

Nesse capítulo faremos uma descrição das metodologias que serão estudadas e aplicadas nesse trabalho.

2.1 Seleção de Variáveis

2.1.1 *Least Absolute Selection and Shrinkage Operator*

O *Least Absolute Selection and Shrinkage Operator* (LASSO) é um método que seleciona covariáveis a partir da estimação dos coeficientes da regressão linear, forçando, por meio de uma restrição adicional no modelo, que as estimativas da maioria desses parâmetros seja zero. Seja o modelo de regressão (James *et al.*, 2013)

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i, \tag{2.1}$$

em que, y_i é a i-ésima observação da variável resposta, x_{ij} é a i-ésima observação da j-ésima covariável, β_0 é o intercepto, β_j é o j-ésimo coeficiente, p é a quantidade de covariáveis a serem consideradas e ϵ_i é o erro aleatório da i-ésima observação.

As estimativas dos valores de $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ são dadas por (Feng *et al.*, 2012)

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \tag{2.2}$$

sendo n o número de observações. Para forçar os valores das estimativas dos coeficientes

para zero, a equação estará sujeita a restrição $\sum_{j=1}^{p} |\beta_j| < t$, denominada Regularização L1, com $t > 0$. Quanto mais aumenta-se o valor de t, maior será a quantidade de parâmetros estimados como zero.

A escolha do parâmetro de regularização t

A seleção das variáveis no modelo e a estimação dos valores dos coeficientes da regressão dependem do valor pré-fixado para t. Na prática, desconhecemos o valor de t e buscamos utilizar o valor que possibilita a seleção de variáveis e estimativas dos parâmetros cujo modelo apresenta a melhor predição. Alguns métodos tem sido propostos na literatura para essa escolha, mas um dos mais utilizados e eficientes é escolhê-lo por validação cruzada no método k-fold (James *et al.*, 2013).

Esse método consiste em separar a amostra em k subconjuntos, geralmente $k = 5$, em que um deles em cada uma de k estimações é escolhido para ser a amostra de validação e os outros de treinamento. Para vários e diferentes valores de t, estimamos o modelo representado na Equação (2.1) com a amostra de treinamento. Com base no modelo estimado, prevemos o valor de y_i para cada indivíduo da base de validação e calculamos o erro total de predição cometido com o modelo. Para cada valor de t, esse processo é realizado k vezes, em que cada subconjunto é selecionado uma vez para ser a amostra de validação. O valor de t escolhido é o valor que minimiza o erro total cometido.

2.1.2 *Stepwise* na regressão linear

O *stepwise* é um método de seleção de variáveis que as inclui ou exclui do modelo (aqui vamos considerar um modelo de regressão linear múltiplo como na Equação (2.1)) observando o quanto cada uma agrega de informação no modelo estimado (James *et al.*, 2013). Seja p o número de variáveis, β_j o parâmetro da j-ésima variável, α_e o nível de significância para entrada de variáveis no modelo e α_s o nível de significância para saída. Apesar do método escolhido nesse trabalho incluir e excluir variáveis usando sucessivos testes de hipóteses, devemos ter muito cuidado ao interpretá-los porque esses testes são individuais. Assim, segue uma série de passos para a aplicação do método:

1. Ajustam-se p modelos, cada um com cada uma das variáveis explicativas, testa-se $H_0 : \beta_1 = 0$ e obtêm-se o p-valor. Se pelo menos um deles for menor que α_e, adiciona-se no modelo a variável cujo p-valor for o menor de todos.

2. Ajustam-se $p-1$ modelos com cada uma das variáveis restantes considerando no modelo a variável incluída no passo anterior, testa-se $H_0 : \beta_2 = 0$ e obtêm-se o p-valor. Se pelo menos um deles for menor que α_e, adiciona-se no modelo a variável cujo p-valor for o menor de todos.

3. Adicionada a segunda variável no modelo, testa-se novamente $H_0 : \beta_1 = 0$ e compara-se com o α_s. Se p-valor$> \alpha_s$ retira-se a primeira variável do modelo.

4. Continua-se assim o processo até que não haja nenhuma inclusão ou exclusão de variáveis.

O método pode ser implementado no SAS através da função *proc glmselect*. Escolhemos o SAS, pois o procedimento é feito pelo p-valor, além de possibilitar a escolha do nível de significância de entrada e saída.

2.2 Medidas para comparação de modelos

Para comparar os diferentes métodos utilizados na seleção e estimação do modelo e analisar a qualidade do modelo ajustado para predição, contamos com o auxílio da validação cruzada no método *Holdout* (Kohavi *et al.*, 1995). Esse método consiste em dividir as observações do banco de dados em duas partes, denominados treino e validação. A separação deve ocorrer de forma aleatória, em geral contendo 70% e 30% das observações, respectivamente. Aplicamos as metodologias de seleção de variáveis no banco de treino e o cálculo da função de risco, AIC e BIC foram feitos no banco de validação, para assim escolher o melhor modelo.

2.2.1 Função de Risco

Seja $g(\mathbf{x})$ a função de predição. A função de risco $R(g)$ é definido como (Izbicki e Santos, 2018)

$$R(g) = E[(y - g(\mathbf{x}))^{\mathbf{2}}], \tag{2.3}$$

em que $(\mathbf{x}, y)$ são os valores das observações não utilizadas na estimação de $g(\mathbf{x})$. A função de risco definida na Equação (2.3) determina o quão bem $g(\mathbf{x})$ está predizendo os valores

da variável resposta. Como temos $g(\mathbf{x})$ diferente para cada modelo, temos diferentes $R(g)$ para cada um deles.

Se o valor de $R(g)$ for baixo, então o modelo correspondente possui um bom poder preditivo, e ao compararmos os valores da função de risco, a escolha pelo melhor modelo se dá naquele que possuir o menor $R(g)$.

Na prática, o valor da função de risco é estimado pelo cálculo do erro médio de predição

$$\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{n}.$$

2.2.2 Critério de Informação de Akaike (AIC)

O AIC é um método que seleciona o melhor modelo dentre os testados pelo pesquisador, cujo modelo selecionado descreve melhor a variável resposta (Akaike, 1974). A seleção se baseia na Equação (2.4)

$$AIC = -2ln(L(\hat{\boldsymbol{\theta}}|\mathbf{x})) + 2p, \tag{2.4}$$

em que $L(\hat{\boldsymbol{\theta}}|\mathbf{x}) = f(\mathbf{x}|\hat{\boldsymbol{\theta}})$ é a função de máxima verossimilhança com $\hat{\boldsymbol{\theta}}$ sendo as estimativas de máxima verossimilhança dos parâmetros do modelo, $\mathbf{x}$ representa os dados $(\mathbf{x}, y)$ e p o número de parâmetros a serem estimados. Para cada modelo testado haverá um valor de AIC correspondente, cuja escolha do melhor modelo é daquele que obtiver o menor valor de AIC.

É importante notar que a penalização $2p$ desfavorece modelos que possuem muitos parâmetros, optando por escolher modelos mais simples que facilitam a interpretação dos coeficientes do modelo no problema em questão. O cálculo do critério depende de uma distribuição para y que originalmente é desconhecida, mas que em um modelo de regressão linear é geralmente assumida como normal.

2.2.3 Critério de Informação Bayesiano (BIC)

O BIC é um critério de seleção de modelos que utiliza a função de verossimilhança para avaliar o ajuste, como o AIC, discutido na Seção 2.2.2. Porém, há diferenças na penalização como mostra a Equação (2.5) (Schwarz *et al.*, 1978)

$$BIC = -2ln(L(\hat{\boldsymbol{\theta}}|\mathbf{x})) + ln(n)p, \tag{2.5}$$

em que n é o número de observações. A escolha do melhor modelo é daquele que obtiver o menor valor de BIC e o mesmo possui característica semelhante ao AIC de desfavorecer modelos com muitos parâmetros e depender de uma distribuição desconhecida.

2.3 Análise de Resíduos e identificação de pontos influentes

Neste trabalho, temos que verificar duas suposições referente ao modelo linear para a aplicação do *stepwise*, a normalidade e homocedasticidade dos erros, e para o LASSO temos que verificar somente a homocedasticidade dos erros. Além disso, há a verificação de pontos influentes que podem modificar significativamente na escolha das variáveis.

A normalidade dos erros pode ser verificada por meio de uma análise visual do Gráfico Quantil-Quantil (QQ-plot), em que a normalidade aparece quando os pontos formam uma linha reta. O teste que utilizamos para confirmar a normalidade foi o teste de Shapiro-Wilk. Para mais detalhes, consulte Wilk e Gnanadesikan (1968) e Shapiro e Wilk (1965).

A visualização da homocedasticidade ocorre quando os resíduos não apresentam tendências e estão aleatoriamente distribuídos em torno do zero, podendo ser observado no gráfico de resíduos por valores ajustados (Neter *et al.*, 1996). O teste utilizado foi o Teste F para duas variâncias (Morettin e BUSSAB, 2017), em que os resíduos são separados em dois grupos usando a mediana como critério, possibilitando a realização do teste tendo como base esta divisão. Se o teste rejeitar a hipótese nula, significa que os dois grupos de resíduos possuem variâncias diferentes, que por consequência, não são homocedásticos.

A metodologia utilizada na identificação de pontos influentes foi a distância de Cook (Cook, 1977). Ela é representada pela Equação (2.6)

$$D_i = \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right], \tag{2.6}$$

em que e_i é o resíduo da i-ésima observação, p é o número de parâmetros ajustados, MSE é o Erro Quadrático Médio e h_{ii} é o i-ésimo elemento da diagonal da matriz de projeção (conhecida como "matriz chapéu") $\mathbf{H} = \mathbf{X}(\mathbf{X^t X})^{-1}\mathbf{X^t}$. A distância D_i pode ser interpretada como uma medida de mudança quando ajustamos o modelo sem a i-ésima observação, significando que quanto maior for essa distância, maior é a influência dessa observação. Na literatura, sugere-se pontos $D_i > 1$ para serem considerados como pontos

influentes.

Capítulo 3

Aplicação em dados do moinho

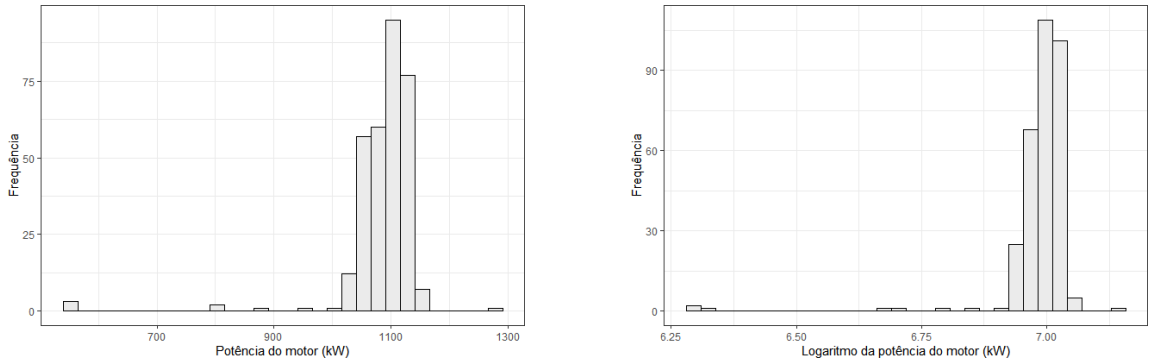
3.1 Banco de dados

Os dados utilizados no trabalho são fornecidos pela empresa de cimento Apodi. O banco de dados para estudos iniciais contém 317 observações de 94 variáveis, coletadas no dia 25 de março de 2019 com intervalos de tempo de aproximadamente 30 segundos utilizando sensores instalados no moinho.

As observações são provenientes do mesmo moinho, portanto elas seriam dependentes, entretanto, neste trabalho, são tratadas como independentes pois especialistas na área acreditam que 30 segundos já é um tempo suficiente para as observações não serem dependentes. Além disso, utilizou-se a produção do mesmo tipo de cimento (CP II-E: Cimento Portland composto com escória granulada de alto forno) para todas as observações.

As variáveis que constam no banco de dados estão associadas à potência do motor do moinho, do ventilador, do elevador do balde, do separador; à temperatura na entrada do moinho, na saída do moinho; ao diferencial de pressão do moinho, à vibração do moinho, à finura do cimento e à altura do leito de material sobre a mesa. Essas variáveis estão divididas em duas categorias, em que "sp" representa aquelas que o operador controla diretamente e "pv" representa as medidas que realmente aconteceram. Algumas das variáveis possuem a unidade de medida em toneladas por hora (ton/h) e percentual (perc) ou metros cúbicos por hora (m3/h) e percentual. A variável de maior interesse no estudo é a potência do motor.

A partir das 94 variáveis, excluímos 28 variáveis que não possuem variabilidade. Então, trabalhamos com as 66 restantes. Afim de comparar a seleção de variáveis e estimação do modelo por diferentes métodos estatísticos, utilizamos a validação cruzada pelo método



(a) potência do motor. (b) Logaritmo da potência do motor.

Figura 3.1: Histograma da variável resposta.

Tabela 3.1: Análise descritiva da variável potência do motor.

Mínimo	1º quartil	Média	Mediana	3º quartil	Máximo
543.8	1066.0	1085.3	1097.9	1116.7	1271.0

Holdout e dividimos aleatoriamente o banco de dados inicial em treino e validação. O banco de dados treino ficou com 222 observações equivalendo a 70% do banco de dados original e o banco de dados validação ficou com 95 observações, equivalendo a 30%. A base de validação foi utilizada apenas para cálculo das métricas na comparação final dos modelos.

3.2 Análise exploratória dos dados

Inicialmente, analisamos o comportamento da variável resposta por meio do histograma, mostrado na Figura 3.1(a). O histograma destaca pontos atípicos inferiores que estão entre 500 e 600 kW e também, pontos atípicos superiores que estão próximos de 1300 kW. Além disso, a Tabela 3.2 traz informações numéricas sobre o comportamento da variável, com o valor mínimo de 543.8, primeiro quartil de 1066.0, média de 1085.3, mediana de 1097.9, terceiro quartil de 1116.7 e valor máximo de 1271.0 kW.

Juntando as informações fornecidas pelo histograma com aquelas fornecidas pela média e mediana, a variável potência do motor parece ser assimétrica, pois a média e a mediana se diferem em 12.6. O coeficiente de assimetria também foi calculado, resultando no valor de -5.4 confirmando a assimetria à esquerda pelo valor negativo e distante de zero. Para mais detalhes sobre assimetrias, veja, por exemplo, Cramér (1999).

A fim de reduzir valores discrepantes no conjunto de dados, utilizamos a transformação

logarítmica na variável resposta. A Figura 3.1(b) mostra a distribuição da variável com a transformação, revelando a diminuição na amplitude da variável, que possui o valor mínimo de 6.299 e valor máximo de 7.148, entretanto a visualização dos pontos atípicos ainda é clara. Como os resultados da seleção de variáveis são muito semelhantes aos sem a transformação, mostramos nesse relatório apenas os resultados obtidos com a variável potência do motor na sua escala original.

3.3 Metodologias comparadas

Para todos os modelos encontrados, foi utilizado o mesmo banco de dados de treino e validação. Com isso, temos que tanto o LASSO como o *stepwise* foram aplicados perante as mesmas condições.

O LASSO realizou o processo de encontrar o parâmetro de regularização com a validação cruzada no método k-fold, utilizando $k = 10$. Para o *stepwise*, o nível de significância de entrada considerado foi de 0.05 e o nível de significância de saída considerado foi de 0.1. Os testes da análise de diagnóstico foram realizados com o nível de significância de 0.05.

3.4 Resultados sem covariáveis defasadas no tempo

Esta seção contém os resultados da seleção de variáveis, incluindo somente as variáveis originais, sem defasagem no tempo. As medidas de adequabilidade do modelo serão apresentadas no final desse capítulo (Seção 3.6) incluindo todos os modelos estimados.

A Tabela 3.2 mostra as variáveis selecionadas pelo LASSO e *stepwise* considerando todas as observações e covariáveis sem defasagem. Para cada uma dessas variáveis, acompanha-se as estimativas dos coeficientes de regressão que podem dar uma ideia da importância de cada variável selecionada para o modelo. Entretanto esses valores não representam exatamente o grau de importância, pois eles são influenciados pela escala da variável e multicolinearidade, ou seja, neste caso não devemos interpretá-los da maneira convencional. Os coeficientes exibidos no método LASSO foram estimados pelo método de mínimos quadrados irrestrito considerando apenas as variáveis importantes segundo o LASSO. O *stepwise* selecionou mais variáveis que o LASSO, e apenas quatro delas (pressão na saída do filtro pv, pressão de entrada no moinho pv, pressão de moagem pv e

Tabela 3.2: Variáveis selecionadas pelo LASSO e *stepwise* e estimativas dos coeficientes. As variáveis em negrito são as selecionadas por ambos os métodos.

LASSO		***Stepwise***	
alimentação de matéria-prima ton/h sp	9.04919	gesso perc pv	-43.3175
pressão na saída do filtro pv	14.08461	gesso ton/h sp	168.7920
pressão de entrada no moinho sp	-21.62958	injeção de água no moinho m3/h pv	168.6210
pressão de entrada no moinho pv	-26.92274	**pressão na saída do filtro pv**	8.2735
pressão de moagem pv	1.69328	diferencial de pressão do filtro pv	-12.0287
altura do rolo 3 em relação à mesa pv	-2.78493	**pressão de entrada no moinho pv**	-36.9800
vibração na caixa de engrenagens 2 pv	21.27657	potência do separador pv	2.7044
pressão na área 7 da mesa pv	0.04719	**pressão de moagem pv**	3.3150
		altura do rolo 3 em relação à mesa pv	-3.1342
		pressão na área 11 da mesa pv	0.5127

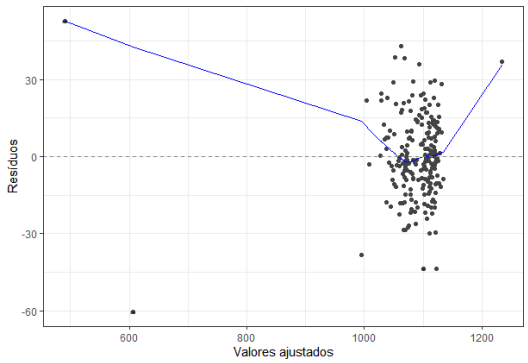
altura do rolo 3 em relação à mesa pv) são selecionadas por ambos. No LASSO, a variável que possui o maior valor da estimativa, em módulo é a pressão de entrada no moinho pv e no *stepwise* é o gesso ton/h sp.

A verificação da suposição da homocedasticidade dos erros do LASSO é mostrada na Figura 3.2(a) usando a base de treino. Se os pontos estiverem aleatoriamente distribuídos no plano em torno do zero significa que os resíduos são homocedásticos. A linha traçada na figura nos ajuda a perceber que há uma tendência nos resíduos, pois ela não está próxima de zero em toda a extensão dos valores ajustados fornecendo uma evidência de que os resíduos não são homocedásticos. O Teste F rejeita a hipótese de homocedasticidade com o p-valor de 0.002211 colaborando com a evidência fornecida pela figura.

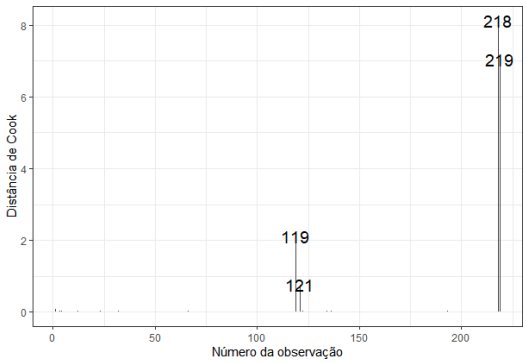
A Figura 3.2(b) mostra a distância de Cook para cada uma das observações, cujas observações 119, 121, 218 e 219 são aquelas que possuem maior destaque e parecem influenciar mais a seleção das variáveis e estimação dos coeficientes de regressão.

A Figura 3.3(a) representa o QQ-plot dos resíduos do modelo selecionado e estimado via *stepwise* em que os pontos estão próximos da reta. Isso mostra que os resíduos parecem ter distribuição normal e o teste de normalidade de Shapiro-Wilk não rejeita a hipótese de normalidade por meio do p-valor de 0.9299. A Figura 3.3(b) mostra uma linha mais próxima de zero e o teste com o p-valor de 0.1848 nos ajuda na conclusão de que não há evidências para dizer que os resíduos não são homocedásticos. Já a Figura 3.3(c) mostra destaque para as observações 119, 218 e 219 como as observações mais influentes.

Feita a análise com todas as observações de treino, decidimos retirar os pontos influen-

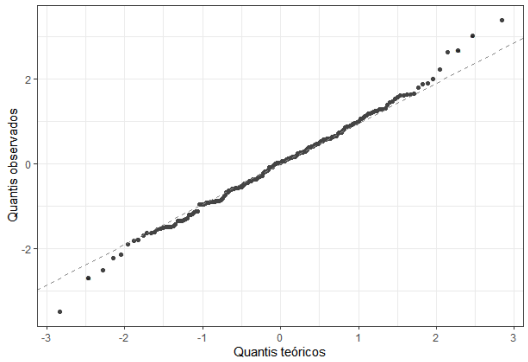


(a) Gráfico de resíduos por valores ajustados.

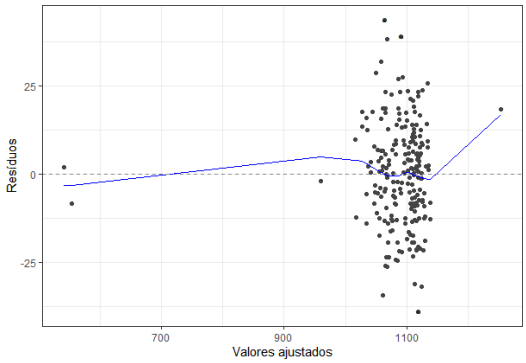


(b) distância de Cook.

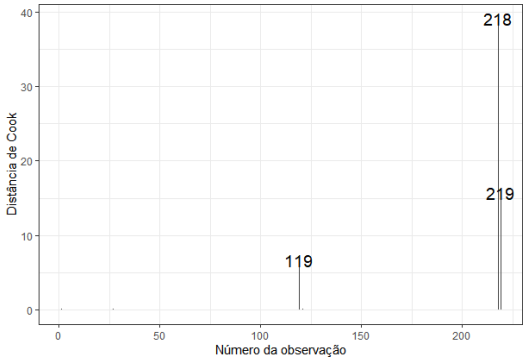
Figura 3.2: Análise de resíduos e pontos influentes da seleção LASSO.



(a) Gráfico QQ-plot.



(b) Gráfico de resíduos por valores ajustados.



(c) distância de Cook

Figura 3.3: Análise de resíduos e pontos influentes da seleção *stepwise*.

Tabela 3.3: Variáveis selecionadas pelo LASSO e *stepwise* sem os pontos influentes e estimativas dos coeficientes. As variáveis em negrito são as selecionadas por ambos os métodos.

LASSO		***Stepwise***	
alimentação de matéria-prima ton/h pv	10.2926	**alimentação de matéria-prima ton/h pv**	12.4789
calcário perc pv	13.0465	**injeção de água no moinho m3/h pv**	216.2448
escória ton/h sp	2.1767	**pressão na saída do filtro pv**	7.5446
gesso perc pv	-18.7227	**pressão de entrada no moinho pv**	-30.2933
injeção de água no moinho m3/h pv	174.0352	velocidade do separador do ponto de ajuste pv	0.1369
pressão na saída do filtro pv	-0.2022	**pressão de moagem pv**	3.2043
pressão de entrada no moinho sp	-12.1406	**so3 perc pv**	-6.5197
pressão de entrada no moinho pv	-22.0812		
pressão na saída do moinho pv	7.0226		
temp do leito do gerador de gás quente pv	0.2889		
temp na saída do gerador de gás quente pv	0.2498		
pressão de moagem pv	2.9883		
pressão na área 2 da mesa pv	0.1919		
pressão na área 8 da mesa pv	-0.2348		
pressão na área 11 da mesa pv	0.4208		
so3 perc pv	-6.7617		

tes indicados pela distância de Cook para verificar se ocorrem muitas mudanças na seleção de variáveis. Pode ocorrer de os sensores do moinho fazerem medições erradas, isto pode ocasionar a existência de pontos atípicos e influentes que não representa o comportamento real das variáveis. A Tabela 3.3 mostra as variáveis selecionadas pelos dois métodos sem os pontos considerados como influentes pela distância de Cook.

Ao contrário do conjunto de variáveis selecionadas com todas as observações de treino, sem os pontos influentes o LASSO selecionou mais variáveis que o *stepwise*. Apesar da diferença no número de variáveis selecionadas, há seis variáveis em comum entre as duas seleções sem pontos influentes, que são alimentação de matéria-prima ton/h pv, injeção de água no moinho m3/h pv, pressão na saída do filtro pv, pressão de entrada no moinho pv, pressão de moagem pv e so3 perc pv em que, praticamente, todas as variáveis selecionadas pelo *stepwise* estão presentes na seleção do LASSO. A maior estimativa de coeficiente está na variável injeção de água no moinho m3/h pv em ambas as seleções.

A Figura 3.4 mostra o gráfico da homocedasticidade dos resíduos para o LASSO, que em comparação ao mesmo gráfico da seleção que contém todas as observações, os pontos aparentam estar mais distribuídos aleatoriamente. O Teste F não rejeita a hipótese de

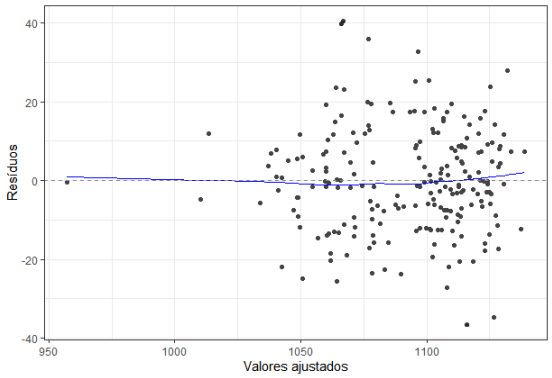
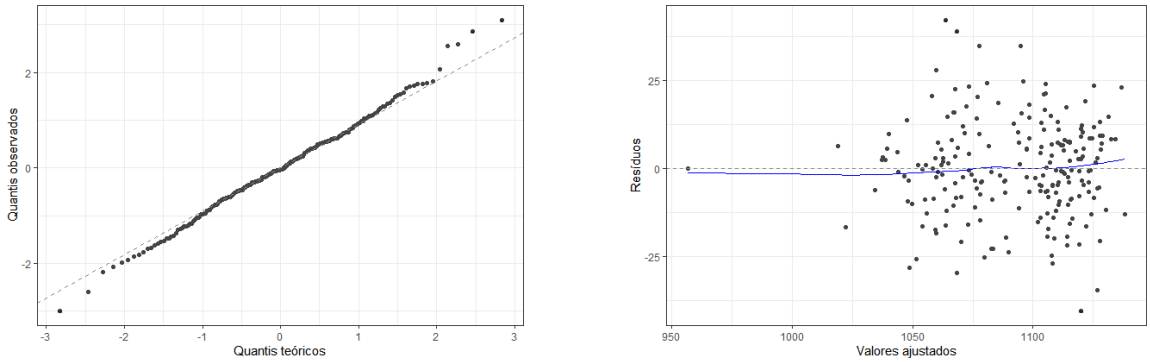


Figura 3.4: Análise de resíduos da seleção LASSO sem pontos influentes.



(a) Gráfico QQ-plot.

(b) Gráfico de resíduos por valores ajustados.

Figura 3.5: Análise de resíduos da seleção *stepwise* sem pontos influentes.

homocedasticidade com o p-valor de 0.5076. A maior distância de Cook registrada foi de 0.14, significando que não há mais pontos influentes.

As suposições de normalidade e homocedasticidade dos erros da seleção *stepwise* sem os pontos influentes aparentam estar atendidas segundo as Figuras 3.5(a) e 3.5(b), respectivamente. O resultado do Teste de Shapiro-Wilk foi de p-valor = 0.7715 e Teste F com p-valor = 0.1216, portanto as duas suposições não foram violadas neste caso. Além disso, o valor máximo da distância de Cook foi de 0.04, indicando que não há mais pontos influentes. Em relação às variáveis selecionadas observamos alteração quando comparamos o mesmo método com e sem pontos influentes. Ou seja, pontos influentes devem ser analisados pois impactam na escolha das variáveis.

3.5 Resultados com covariáveis defasadas no tempo

As variáveis defasadas foram construídas levando em consideração quatro tempos de atraso: 30, 60, 90 e 120 segundos, em relação à potência do motor observada em cada

Tabela 3.4: Variáveis selecionadas pelo LASSO com covariáveis defasadas no tempo e estimativas dos coeficientes. As variáveis em negrito são as selecionadas por ambos os métodos.

LASSO			
0 segundos		**30 segundos**	
pressão na saída do filtro pv	18.5451	gesso ton/h pv	92.5852
pressão de entrada no moinho sp	-13.8599		
pressão de entrada no moinho pv	-21.9002		
pressão de moagem pv	1.3301		
altura do rolo 3 em relação à mesa pv	-2.7765		
vibração na caixa de engrenagens 2 pv	18.3270		
pressão na área 7 da mesa pv	0.1173		

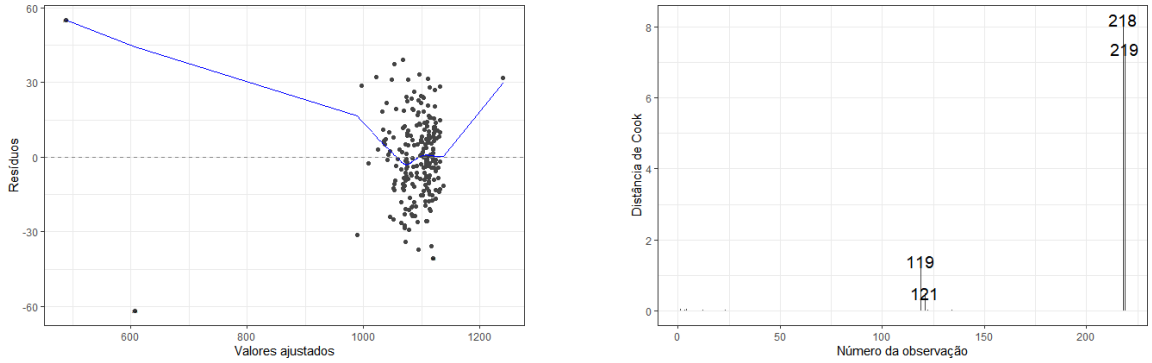
momento. Cada covariável foi deslocada em uma, duas, três e quatro observações, ou seja, considerando i o número da observação, temos que para o atraso de 30 segundos, Y_i tem como correspondente X_{i-1}, para o atraso de 60 segundos, temos como correspondente X_{i-2} e assim por diante. A seleção de variáveis foi realizada juntando todas essas variáveis mais o conjunto de variáveis original. A seleção de variáveis foi realizada da mesma forma que a seleção sem covariáveis defasadas. Vale notar que agora existe uma associação muito grande entre as variáveis, pois é a mesma variável que está sendo utilizada para gerar as defasagens.

As Tabelas 3.4 e 3.5 mostram as variáveis selecionadas pelo LASSO e *stepwise*. O LASSO selecionou muito menos variáveis que o *stepwise* e a maior parte das variáveis selecionadas pelos dois métodos não possui defasagem ou possui defasagem de 30 segundos. As quatro variáveis selecionadas pelos dois métodos são pressão na saída do filtro pv, pressão de entrada no moinho pv, pressão de moagem pv e altura do rolo 3 em relação à mesa pv sem defasagem, que são as mesmas 4 covariáveis selecionadas por ambos métodos na ausência de covariáveis defasadas. A variável com a maior estimativa é gesso ton/h pv com defasgem de 30 segundos no LASSO e injeção de água no moinho m3/h pv sem defasagem no *stepwise*.

Para a seleção LASSO, a Figura 3.6(a) mostra o gráfico de resíduos por valores ajustados para termos uma ideia de como os resíduos estão distribuídos. Junto com o Teste F que nos deu o resultado de p-valor = 0.0002517, temos evidências de que os resíduos

Tabela 3.5: Variáveis selecionadas pelo *stepwise* com covariáveis defasadas no tempo e estimativas dos coeficientes. As variáveis em negrito são as selecionadas por ambos os métodos.

Stepwise			
0 segundos		**30 segundos**	
clínquer perc pv	-7.6848	clínquer ton/h pv	17.8185
injeção de água no moinho perc sp	-19.6755	pressão na saída do filtro pv	-14.8706
injeção de água no moinho m3/h pv	748.1530	pressão na área 2 da mesa pv	1.0967
pressão na saída do filtro pv	27.3353	pressão na área 3 da mesa pv	-0.6195
pressão de entrada no moinho pv	-38.1008	pressão na área 8 da mesa pv	1.3196
pressão de moagem pv	3.0897	pressão na área 11 da mesa pv	1.0275
altura do rolo 3 em relação à mesa pv	-2.1992		
pressão na área 11 da mesa pv	0.4928		
60 segundos		**90 segundos**	
pressão na área 5 da mesa pv	-0.4074	cinzas volantes 2 pv	-154.2352
pressão na área 7 da mesa pv	0.7492	altura do rolo 1 em relação à mesa pv	3.2297
		altura do rolo 3 em relação à mesa pv	-2.1341
120 segundos			
vibração na caixa de engrenagens 1 pv	-4.8596		
vibração na caixa de engrenagens 2 pv	-24.8952		



(a) Gráfico de resíduos por valores ajustados. (b) distância de Cook.

Figura 3.6: Análise de resíduos e pontos influentes da seleção LASSO com defasagem.

não são homocedásticos. A Figura 3.6(b) nos mostra novamente os pontos 119, 121, 218 e 219 como pontos influentes.

Para a seleção *stepwise*, com a análise da Figura 3.7(a) junto com o teste de Shapiro-Wilk de p-valor = 0.9439, não temos evidências para dizer que os resíduos não apresentam distribuição normal. A Figura 3.7(b) junto com o Teste F de p-valor = 0.3499, não temos evidências para dizer que os resíduos não são homocedásticos. A Figura 3.7(c) indica os pontos 1, 119, 218 e 219 como pontos influentes.

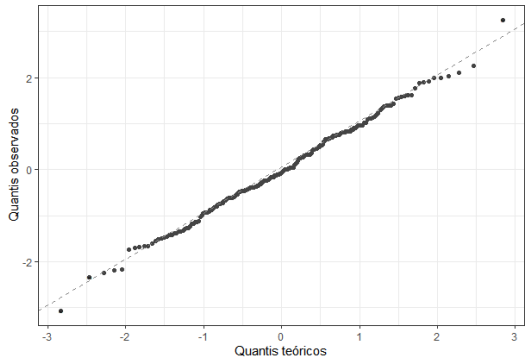
Com a informação de quais pontos são influentes, retiramos-os da análise e aplicamos os métodos de seleção novamente para verificar suas influências na seleção de variáveis.

As Tabelas 3.6 e 3.7 mostram quais variáveis foram selecionadas. Ao contrário da situação anterior, o *stepwise* selecionou menos variáveis que o LASSO, e as variáveis em comum são alimentação de matéria-prima ton/h pv, pressão de entrada no moinho pv e pressão de moagem pv para variáveis sem defasagem, cinzas volantes 2 pv, pressão na saída do moinho pv, pressão na área 1 da mesa pv, pressão na área 8 da mesa pv e pressão na área 11 da mesa pv para variáveis com defasagem de 30 segundos, cinzas volantes 2 pv, altura do rolo 1 em relação à mesa pv e altura do rolo 3 em relação à mesa pv para variáveis de 90 segundos e calcário perc pv, pressão de moagem pv e vibração na caixa de engrenagens 1 pv para variáveis de 120 segundos, que totalizam 14 variáveis. Para ambas as seleções, a variável que possui a maior estimativa do coeficiente é cinzas volantes 2 pv com defasagem de 90 segundos.

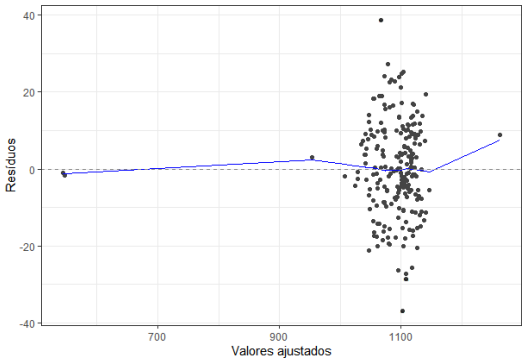
A análise de resíduos para o LASSO é mostrada na Figura 3.8 indicando uma dispersão aleatória dos pontos, e junto com o Teste F de p-valor = 0.6262 podemos concluir que não há evidências para rejeitar a suposição de homocedasticidade dos erros. A maior distância de Cook registrada é de 1.11.

Tabela 3.6: Variáveis selecionadas pelo LASSO com covariáveis defasadas no tempo sem pontos influentes e estimativas dos coeficientes. As variáveis em negrito são as selecionadas por ambos os métodos.

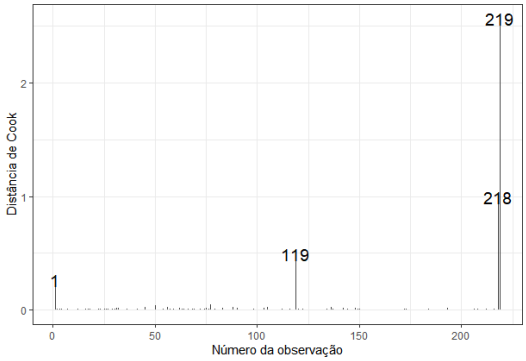
LASSO			
0 segundos		**30 segundos**	
alimentação de matéria-prima ton/h pv	5.45950	clínquer ton/h pv	5.52579
calcário ton/h pv	17.49990	gesso ton/h pv	6.42306
pressão de entrada no moinho sp	-1.75191	**cinzas volantes 2 pv**	108.01256
pressão de entrada no moinho pv	-24.45722	pressão de entrada no moinho sp	-13.76175
pressão na saída do moinho pv	-1.36600	**pressão na saída do moinho pv**	7.77682
pressão de moagem pv	2.99977	altura do rolo 3 em relação à mesa pv	-0.58915
pressão na área 2 da mesa pv	0.30016	auxiliares de moagem pv	10.85032
pressão na área 11 da mesa pv	0.47385	**pressão na área 1 da mesa pv**	0.63129
so3 perc pv	-0.89786	**pressão na área 8 da mesa pv**	0.73597
		pressão na área 11 da mesa pv	0.81690
60 segundos		**90 segundos**	
escória perc pv	10.07942	alimentação de matéria-prima perc pv	7.42490
pressão na área 7 da mesa pv	0.37606	escória ton/h pv	-8.66627
pressão na área 11 da mesa pv	0.32510	**cinzas volantes 2 pv**	-157.39324
pressão na área 12 da mesa pv	0.18480	pressão de moagem pv	0.40864
		altura do rolo 1 em relação à mesa pv	2.17754
		altura do rolo 3 em relação à mesa pv	-1.99286
		auxiliares de moagem pv	-1.53607
		so3 perc pv	-0.19125
120 segundos			
calcário perc pv	-20.60810		
escória ton/h pv	-0.73772		
gesso ton/h pv	16.18444		
pressão na saída do filtro pv	1.46971		
pressão de entrada no moinho sp	-13.28450		
temp de entrada no moinho pv	-3.40607		
pressão de moagem pv	0.70818		
vibração na caixa de engrenagens 1 pv	-5.47658		
partículas expelidas pv	0.02891		



(a) Gráfico QQ-plot.



(b) Gráfico de resíduos por valores ajustados.



(c) distância de Cook

Figura 3.7: Análise de resíduos e pontos influentes da seleção *stepwise* com defasagem.

Tabela 3.7: Variáveis selecionadas pelo *stepwise* com covariáveis defasadas no tempo sem pontos influentes e estimativas dos coeficientes. As variáveis em negrito são as selecionadas por ambos os métodos.

Stepwise			
0 segundos		**30 segundos**	
alimentação de matéria-prima ton/h pv	10.9942	**cinzas volantes 2 pv**	97.7618
pressão de entrada no moinho pv	-32.6586	**pressão na saída do moinho pv**	11.3191
pressão de moagem pv	3.4193	**pressão na área 1 da mesa pv**	0.7195
		pressão na área 8 da mesa pv	0.8815
		pressão na área 11 da mesa pv	0.7513
90 segundos		**120 segundos**	
cinzas volantes 2 pv	-147.3213	**calcário perc pv**	-22.3488
altura do rolo 1 em relação à mesa pv	2.7423	**pressão de moagem pv**	1.3918
altura do rolo 3 em relação à mesa pv	-1.9968	**vibração na caixa de engrenagens 1 pv**	-6.0674

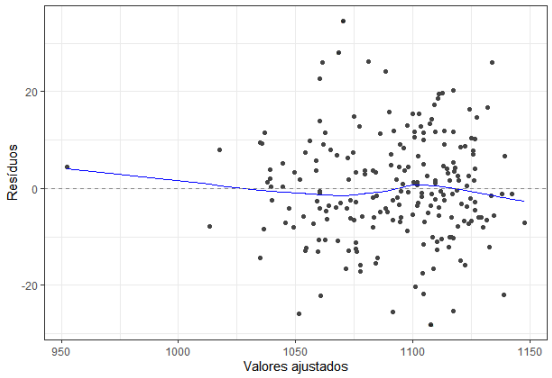
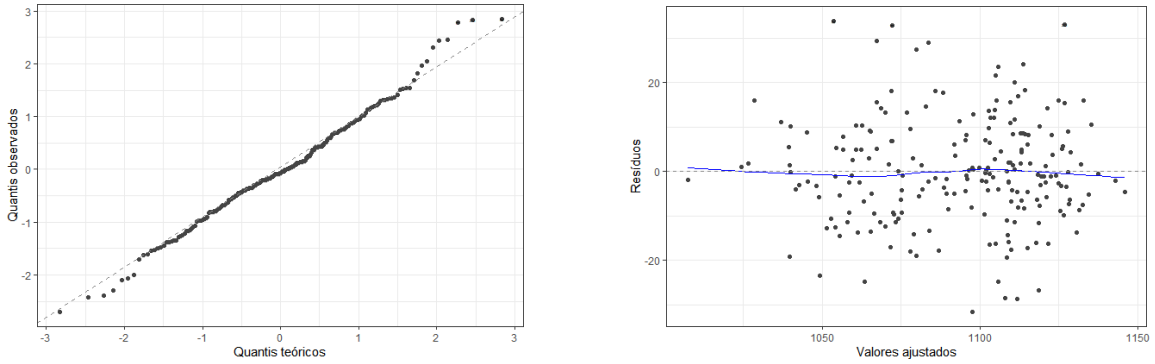


Figura 3.8: Análise de resíduos da seleção LASSO com defasagem e sem pontos influentes.



(a) Gráfico QQ-plot.

(b) Gráfico de resíduos por valores ajustados.

Figura 3.9: Análise de resíduos da seleção *stepwise* com defasagem e sem pontos influentes.

Para o *stepwise*, com a Figura 3.9(a) que mostra os pontos próximos à linha, junto com o teste de Shapiro-Wilk de p-valor = 0.1988, podemos concluir que não há evidências para rejeitar a suposição de normalidade dos erros. A Figura 3.9(b) mostra os pontos espalhados aleatoriamente pelo plano evidenciando a independência e homocedasticidade entre os resíduos, e juntamente com o Teste F de p-valor = 0.1195, não há evidências de que a suposição está violada. A maior distância de Cook é de 0.06, significando que não há mais pontos influentes.

3.6 Comparação entre os resultados

A Tabela 3.8 mostra os valores da função de risco, AIC e BIC para cada um dos modelos encontrados, calculados na base de validação. As colunas especificam o modelo que estamos referenciando, no qual a primeira divisória é se o modelo possui defasagens ou não nas covariáveis, a segunda divisória é se o modelo possui todas as observações ou não, por último, especifica o método de seleção utilizado. Em relação as linhas, a primeira

mostra os valores da função de risco, a segunda os valores de AIC e terceira o BIC.

Tabela 3.8: Valores da função de risco, AIC e BIC para cada um dos modelos encontrados.

	Sem defasagem				Com defasagem			
	Todas as observações		Sem pontos influentes		Todas as observações		Sem pontos influentes	
	LASSO	*Stepwise*	LASSO	*Stepwise*	LASSO	*Stepwise*	LASSO	*Stepwise*
R(g)	262.6173	237.1658	221.6901	210.5324	301.4253	484.1292	249.2098	270.9317
AIC	792.4060	782.8238	776.4808	771.6266	805.3615	849.9015	787.4802	795.3359
BIC	794.9493	785.3671	779.0241	774.1699	807.9048	852.4448	790.0235	797.8792

Note que os modelos com covariáveis defasadas sempre se mostraram menos adequados que o seu correspondente sem covariáveis defasadas. Isso pode indicar a inexistência de defasagens no tempo, mas pode haver uma inadequação na abordagem realizada ao juntar todas as covariáveis defasadas em um único conjunto.

Para os três critérios de escolha, o *stepwise* sem pontos influentes e sem defasagem obteve os menores valores, significando que os três critérios indicam para um mesmo modelo escolhendo-o como aquele que se ajusta melhor aos dados e que possui melhor poder preditivo. Além disso, as suposições de homocedasticidade e normalidade dos erros para este modelo foram atendidas.

Apesar do modelo mais adequado ter sido observado na metodologia *stepwise*, o LASSO (sem defasagem e sem pontos influentes) teve desempenho muito parecido em relação aos valores dos critérios e se mostrou mais estável entre as quatro situações testadas.

Portanto, para o banco de dados considerado nessa situação, indicamos as variáveis selecionadas pelos *stepwise* sem pontos influentes e sem defasagem como sendo as variáveis que realmente impactam na potência do motor e sugeríamos as variávies adicionais selecionadas pelo LASSO sem defasagem e pontos influentes para estudos mais refinados de influência.

Capítulo 4

Conclusão e estudos futuros

Neste trabalho, fizemos uma comparação entre duas metodologias de seleção de variáveis, o LASSO e o *stepwise*, através das quais abordamos quatro situações diferentes para cada tipo de seleção: variáveis com ou sem defasagens no tempo e considerando todas as observações ou descartando os pontos influentes. Aplicamos as metodologias em um banco de dados de moinho de cimento fornecida pela empresa Cimento Apodi, e concluímos que a melhor seleção de variáveis para o caso considerado foi do *stepwise* sem pontos influentes e sem defasagem, decidida através de três critérios (Função de risco, AIC e BIC) que se concordaram.

A presença de pontos atípicos e influentes alteram a escolha das covariáveis mais importantes e deve ser analisada e considerada. A suposição de independência entre as observações não necessariamente está atendida, porque os dados se tratam de medidas no mesmo moinho ao longo do tempo. Aqui consideramos que essa suposição é atendida, mas métodos que consideram observações correlacionadas podem ser estudados no futuro. Além disso, as covariáveis defasadas no tempo não aumentaram o poder preditivo do modelo, tornando o modelo estimado mais complexo e menos adequado aos dados. Um motivo para esse fato é que as covariáveis são muito correlacionadas. Além disso, as estimativas dos coeficientes obtidas a partir das variáveis selecionadas não devem ser interpretadas da maneira usual devido à multicolinearidade, podendo levar a conclusões incorretas. Como sugestão, pode-se lidar com esta multicolinearidade e refinar as metodologias para apresentarem um melhor desempenho no caso considerado, selecionar melhor as variáveis e estimar seus coeficientes para tentar uma interpretação.

Das 66 variáveis iniciais, conseguimos escolher 7 que mais impactam a potência do motor. Esperamos que a empresa consiga aproveitar os nossos resultados aplicando-os no

algoritmo de otimização, atingindo o objetivo de aumentar a produção de cimento.

Apêndice A

Códigos utilizados no trabalho

A.1 Análise descritiva

```
dados <- read.csv('sem-variab.csv')

library(ggplot2) #ggplot
library(e1071) # skewness

# histograma da potência do motor
ggplot(dados, aes(mill_motor_pwr_kw_pv)) + geom_histogram(color="black",
fill="gray92") + labs(x="Potência do motor (kW)", y = "Frequência") +
theme_bw()

# descritiva da potência do motor
summary(dados$mill_motor_pwr_kw_pv)

# medida de assimetria da potência do motor
skewness(dados$mill_motor_pwr_kw_pv,type = 1)

# histograma do log da potência do motor
ggplot(dados, aes(log(mill_motor_pwr_kw_pv))) + geom_histogram(color="black",
fill="gray92") + labs(x="Logaritmo da potência do motor (kW)",
y = "Frequência") +  theme_bw()
```

```
# descritiva do log da potência do motor
summary(log(dados$mill_motor_pwr_kw_pv))
```

A.2 Seleção LASSO e *stepwise* sem defasagem

Código SAS

```
* com pontos influentes;
proc glmselect data=work.import;
   model millmotorpwrkwpv=totalfeedtonhsp--millinjectionwaterm3hpv
   mainfanpwrkwpv--finenesspercpv /
   selection=stepwise(select = SL SLE=0.05 SLS=0.1);
run;

* sem pontos influentes;
proc glmselect data=work.import1;
   model millmotorpwrkwpv=totalfeedtonhsp--millinjectionwaterm3hpv
   mainfanpwrkwpv--finenesspercpv /
   selection=stepwise(select = SL SLE=0.05 SLS=0.1);
run;
```

Código R

```
library(glmnet)
library(corrplot)
library(ggfortify)
library(gnm)

treino <- read.csv('treino.csv')
validacao <- read.csv('validação.csv')

################
# Selecao: LASSO
################
```

```
sem_motor <- subset(treino, select = -c(mill_motor_pwr_kw_pv))

set.seed(12345)
aj_lasso1 <- cv.glmnet(as.matrix(sem_motor), treino$mill_motor_pwr_kw_pv,
alpha = 1)
lambda1 <- aj_lasso1$lambda.min
coef(aj_lasso1, s = lambda1)

aj1 <- lm(mill_motor_pwr_kw_pv ~ total_feed_ton.h_sp +
main_bf_out_press_mbar_pv + mill_in_pres_mbar_sp + mill_in_pres_mbar_pv +
grinding_pressure_bar_pv + roller_3_bed_depth_mm_pv +
gearbox_2_vibration_mm.s_pv + thrust_pad_7_press_bar_pv, data = treino)

validacao1 <- subset(validacao, select = -c(mill_motor_pwr_kw_pv))
pred1 <- predict(aj1, newdata = validacao1)
risco1 <- mean( (validacao$mill_motor_pwr_kw_pv-pred1)^2 )

aj_1p <- gnm (mill_motor_pwr_kw_pv ~ total_feed_ton.h_sp +
main_bf_out_press_mbar_pv + mill_in_pres_mbar_sp + mill_in_pres_mbar_pv +
grinding_pressure_bar_pv + roller_3_bed_depth_mm_pv +
gearbox_2_vibration_mm.s_pv + thrust_pad_7_press_bar_pv,
data = validacao, constrain = "*", constrainTo = coef(aj1) )

AIC(aj_1p)
BIC(aj_1p)

####################
# Selecao: Stepwise
####################

aj3 <- lm(mill_motor_pwr_kw_pv ~ gypsum_perc_pv + gypsum_ton.h_sp +
mill_injection_water_m3.h_pv + main_bf_out_press_mbar_pv +
main_bf_dp_mbar_pv + mill_in_pres_mbar_pv + separator_pwr_kw_pv +
```

```
grinding_pressure_bar_pv + roller_3_bed_depth_mm_pv +
thrust_pad_11_press_bar_pv, data = treino)

validacao1 <- subset(validacao, select = -c(mill_motor_pwr_kw_pv))
pred3 <- predict(aj3, newdata = validacao1)
risco3 <- mean((validacao$mill_motor_pwr_kw_pv - pred3)^2)

aj_3p <- gnm (mill_motor_pwr_kw_pv ~ gypsum_perc_pv + gypsum_ton.h_sp +
mill_injection_water_m3.h_pv + main_bf_out_press_mbar_pv +
main_bf_dp_mbar_pv + mill_in_pres_mbar_pv +
separator_pwr_kw_pv + grinding_pressure_bar_pv +
roller_3_bed_depth_mm_pv + thrust_pad_11_press_bar_pv,
data = validacao, constrain = "*", constrainTo = coef (aj3))

AIC(aj_3p)
BIC(aj_3p)

############################
# Analise de Residuos: Lasso
############################

# Erros possuem variancia constante? Nao
autoplot(aj1, which = 1, ncol = 1, label.size = 1) + theme_bw() +
ggtitle(NULL) + xlab('Valores ajustados') + ylab('Resíduos')

var.test(residuals(aj1)[treino$mill_motor_pwr_kw_pv >
median(treino$mill_motor_pwr_kw_pv)],
residuals(aj1)[treino$mill_motor_pwr_kw_pv <
median(treino$mill_motor_pwr_kw_pv)])

# Erros tem distribuicao normal? Sim
autoplot(aj1, which = 2, ncol = 1, label.size = 1)
```

```
shapiro.test(residuals(aj1))

# Pontos influentes (Distancia de Cook)? Pontos 119, 121, 218, 219
autoplot(aj1, which = 4, ncol = 1, label.size = 5, label.n = 4) + theme_bw() +
ggtitle(NULL) + xlab('Número da observação') + ylab('Distância de Cook')

################################
# Analise de Residuos: Stepwise
################################

# Erros possuem variancia constante? Sim
autoplot(aj3, which = 1, ncol = 1, label.size = 1) + theme_bw() +
ggtitle(NULL) + xlab('Valores ajustados') + ylab('Resíduos')

var.test(residuals(aj3)[treino$mill_motor_pwr_kw_pv >
median(treino$mill_motor_pwr_kw_pv)],
residuals(aj3)[treino$mill_motor_pwr_kw_pv <
median(treino$mill_motor_pwr_kw_pv)])

# Erros tem distribuicao normal? Sim
autoplot(aj3, which = 2, ncol = 1, label.size = 1) + theme_bw() +
ggtitle(NULL) + xlab('Quantis teóricos') + ylab('Quantis observados')

shapiro.test(residuals(aj3))

# Pontos influentes (Distancia de Cook)? Pontos 119, 218, 219
autoplot(aj3, which = 4, ncol = 1, label.size = 5, label.n = 3) + theme_bw() +
ggtitle(NULL) + xlab('Número da observação') + ylab('Distância de Cook')


#############################
# Selecao sem outliers: LASSO
#############################
```

```
treino_lasso_or <- treino[-c(119, 121, 218, 219),]

sem_motor1 <- subset(treino_lasso_or, select = -c(mill_motor_pwr_kw_pv))

set.seed(12345)
aj_lasso11 <- cv.glmnet(as.matrix(sem_motor1),
treino_lasso_or$mill_motor_pwr_kw_pv, alpha = 1)
lambda11 <- aj_lasso11$lambda.min
coef(aj_lasso11, s = lambda11)

aj11 <- lm(mill_motor_pwr_kw_pv ~ total_feed_ton.h_pv + limestone_perc_pv +
slag_ton.h_sp + gypsum_perc_pv + mill_injection_water_m3.h_pv +
main_bf_out_press_mbar_pv + mill_in_pres_mbar_sp + mill_in_pres_mbar_pv +
mill_out_pres_mbar_pv + icon_bed_temp_c_pv + icon_exit_temp_c_pv +
grinding_pressure_bar_pv + thrust_pad_2_press_bar_pv +
thrust_pad_8_press_bar_pv + thrust_pad_11_press_bar_pv + so3_perc_pv,
data = treino_lasso_or)


validacao1 <- subset(validacao, select = -c(mill_motor_pwr_kw_pv))
pred11 <- predict(aj11, newdata = validacao1)
risco11 <- mean((validacao$mill_motor_pwr_kw_pv-pred11)^2)

aj_11p <- gnm (mill_motor_pwr_kw_pv ~ total_feed_ton.h_pv + limestone_perc_pv +
slag_ton.h_sp + gypsum_perc_pv + mill_injection_water_m3.h_pv +
main_bf_out_press_mbar_pv + mill_in_pres_mbar_sp + mill_in_pres_mbar_pv +
mill_out_pres_mbar_pv + icon_bed_temp_c_pv + icon_exit_temp_c_pv +
grinding_pressure_bar_pv + thrust_pad_2_press_bar_pv +
thrust_pad_8_press_bar_pv + thrust_pad_11_press_bar_pv + so3_perc_pv,
data = validacao, constrain = "*", constrainTo = coef (aj11))

AIC(aj_11p)
```

```
BIC(aj_11p)

################################
# Selecao sem outliers: Stepwise
################################

treino_step_or <- treino[-c(119, 218, 219),]

aj3a <- lm(mill_motor_pwr_kw_pv ~ total_feed_ton.h_pv +
mill_injection_water_m3.h_pv + main_bf_out_press_mbar_pv +
mill_in_pres_mbar_pv + separator_speed_rpm_pv + grinding_pressure_bar_pv +
so3_perc_pv, data = treino_step_or)
round(aj3a$coefficients,4)

pred3a <- predict(aj3a, newdata = validacao1)
risco3a <- mean((validacao$mill_motor_pwr_kw_pv-pred3a)^2)

aj_3ap <- gnm (mill_motor_pwr_kw_pv ~ total_feed_ton.h_pv +
mill_injection_water_m3.h_pv + main_bf_out_press_mbar_pv +
mill_in_pres_mbar_pv + separator_speed_rpm_pv + grinding_pressure_bar_pv +
so3_perc_pv, data = validacao, constrain = "*", constrainTo = coef (aj3a))

AIC(aj_3ap)
BIC(aj_3ap)

########################################
# Analise de residuos sem outliers: Lasso
########################################

# Erros possuem varincia constante? Sim
autoplot(aj11, which = 1, ncol = 1, label.size = 1) + theme_bw() +
ggtitle(NULL) + xlab('Valores ajustados') + ylab('Resíduos')
```

```
var.test(residuals(aj11)[treino_lasso_or$mill_motor_pwr_kw_pv >
median(treino_lasso_or$mill_motor_pwr_kw_pv)],
residuals(aj11)[treino_lasso_or$mill_motor_pwr_kw_pv <
median(treino_lasso_or$mill_motor_pwr_kw_pv)])

# Erros tem distribuicao normal? Sim
autoplot(aj11, which = 2, ncol = 1, label.size = 1)

shapiro.test(residuals(aj11))

# Pontos influentes (Distancia de Cook)? Pontos nenhum
autoplot(aj11, which = 4, ncol = 1, label.size = 5, label.n = 1) +
theme_bw() + ggtitle(NULL) + xlab('Número da observação') +
ylab('Distância de Cook')
max(cooks.distance(aj11))

##############################################
# Analise de residuos sem outliers: Stepwise
##############################################

# Erros possuem variancia constante? Sim
autoplot(aj3a, which = 1, ncol = 1, label.size = 1) + theme_bw() +
ggtitle(NULL) + xlab('Valores ajustados') + ylab('Resíduos')

var.test(residuals(aj3a)[treino$mill_motor_pwr_kw_pv >
median(treino$mill_motor_pwr_kw_pv)],
residuals(aj3a)[treino$mill_motor_pwr_kw_pv <
median(treino$mill_motor_pwr_kw_pv)])

# Erros tem distribuicao normal? Sim
autoplot(aj3a, which = 2, ncol = 1, label.size = 1) + theme_bw() +
ggtitle(NULL) + xlab('Quantis teóricos') + ylab('Quantis observados')
```

```
shapiro.test(residuals(aj3a))

# Pontos influentes (Distancia de Cook)? Pontos nenhum
autoplot(aj3a, which = 4, ncol = 1, label.size = 5, label.n = 3) + theme_bw() +
ggtitle(NULL) + xlab('Número da observação') + ylab('Distância de Cook')
max(cooks.distance(aj3a))
```

A.3 Seleção LASSO e *stepwise* com defasagem

Código SAS

```
* Com pontos influentes;
proc glmselect data=WORK.IMPORT1;
   model millmotorpwrkwpv=totalfeedtonhsp--finenesspercpv4 /
   selection=stepwise(select = SL SLE=0.05 SLS=0.1);
run;

* Sem pontos influentes;
proc glmselect data=WORK.IMPORT2;
   model millmotorpwrkwpv=totalfeedtonhsp--finenesspercpv4 /
   selection=stepwise(select = SL SLE=0.05 SLS=0.1);
run;
```

Código R

```
library(glmnet)
library(corrplot)
library(ggfortify)
library(gnm)

treino_1 <- read.csv('defasagem-total-treino.csv')
validacao <- read.csv('defasagem-total-validação.csv')

################
# Selecao: LASSO
```

```
################

treino_1m <- subset(treino_1, select = -c(mill_motor_pwr_kw_pv))

set.seed(12345)
lasso_1 <- cv.glmnet(as.matrix(treino_1m), treino_1$mill_motor_pwr_kw_pv,
alpha = 1)
lambda_1 <- lasso_1$lambda.min
coef(lasso_1, s = lambda_1)

aj_1l <- lm(mill_motor_pwr_kw_pv ~ main_bf_out_press_mbar_pv +
mill_in_pres_mbar_sp + mill_in_pres_mbar_pv + grinding_pressure_bar_pv +
roller_3_bed_depth_mm_pv + gearbox_2_vibration_mm.s_pv +
thrust_pad_7_press_bar_pv + gypsum_ton.h_pv.1, data = treino_1)

aj_1lp <- gnm (mill_motor_pwr_kw_pv ~ main_bf_out_press_mbar_pv +
mill_in_pres_mbar_sp + mill_in_pres_mbar_pv + grinding_pressure_bar_pv +
roller_3_bed_depth_mm_pv + gearbox_2_vibration_mm.s_pv +
thrust_pad_7_press_bar_pv + gypsum_ton.h_pv.1, data = validacao,
constrain = "*", constrainTo = coef(aj_1l))

AIC(aj_1lp)
BIC(aj_1lp)

validacao_m <- subset(validacao, select = -c(mill_motor_pwr_kw_pv))
pred_1l <- predict(aj_1l, newdata = validacao_m)
risco_1l <- mean((validacao$mill_motor_pwr_kw_pv-pred_1l)^2)

###################
# Selecao: Stepwise
###################

aj_1s <- lm(mill_motor_pwr_kw_pv ~ clinker_perc_pv +
```

```
mill_injection_water_perc_sp + mill_injection_water_m3.h_pv +
main_bf_out_press_mbar_pv + mill_in_pres_mbar_pv + grinding_pressure_bar_pv +
roller_3_bed_depth_mm_pv +  thrust_pad_11_press_bar_pv + clinker_ton.h_pv.1 +
main_bf_out_press_mbar_pv.1 + thrust_pad_2_press_bar_pv.1 +
thrust_pad_3_press_bar_pv.1 + thrust_pad_8_press_bar_pv.1 +
thrust_pad_11_press_bar_pv.1 + thrust_pad_5_press_bar_pv.2 +
thrust_pad_7_press_bar_pv.2 + fly_ash_2_ton.h_pv.3 +
roller_1_bed_depth_mm_pv.3 + roller_3_bed_depth_mm_pv.3 +
gearbox_1_vibration_mm.s_pv.4 + gearbox_2_vibration_mm.s_pv.4,
data = treino_1)

pred_1s <- predict(aj_1s, newdata = validacao_m)
risco_1s <- mean((validacao$mill_motor_pwr_kw_pv-pred_1s)^2)

aj_1sp <- gnm (mill_motor_pwr_kw_pv ~ clinker_perc_pv +
mill_injection_water_perc_sp + mill_injection_water_m3.h_pv +
main_bf_out_press_mbar_pv + mill_in_pres_mbar_pv + grinding_pressure_bar_pv +
roller_3_bed_depth_mm_pv + thrust_pad_11_press_bar_pv + clinker_ton.h_pv.1 +
main_bf_out_press_mbar_pv.1 + thrust_pad_2_press_bar_pv.1 +
thrust_pad_3_press_bar_pv.1 + thrust_pad_8_press_bar_pv.1 +
thrust_pad_11_press_bar_pv.1 + thrust_pad_5_press_bar_pv.2 +
thrust_pad_7_press_bar_pv.2 + fly_ash_2_ton.h_pv.3 +
roller_1_bed_depth_mm_pv.3 + roller_3_bed_depth_mm_pv.3 +
gearbox_1_vibration_mm.s_pv.4 + gearbox_2_vibration_mm.s_pv.4,
data = validacao, constrain = "*", constrainTo = coef (aj_1s))

AIC(aj_1sp)
BIC(aj_1sp)

##############################
# Analise de Residuos: Lasso
##############################
```

```
# Erros possuem variancia constante? Nao
autoplot(aj_1l, which = 1, ncol = 1, label.size = 1) + theme_bw() +
ggtitle(NULL) + xlab('Valores ajustados') + ylab('Resíduos')

var.test(residuals(aj_1l)[treino_1$mill_motor_pwr_kw_pv >
median(treino_1$mill_motor_pwr_kw_pv)],
residuals(aj_1l)[treino_1$mill_motor_pwr_kw_pv <
median(treino_1$mill_motor_pwr_kw_pv)])

# Erros tem distribuicao normal? Sim
autoplot(aj_1l, which = 2, ncol = 1, label.size = 1)

shapiro.test(residuals(aj_1l))

# Pontos influentes (Distancia de Cook)? Pontos 119, 121, 218 e 219
autoplot(aj_1l, which = 4, ncol = 1, label.size = 5, label.n = 4) +
theme_bw() + ggtitle(NULL) + xlab('Número da observação') +
ylab('Distância de Cook')



################################
# Analise de Residuos: Stepwise
################################

# Erros possuem variancia constante? Sim
autoplot(aj_1s, which = 1, ncol = 1, label.size = 1) + theme_bw() +
ggtitle(NULL) + xlab('Valores ajustados') + ylab('Resíduos')

var.test(residuals(aj_1s)[treino_1$mill_motor_pwr_kw_pv >
median(treino_1$mill_motor_pwr_kw_pv)],
residuals(aj_1s)[treino_1$mill_motor_pwr_kw_pv <
median(treino_1$mill_motor_pwr_kw_pv)])
```

```
# Erros tem distribuicao normal? Sim
autoplot(aj_1s, which = 2, ncol = 1, label.size = 1) + theme_bw() +
ggtitle(NULL) + xlab('Quantis teóricos') + ylab('Quantis observados')

shapiro.test(residuals(aj_1s))

# Pontos influentes (Distancia de Cook)? Pontos 1, 119, 218 e 219
autoplot(aj_1s, which = 4, ncol = 1, label.size = 5, label.n = 4) +
theme_bw() + ggtitle(NULL) + xlab('Número da observação') +
ylab('Distância de Cook')

##############################
# Selecao sem outliers: LASSO
##############################

treino_2l <- treino_1[-c(119, 121, 218, 219),]

treino_2lm <- subset(treino_2l, select = -c(mill_motor_pwr_kw_pv))

set.seed(12345)
lasso_2 <- cv.glmnet(as.matrix(treino_2lm), treino_2l$mill_motor_pwr_kw_pv,
alpha = 1)
lambda_2 <- lasso_2$lambda.min
plot(lasso_2)
coef(lasso_2, s = lambda_2)

aj_2l <- lm(mill_motor_pwr_kw_pv ~ total_feed_ton.h_pv + limestone_ton.h_pv +
mill_in_pres_mbar_sp + mill_in_pres_mbar_pv + mill_out_pres_mbar_pv +
grinding_pressure_bar_pv + thrust_pad_2_press_bar_pv +
thrust_pad_11_press_bar_pv + so3_perc_pv + clinker_ton.h_pv.1 +
gypsum_ton.h_pv.1 + fly_ash_2_ton.h_pv.1 + mill_in_pres_mbar_sp.1 +
mill_out_pres_mbar_pv.1 + roller_3_bed_depth_mm_pv.1 +
grinding_aid_addition_l.h_pv.1 + thrust_pad_1_press_bar_pv.1 +
```

```
thrust_pad_8_press_bar_pv.1 + thrust_pad_11_press_bar_pv.1 +
slag_perc_pv.2 + thrust_pad_7_press_bar_pv.2 +
thrust_pad_11_press_bar_pv.2 + thrust_pad_12_press_bar_pv.2 +
total_feed_perc_pv.3 + slag_ton.h_pv.3 + fly_ash_2_ton.h_pv.3 +
grinding_pressure_bar_pv.3 + roller_1_bed_depth_mm_pv.3 +
roller_3_bed_depth_mm_pv.3 + grinding_aid_addition_l.h_pv.3 +
so3_perc_pv.3 + limestone_perc_pv.4 + slag_ton.h_pv.4 +
gypsum_ton.h_pv.4 + main_bf_out_press_mbar_pv.4 + mill_in_pres_mbar_sp.4 +
mill_in_temp_c_pv.4 + grinding_pressure_bar_pv.4 +
gearbox_1_vibration_mm.s_pv.4 + hopper_level_mg.m3_pv.4, data = treino_2l)

validacao_m <- subset(validacao, select = -c(mill_motor_pwr_kw_pv))
pred_2l <- predict(aj_2l, newdata = validacao_m)
risco_2l <- mean((validacao$mill_motor_pwr_kw_pv-pred_2l)^2)

aj_2lp <- gnm (mill_motor_pwr_kw_pv ~ total_feed_ton.h_pv +
limestone_ton.h_pv + mill_in_pres_mbar_sp + mill_in_pres_mbar_pv +
mill_out_pres_mbar_pv + grinding_pressure_bar_pv +
thrust_pad_2_press_bar_pv +  thrust_pad_11_press_bar_pv + so3_perc_pv +
clinker_ton.h_pv.1 +  gypsum_ton.h_pv.1 + fly_ash_2_ton.h_pv.1 +
mill_in_pres_mbar_sp.1 + mill_out_pres_mbar_pv.1 +
roller_3_bed_depth_mm_pv.1 + grinding_aid_addition_l.h_pv.1 +
thrust_pad_1_press_bar_pv.1 +  thrust_pad_8_press_bar_pv.1 +
thrust_pad_11_press_bar_pv.1 + slag_perc_pv.2 +
thrust_pad_7_press_bar_pv.2 + thrust_pad_11_press_bar_pv.2 +
thrust_pad_12_press_bar_pv.2 + total_feed_perc_pv.3 + slag_ton.h_pv.3 +
fly_ash_2_ton.h_pv.3 + grinding_pressure_bar_pv.3 +
roller_1_bed_depth_mm_pv.3 + roller_3_bed_depth_mm_pv.3 +
grinding_aid_addition_l.h_pv.3 + so3_perc_pv.3 + limestone_perc_pv.4 +
slag_ton.h_pv.4 + gypsum_ton.h_pv.4 + main_bf_out_press_mbar_pv.4 +
mill_in_pres_mbar_sp.4 + mill_in_temp_c_pv.4 + grinding_pressure_bar_pv.4 +
gearbox_1_vibration_mm.s_pv.4 + hopper_level_mg.m3_pv.4,
data = validacao, constrain = "*", constrainTo = coef (aj_2l))
```

```
AIC(aj_2lp)
BIC(aj_2lp)

################################
# Selecao sem outliers: Stepwise
################################

treino_2s <- treino_1[-c(1, 119, 218, 219),]

treino_2sm <- subset(treino_2s, select = -c(mill_motor_pwr_kw_pv))

aj_2s <- lm(mill_motor_pwr_kw_pv ~ total_feed_ton.h_pv + mill_in_pres_mbar_pv +
grinding_pressure_bar_pv + fly_ash_2_ton.h_pv.1 + mill_out_pres_mbar_pv.1 +
thrust_pad_1_press_bar_pv.1 + thrust_pad_8_press_bar_pv.1 +
thrust_pad_11_press_bar_pv.1 + fly_ash_2_ton.h_pv.3 +
roller_1_bed_depth_mm_pv.3 + roller_3_bed_depth_mm_pv.3 +
limestone_perc_pv.4 + grinding_pressure_bar_pv.4 +
gearbox_1_vibration_mm.s_pv.4, data = treino_2s)

pred_2s <- predict(aj_2s, newdata = validacao_m)
risco_2s <- mean((validacao$mill_motor_pwr_kw_pv-pred_2s)^2)

aj_2sp <- gnm (mill_motor_pwr_kw_pv ~ total_feed_ton.h_pv +
mill_in_pres_mbar_pv + grinding_pressure_bar_pv + fly_ash_2_ton.h_pv.1 +
mill_out_pres_mbar_pv.1 + thrust_pad_1_press_bar_pv.1 +
thrust_pad_8_press_bar_pv.1 + thrust_pad_11_press_bar_pv.1 +
fly_ash_2_ton.h_pv.3 + roller_1_bed_depth_mm_pv.3 + roller_3_bed_depth_mm_pv.3 +
limestone_perc_pv.4 + grinding_pressure_bar_pv.4 + gearbox_1_vibration_mm.s_pv.4,
data = validacao, constrain = "*", constrainTo = coef (aj_2s))

AIC(aj_2sp)
BIC(aj_2sp)
```

```
########################################
# Analise de Residuos sem outlier: Lasso
########################################

# Erros possuem variancia constante? Sim
autoplot(aj_2l, which = 1, ncol = 1, label.size = 1) + theme_bw() +
ggtitle(NULL) + xlab('Valores ajustados') + ylab('Resíduos')

var.test(residuals(aj_2l)[treino_2l$mill_motor_pwr_kw_pv >
median(treino_2l$mill_motor_pwr_kw_pv)],
residuals(aj_2l)[treino_2l$mill_motor_pwr_kw_pv <
median(treino_2l$mill_motor_pwr_kw_pv)])

# Erros tem distribuicao normal? Sim
autoplot(aj_2l, which = 2, ncol = 1, label.size = 1)

shapiro.test(residuals(aj_2l))

# Pontos influentes (Distancia de Cook)? Pontos 1
autoplot(aj_2l, which = 4, ncol = 1, label.size = 5, label.n = 1)
max(cooks.distance(aj_2l))

##########################################
# Analise de Residuos sem outlier: Stepwise
##########################################

# Erros possuem variancia constante? Sim
autoplot(aj_2s, which = 1, ncol = 1, label.size = 1) + theme_bw() +
ggtitle(NULL) + xlab('Valores ajustados') + ylab('Resíduos')

var.test(residuals(aj_2s)[treino_2s$mill_motor_pwr_kw_pv >
median(treino_2s$mill_motor_pwr_kw_pv)],
```

```
residuals(aj_2s)[treino_2s$mill_motor_pwr_kw_pv <
median(treino_2s$mill_motor_pwr_kw_pv)])

# Erros tem distribuicao normal? Sim
autoplot(aj_2s, which = 2, ncol = 1, label.size = 1) + theme_bw() +
ggtitle(NULL) + xlab('Quantis teóricos') + ylab('Quantis observados')

shapiro.test(residuals(aj_2s))

# Pontos influentes (Distancia de Cook)? Pontos 121 e 212
autoplot(aj_2s, which = 4, ncol = 1, label.size = 5, label.n = 2)
max(cooks.distance(aj_2s))
```

Referências Bibliográficas

Akaike, H. (1974). A new look at the statistical model identification. In *Selected Papers of Hirotugu Akaike*, pages 215–222. Springer.

Apodi, C. (2011). Cimento apodi. `http://cimentoapodi.com.br/`. Acessado em: 08-04-2019.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, **19**(1), 15–18.

Cramér, H. (1999). *Mathematical methods of statistics*, volume 9. Princeton university press.

Feng, Z. Z., Yang, X., Subedi, S. e McNicholas, P. D. (2012). The lasso and sparse least squares regression methods for snp selection in predicting quantitative traits. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, **9**(2), 629–636.

Izbicki, R. e Santos, T. (2018). Machine learning sob a ótica estatística: uma abordagem preditivista para a estatística com exemplos em r, 2018 [versão em desenvolvimento].

James, G., Witten, D., Hastie, T. e Tibshirani, R. (2013). *An introduction to statistical learning*, volume 112. Springer.

Jorgensen, S. W. (2004). Cement grinding vertical roller mills versus ball mills. In *13th Arab-International Cement Conference and Exhibition*.

Kohavi, R. *et al.* (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada.

Morettin, P. A. e BUSSAB, W. O. (2017). *Estatística básica*. Editora Saraiva.

Neter, J., Kutner, M. H., Nachtsheim, C. J. e Wasserman, W. (1996). *Applied linear statistical models*, volume 4. Irwin Chicago.

Pal, S. K. e Mitra, S. (1992). Multilayer perceptron, fuzzy sets, and classification. *IEEE Transactions on neural networks*, **3**(5), 683–697.

Schwarz, G. *et al.* (1978). Estimating the dimension of a model. *The annals of statistics*, **6**(2), 461–464.

Shapiro, S. S. e Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**(3/4), 591–611.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**(1), 267–288.

Wilk, M. B. e Gnanadesikan, R. (1968). Probability plotting methods for the analysis for the analysis of data. *Biometrika*, **55**(1), 1–17.