FERRAMENTAS GRÁFICAS NO PROCESSO DE SELEÇÃO DE VARIÁVEIS

Lucas Ragiotto

Dissertação apresentada à Universidade Estadual Paulista "Júlio de Mesquita Filho" para a obtenção do título de Mestre em Biometria.

BOTUCATU
São Paulo - Brasil
Março – 2019

FERRAMENTAS GRÁFICAS NO PROCESSO DE SELEÇÃO DE VARIÁVEIS

Lucas Ragiotto

Orientadora: Prof. Dr. **Luzia Aparecida Trinca**

Dissertação apresentada à Universidade Estadual Paulista "Júlio de Mesquita Filho" para a obtenção do título de Mestre em Biometria.

BOTUCATU
São Paulo - Brasil
Março – 2019

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO TÉCNICA DE BIBLIOTECA E DOCUMENTAÇÃO - CÂMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: ROSANGELA APARECIDA LOBO-CRB 8/7500

Ragiotto, Lucas.
Ferramentas gráficas no processo de seleção de variáveis / Lucas Ragiotto. - Botucatu, 2019

Dissertação (mestrado) - Universidade Estadual Paulista "Júlio de Mesquita Filho", Instituto de Biociências de Botucatu
Orientador: Luzia Aparecida Trinca
Capes: 90100000

1. Bootstrap (Estatística). 2. Modelos lineares. 3. Modelos logísticos. 4. Variáveis (Matemática).

Palavras-chave: Bootstrap; Critério de informação generalizado; Método de cerca; Regressão linear; Regressão logística.

Agradecimentos

Aos meus pais, Osmar e Andrea, e irmã, Luhanna, por todo o incentivo e carinho.

A Aline, que todos os dias, sem exceção, me apoiou, ouviu e aconselhou.

À Professora Doutora Luzia Trinca, minha orientadora, pela compreensão, amizade, disposição, incentivo e oportunidade.

À Professora Doutora Berenice Camargo Damasceno e o Professor Doutor Luciano Barbanti, meus orientadores da graduação, que me aconselharam e auxiliaram.

Aos amigos do departamento de Bioestatística, entre eles os funcionários, professores e estudantes.

Às Professoras Doutoras Júlia Maria Pavan Soler e Miriam Harumi Tsunemi pelas contribuições.

À Fundação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES pelo apoio financeiro, que foi essencial para o desenvolvimento de toda a pesquisa realizada durante o mestrado.

Novamente, agradeço por todo suporte. Todos foram essenciais!

Sumário

Página

LISTA DE FIGURAS **vi**

LISTA DE TABELAS **viii**

RESUMO **x**

SUMMARY **xii**

1 INTRODUÇÃO **1**

2 SELEÇÃO DE VARIÁVEIS EM MODELOS DE REGRESSÃO **3**

2.1 Modelos Lineares 3

2.2 Modelo de regressão Binomial 8

2.3 Diagnósticos 14

2.4 Seleção de Variáveis 19

2.4.1 Métodos e critérios usuais 19

2.4.2 O método *fence* 23

2.4.3 Ferramentas gráficas 24

3 MATERIAL E MÉTODOS **28**

3.1 Aplicação 1: estudo sobre peso de recém-nascidos (RN) prematuros . . . 28

3.2 Aplicação 2: probabilidade de prenhez em inseminação artificial (IA) de vacas 32

4 RESULTADOS **39**

4.1 Aplicação 1: estudo sobre o peso de recém-nascidos (RN) prematuros . . 39

4.2 Aplicação 2: probabilidade de prenhez em inseminação artificial (IA) de vacas 53

5 CONSIDERAÇÕES FINAIS **69**

ANEXOS **72**

REFERÊNCIAS BIBLIOGRÁFICAS **76**

Lista de Figuras

Página

1 Valores da correlação entre as regressoras quantitativas e a resposta peso, dados da Aplicação 1 ($n = 90$) 30

2 Valores da correlação entre as regressoras quantitativas e da resposta por lote, dados da Aplicação 2 ($N = 65$) 38

3 Gráfico de resíduos do modelo de efeitos principais ajustado para a variável - peso de RN: envelope quantil-quantil Normal (à esquerda) e resíduos em função dos valores ajustados (à direita). 40

4 Gráfico de resíduos do modelo de efeitos principais ajustado para a variável - ln(*peso*) de RN: envelope quantil-quantil Normal (à esquerda) e resíduos em função dos valores ajustados (à direita). 40

5 Gráfico para diagnóstico de influência do ajuste do modelo de efeitos principais para ln(*peso*) de RN. O raio dos círculos são proporcionais à distância de Cook. 41

6 Gráfico de inclusão de variáveis (VIP) em modelos de efeitos principais para ln(*peso*) de RN, com destaque da curva de RV. 43

7 Gráficos do valor do componente de perda contra a dimensão do modelo. Modelos de efeitos principais para ln(*peso*) de RN. Destaque para as regressoras `IG` (à esquerda), `GR` (à direita) e `pré_natal` (inferior). . . . 44

8 Gráfico de probabilidades de seleção de modelos em função do número de parâmetros com destaque para a regressora pré-natal. Modelos de efeitos principais para ln(*peso*) de RN, utilizando *bootstrap* ponderado. 46

9 Gráficos de probabilidades de seleção de modelos usando o método *fence*. Modelos de efeitos principais para o ln(*peso*) de RN. 47
10 Gráficos VIP dos termos de interação para os modelos 2 (à esquerda) e 3 (à direita) para o ln(*peso*) de RN. 49
11 Retas ajustadas para o ln(*peso*) de RN segundo o modelo 2. 51
12 Retas ajustadas para o ln(*peso*) de RN segundo o modelo 3. 52
13 Gráficos de diagnóstico para o ajuste do modelo de efeitos principais para a probabilidade de prenhez sob IA, dados agregados completos (N= 72). 54
14 Gráficos de diagnóstico para o ajuste do modelo de efeitos principais para a probabilidade de prenhez sob IA, dados agregados reduzidos (N= 65). . 55
15 Gráfico de inclusão de variáveis (VIP) para o modelo linearizado. 58
16 Gráfico de probabilidade de seleção de modelos em função do número de parâmetros, com destaque para as regressoras, visualizadas da esquerda para a direita e de cima a baixo, `strr`, `linr`, `motr`, `alhr` e `progrr`, dados linearizados. 60
17 Gráfico de probabilidade de seleção de modelos, dados linearizados. . . . 62
18 Gráficos VIP para a interação dos modelos 3 (à esquerda) e 5 (à direita), dados na Tabela 15, dados linearizados. 63
19 Gráficos VIP para a interação dos modelos 4 (à esquerda) e 6 (à direita), dados na Tabela 15, dados linearizados. 64
20 Diagnóstico do modelo com efeitos da variável de blocos, `str` e `mot`, dados reduzidos (N= 65). 66
21 Diagnóstico do modelo com efeitos da variável de blocos, `str`, `mot` e `lin`, dados reduzidos (N= 65). 67

Lista de Tabelas

Página

1 Tabela ANOVA 7

2 Exemplo da ANODEV com um modelo que contém duas regressoras contínuas e sua interação 12

3 Medidas descritivas das variáveis quantitativas, dados da Aplicação 1 ($n = 90$) 29

4 Medidas descritivas para as variáveis binárias, dados da Aplicação 1 ($n = 90$) 30

5 Medidas descritivas das variáveis quantitativas (características do sêmem), dados da Aplicação 2 ($N = 65$) 36

6 Medidas descritivas da variável resposta por lote, dados da Aplicação 2 ($N = 65$) 37

7 ANOVA do modelo de regressão ajustado para o peso de RN 39

8 ANOVA do modelo de regressão ajustado com $y = \ln(peso)$ de RN 40

9 Modelos selecionados pelos métodos usuais segundo o critério, para o $\ln(peso)$ de RN 42

10 Modelos com probabilidade de seleção superior a $0,20$ visualizados na Figura 8. Modelos de efeitos principais para $\ln(peso)$ de RN 46

11 Modelos selecionados conforme aplicação do `mplot` para $\ln(peso)$ de RN . 48

12 Estimativas dos parâmetros e intervalos de confiança (95%) para os ajustes dos modelos selecionados (Tabela 11) para $\ln(peso)$ de RN 49

13 Regressoras presentes nos modelos selecionados segundo os métodos usuais de seleção de varíveis, dados de prenhez sob IA agregados (N= 65) . . 56

14 Regressoras presentes no modelo linearizado segundo os métodos usuais para seleção de variáveis, dados de prenhez sob IA 57
15 Modelos com destaque probabilístico, conforme Figura 16, dados linearizados 61
16 Modelos com destaque probabilístico, encontrados no gráfico de probabilidades de seleção de modelos em função do número de parâmetros (não apresentado) 64
17 Parte preditiva dos modelos utilizando-se o `mplot`, dados linearizados . . 65
18 Regressoras presentes nos modelos finais, feita a seleção de variáveis com a metologia usual do modelo linearizado 65
19 Estimativas e IC's (95%) para os parâmetros do ajuste final, dados reduzidos (N= 65) 66
20 Estimativas da razão de chances e respectivas IC's (95%), para o modelo incluindo a variável de blocos, `mot`, `str` e `lin`, para os dados de prenhez sob IA 68

FERRAMENTAS GRÁFICAS NO PROCESSO DE SELEÇÃO DE VARIÁVEIS

Autor: LUCAS RAGIOTTO

Orientadora: Prof. Dr. LUZIA APARECIDA TRINCA

RESUMO

Em problemas de regressão, na busca por um modelo parcimonioso, o pesquisador pode se deparar com adversidades, por exemplo, a existência de colinearidade entre as regressoras, dificultando a seleção de variáveis. Dessa forma, com a implementação de ferramentas inspiradas nas propostas de Murray et al. (2013), Müller & Welsh (2010) e Jiang et al. (2009) no pacote `mplot` (Tarr et al., 2018) no *software* `R`, pode-se, gráfica e interativamente, estudar em detalhes a estabilidade e a importância de inclusão de covariáveis para a construção de modelos. Neste trabalho, medidas de estabilidade e probabilidade de inclusão de variáveis foram obtidas pelo método *bootstrap*. Medidas resumo de qualidade do ajuste são baseadas no critério de informação generalizado, que incorpora, como casos particulares, os critérios de informação de Akaike e o Bayesiano, e reflete a perda (associada ao ajuste de um modelo simplificado) mais uma penalização à complexidade do modelo. Ao aplicar

a teoria de seleção de variáveis, utilizando as ferramentas gráficas no ajuste de um modelo de regressão linear Normal e regressão Binomial, foi possível reconhecer seu potencial e utilidade no processo de formulação de modelos, no qual a incorporação de conhecimento do especialista da área pode ser feita de maneira natural, já que o processo não é automático. Isso é mais um diferencial em relação aos métodos usuais de seleção de variáveis que também foram aplicados aos mesmos conjuntos de dados para efeito de discussão.

GRAPHICAL TOOLS IN THE PROCESS OF VARIABLE SELECTION

Author: LUCAS RAGIOTTO

Adviser: Prof. Dr. LUZIA APARECIDA TRINCA

SUMMARY

In regression analysis, the search of a parsimonious model can be difficult due to collinearities among variables and other problems. Murray et al. (2013), Müller & Welsh (2010) e Jiang et al. (2009) proposed tools for model stability and variable inclusion plots that were refined and implemented in the `mplot` package of Tarr et al. (2018), which allows interactive graphs and summaries of information relevant to model building. Stability measures and the probability of variable inclusion are obtained through bootstrapping. Goodness of fit measures are based on the generalized information criterion, which includes as particular cases the Akaike and Bayesian information criteria, given by a measure of loss of the fit and a penalization due to model complexity. Applying the method to fit a Normal linear regression and a Binomial regression revealed its great potential and usefulness for model building, allowing expertise knowledge to be incorporated since the selection model is not au-

tomated. This is a further contrast to the usual selection methods which were also applied to the same datasets in order to discuss the differences.

1 INTRODUÇÃO

Conhecimento científico em diversas áreas é adquirido através da coleta e análise estatística de dados que, frequentemente, tem suporte num modelo. Conforme bem colocado por Davison (2003), a formulação do modelo envolve sensatez, experiência, tentativa e erro. Não raramente o pesquisador pode se deparar com vários modelos competitivos para o problema em mãos. Um princípio utilizado na escolha de um modelo é o da parcimônia que favorece modelos simples em detrimento de complexos quando as qualidades de seus ajustes são aproximadamente equivalentes. No caso de modelos de regressão linear múltipla ou linear generalizado, a aplicação do princípio da parcimônia nem sempre é imediata ou fácil e existem diferentes estratégias de visitas aos possíveis modelos, assim como vários critérios de medidas de qualidade do ajuste. Devido ao grande número de modelos que precisam ser avaliados, as estratégias usuais são automatizadas e seguem um critério de ordenação dos modelos pré-especificado. As estratégias mais populares são os métodos tipo *stepwise* e o *all subsets*, que avalia todos os subconjuntos possíveis. Embora úteis, tais estratégias vêm sofrendo críticas, principalmente pela não incorporação do conhecimento do pesquisador especialista da área de aplicação. Outra crítica é a instabilidade dos métodos que podem levar a modelos diferentes quando há pequenas perturbações nos dados. Uma terceira dificuldade é que, embora todas as estratégias sejam baseadas na parcimônia, cada uma pode levar a um modelo distinto, principalmente na presença de colinearidade entre as variáveis regressoras (Tarr et al., 2018).

Com o intuito de proporcionar melhor entendimento sobre a importância relativa aos diversos modelos estatísticos possíveis para explicar determi-

nada característica ou fator, posteriormente chamada de resposta, Müller & Welsh (2010) e Murray et al. (2013) propuseram os gráficos de inclusão de variáveis e medidas resumo da qualidade do ajuste, variando-se a penalidade imposta à complexidade do modelo. Tarr et al. (2018) implementaram essas ideias e agregaram outras metodologias no pacote `mplot` do R (R Core Team, 2017), com excelentes recursos para construção de gráficos dinâmicos, permitindo o estudo de estabilidade de modelos e a exploração detalhada da importância de cada variável regressora. Sob a metodologia proposta por estes autores, a obtenção de medidas de estabilidade e probabilidades de inclusão de variáveis é realizada via reamostragem pelo método *bootstrap*.

O objetivo dessa dissertação é revisar e explorar o uso desses recursos gráficos na formulação de modelos de regressão, assim como levantar as possíveis limitações do pacote `mplot`, lançando mão de duas aplicações. Uma delas é no estudo da relação entre o peso de recém-nascidos prematuros em função de outras covariáveis relacionadas à gestação e a outra é no estudo da relação entre a probabilidade de prenhes de vacas inseminadas artificialmente em função de diversas características do sêmen. No primeiro caso, usa-se o modelo linear Normal e no segundo um modelo de regressão logística, que envolveu diversos desafios como a inclusão de um fator de blocos, não submetido ao processo de seleção, e inclusão de termos de interação entre regressoras. Para fins de discussão, os métodos de seleção de variáveis usuais também foram aplicados em ambos conjuntos de dados.

O trabalho está organizado em capítulos. No Capítulo 2 apresenta-se uma breve introdução aos modelos de regressão linear e logística e aos métodos de seleção de variáveis, incluindo as ferramentas disponibilizadas no `mplot`. O Capítulo 3 descreve os conjuntos de dados das duas aplicações e apresenta os métodos utilizados para a aplicação das ferramentas de seleção. Os resultados são apresentados e discutidos no Capítulo 4. As considerações finais se encontram no Capítulo 5.

2 SELEÇÃO DE VARIÁVEIS EM MODELOS DE REGRESSÃO

2.1 Modelos Lineares

A investigação, análise e interpretação de características tem fator importante nas decisões de diversas áreas de estudo. Devido à evolução tecnológica e computacional, é possível observar tais características e coletar suas informações com certa praticidade. Uma técnica estatística que auxilia na investigação e modelagem dos problemas que envolvem essas características, chamadas de variáveis, é definida como análise de regressão. Seu objetivo é investigar a contribuição de certas variáveis na variação de uma variável de interesse, chamada de resposta. Existe uma classe ampla de modelos de regressão, sendo que a sub-classe dos modelos lineares oferece grande potencial na condução deste tipo de estudo.

Para a aplicação desta teoria e construção de um modelo matemático que visa relacionar as variáveis define-se n como o número de realizações do experimento ou amostra, $\mathbf{Y}$ o vetor $(n \times 1)$ aleatório cujas observações serão consideradas respostas e $\mathbf{x}_j$ um vetor $(n \times 1)$ da j-ésima variável $(j = 1, \ldots, k)$ a qual deseja-se analisar sua relação com a resposta, chamada de covariável, regressora ou variável independente. As variáveis regressoras podem ser qualitativas ou quantitativas. Os coeficientes da regressão ou parâmetros, β_0, β_1, ..., β_{p-1}, são constantes estimáveis, no qual β_0 é o intercepto e β_1, ..., β_{p-1} são os coeficientes das regressoras, alocadas em um vetor $\boldsymbol{\beta}$ $(p \times 1)$, com $p = k+1$. Cada β_j $(j = 1, 2, \ldots, k)$ representa a variação nos valores esperados da resposta conforme há mudança de uma unidade na regressora $\mathbf{x}_j$, desde que os valores das outras estejam fixados. Os vetores das covariáveis

são organizados em um matriz $\mathbf{X}$ $(n \times p)$, com $n > p$, cuja primeira coluna tem seus valores iguais a 1 no modelo com intercepto. Visto que vários fatores não observados na amostra ou experimento contribuem na oscilação dos valores da resposta, deve-se incluir um erro, neste caso denotado por $\boldsymbol{\varepsilon}$ $(n \times 1)$. Dado os componentes de um modelo linear, tem-se sua estrutura, definida da forma

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{1}$$

Note que o termo "linear" é justificado pela forma que o vetor de parâmetros aparece no modelo. Quando o modelo contém apenas uma regressora quantitativa ele é conhecido também como regressão linear simples e com isso os valores de $\boldsymbol{\beta}$ a serem estimados são o intercepto e o coeficiente da reta.

Quando o modelo linear especificado em (1) pressupõe que $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ ele é chamado de Normal, no qual $\mathbf{I}$ é a matriz identidade $(n \times n)$ indicando que o erro se distribui normalmente com média e correlação 0 e variância constante (homocedasticidade). Correlação nula e normalidade implicam em independência. Consequentemente, tem-se

$$E(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta} \quad \text{e} \quad Var(\mathbf{Y}|\mathbf{X}) = \sigma^2\mathbf{I}, \tag{2}$$

então, $\mathbf{Y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$.

Nessa perspectiva, deve-se estimar os parâmetros do modelo e para tal pode-se utilizar de dois métodos que conduzem ao mesmo resultado. Demonstrações e apresentação dessas metodologias de ajuste do modelo, propriedades dos estimadores e exploração da qualidade do ajuste estão presentes em inúmeros livros de autores clássicos como Montgomery et al. (2012), Draper & Smith (2014), Charnet et al. (2015) e Seber & Lee (2012), por exemplo. Um dos métodos de estimação é o dos "mínimos quadrados" que minimiza a soma dos erros ao quadrado. O outro método maximiza a função de verossimilhança da amostra $\mathbf{y}$ que é dada pelo produto das funções de densidades Normais de Y_i avaliadas nos valores y_i $(i = 1, \ldots, n)$. Sob as pressuposições do modelo em (1), para ambos os métodos, o estimador de $\boldsymbol{\beta}$ é

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

para $\mathbf{X}$ de posto completo.

Por conseguinte, é imediato verificar que

$$E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

e

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1},$$

ou seja, $\hat{\boldsymbol{\beta}}$ é não viciado para $\boldsymbol{\beta}$ e sua variância depende da matriz do modelo $\mathbf{X}$ e da variância do erro σ^2.

No entanto, vale notar que a propriedade de não tendenciosidade de $\hat{\boldsymbol{\beta}}$ só é válida se a especificação de $E(\mathbf{Y}|\mathbf{X})$ em (2) for correta. No caso do modelo real ser $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\varepsilon}$ mas se na fase de ajuste, o segundo termo $(\mathbf{A}\boldsymbol{\theta})$ for ignorado, então $E(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\theta}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}\boldsymbol{\theta}$ e $\hat{\boldsymbol{\beta}}$ só é não viciado para $\boldsymbol{\beta}$ se $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{A} = \mathbf{0}$, ou seja, $\mathbf{A}$ e $\mathbf{X}$ são ortogonais, o que raramente acontece em estudos observacionais.

O vetor dos valores ajustados $\hat{y}_i$'s, $i = 1, \ldots, n$, é dado por

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y} \tag{3}$$

em que, $\mathbf{y}$ é o vetor de observações da variável resposta $\mathbf{Y}$ e $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, de dimensão $(n \times n)$, conhecida como matriz de projeção ortogonal já que ela projeta, ortogonalmente, $\mathbf{y}$ em $\hat{\mathbf{y}}$. Logo, a diferença entre os vetores de valores ajustados e observados de $\mathbf{Y}$, definido como resíduo, é $\hat{\boldsymbol{\varepsilon}} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$.

Para a estimação de σ^2, utiliza-se a soma dos quadrados dos resíduos dada por

$$\sum_{i=1}^{n} \hat{\varepsilon}_i^2 = SQ_{res} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}$$

que, também pode ser escrita como

$$SQ_{res} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}.$$

Sendo assim, a estimativa não viciada de σ^2 é dada por

$$\hat{\sigma}^2 = \frac{SQ_{res}}{n-p}.$$

Logo, $\widehat{Var}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$.

Uma vez estimados os parâmetros, deve-se buscar pela boa adequação do modelo e identificar as regressoras de importância. Para tal, os pressupostos do modelo devem ser satisfeitos e a literatura recomenda diversas técnicas exploratórias e gráficas. Na Seção 2.3, um resumo das técnicas que auxiliam nesse diagnóstico é apresentado.

Para testar a significância da regressão, ou seja, determinar se existe relação linear entre a resposta e as covariáveis, utiliza-se o teste de hipótese global associado à tabela ANOVA, o qual testa a hipótese nula $H_0 : \beta_1 = .. = \beta_{p-1} = \mathbf{0}$ contra a hipótese alternativa $H_1 : \beta_j \neq 0$, para algum j, e busca-se evidências de que ao menos uma regressora é significante na explicação da resposta. Para sua construção define-se a tabela ANOVA, apresentada na Tabela 1, cujas quantidades são obtidas por

$$SQ_{reg} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n},$$

$$SQ_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y},$$

$$SQ_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n},$$

nas quais SQ_{reg} ou Soma dos Quadrados da Regressão é a soma de quadrados dos valores de $\hat{\mathbf{y}}$, SQ_T é a soma de quadrados dos valores de $\mathbf{y}$, ambas corrigidas pela média, e $SQres$ ou Soma dos Quadrados dos Resíduos é a diferença entre as duas primeiras. Sob normalidade e sob H_0, cada soma de quadrados está associada a uma distribuição qui-quadrado central com parâmetros específicos. Os graus de liberdade de SQ_{reg}, SQ_{res} e SQ_T são $p-1$, $n-p$ e $n-1$, respectivamente. Os quadrados médios são obtidos pela divisão da soma dos quadrados pelos números de graus de liberdade (G.L.) correspondentes. Assim, a estatística $\frac{QM_{reg}}{QM_{res}}$, sob H_0, segue a distribuição F

de Snedecor com $p-1$ e $n-p$ graus de liberdade. Se seu valor observado, F_0, for superior ao quantil de $F_{1-\alpha;\ p-1;\ n-p}$ da distribuição de referência, então, H_0 é rejeitada, ao nível α de significância.

Tabela 1. Tabela ANOVA

Variação	Soma dos Quadrados	G.L.	Quadrado Médio	F_0
Regressão	SQ_{reg}	$p-1$	QM_{reg}	QM_{reg}/QM_{res}
Resíduo	SQ_{res}	$n-p$	QM_{res}	
Total	SQ_T	$n-1$		

Testes de hipóteses também podem ser feitos para cada parâmetro, nos quais a contribuição de uma regressora dado as outras no modelo é avaliada. Estes se baseiam na estatística $T = \hat{\beta}_j / \widehat{EP}(\hat{\beta}_j)$, que sob H_0 segue a distribuição *t-Student* com $n-p$ graus de liberdade, em que $\widehat{EP}(\hat{\beta}_j)$ é a estimativa do erro-padrão de $\hat{\beta}_j$, dado pela raiz quadrada do valor da diagonal principal de $\widehat{Var}(\hat{\boldsymbol{\beta}})$ na posição $j+1$ para $j = 0, \ldots, p$.

Intervalos de confiança também são construídos para os parâmetros. O intervalo com $100(1-\alpha)\%$ de confiança para o j-ésimo coeficiente β_j da regressão (parâmetro) é expresso da forma

$$\hat{\beta}_j \pm |t_{\alpha/2,\ n-p}|\ \widehat{EP}(\hat{\beta}_j),$$

com $j = 0, \ldots, p$ e $t_{\alpha/2,\ n-p}$ o quantil $\alpha/2$ da distribuição *t-Student* com $n-p$ graus de liberdade.

O intervalo de confiança de $100(1-\alpha)\%$ para o valor esperado da resposta quando as regressoras assumem um ponto $\mathbf{x}_0$, $E(\mathbf{Y}|\mathbf{x}_0)$, no espaço das regressoras é dado por

$$\hat{y}_0 \pm |t_{\alpha/2,\ n-p}|\ \sqrt{\hat{\sigma}^2 \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0},$$

com $\mathbf{x}_0' = [1, x_{01}, x_{02}, \ldots, x_{0k}]$ sendo um ponto particular. O mesmo pode ser calculado para uma nova observação, y_0, logo um intervalo de predição de $100(1-\alpha)\%$ é

dado por

$$\hat{y}_0 \pm |t_{\alpha/2,\ n-p}| \sqrt{\hat{\sigma}^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)}.$$

Todo ajuste de modelo de regressão deve ser submetido à análise de resíduos e diagnóstico afim de validar os pressupostos do modelo. Na Seção 2.3 apresenta-se as principais ferramentas gráficas para esses diagnósticos.

Todavia, o modelo linear Normal nem sempre é uma ferramenta viável na construção de modelos de predição, pois a variável resposta pode ser caracterizada por observações do tipo binárias ou de contagem e não se adequa às suposições feitas no modelo em (1). Outra dificuldade encontrada é que a função que relaciona a resposta esperada e $\mathbf{X}\boldsymbol{\beta}$ não se apresenta de forma linear. Uma alternativa de modelar os dados via modelos lineares é através da aplicação de transformações na resposta de forma a aproximar à normalidade. Felizmente, na década de 1970, Nelder & Wedderburn (1972) introduziriam os modelos lineares generalizados (MLG), que englobam diversas distribuições para a variável resposta dentro da família exponencial. Sua teoria pode ser estudada com detalhes em Paula (2013), McCullagh & Nelder (1989), Hosmer Jr et al. (2013), Chatfield et al. (2010) e Dobson & Barnett (2018). Neste trabalho, devido à ilustração utilizada exigir um modelo que lida com uma resposta do tipo binária, apresenta-se, na Seção 2.2, um resumo dos principais tópicos para ajuste de um modelo deste tipo.

2.2 Modelo de regressão Binomial

É comum, principalmente nas áreas médicas e biológicas, a coleta de dados cujas variáveis resposta são do tipo categórico. Existem diversas estratégias de estudo para dados deste tipo, como, por exemplo, análise de tabelas de contingência e alguns tipos de MLG's. A acomodação de variáveis explanatórias quantitativas é mais natural dentro da classe de modelos de regressão generalizado. No caso de respostas binárias o modelo de regressão Binomial, um caso particular de MLG, permite modelar a probabilidade de sucesso em um ensaio básico via um preditor. Para seu ajuste os dados podem se apresentar de duas formas: a resposta na ob-

servação i é binária ($Y_i = 1$ ou $Y_i = 0$) ou a resposta é a contagem ou proporção amostral do número de sucessos observada em um agregado e existem vários agregados. A agregação é possível quando existem repetições do conjunto de valores das covariáveis. Em ambos os casos, o interesse está em modelar a probabilidade de sucesso, e as estimativas dos parâmetros são equivalentes embora algumas medidas de diagnóstico do ajuste possam não ter interpretações idênticas. Na exposição que segue não se fará distinção entre os casos. A suposição é a de que os ensaios que geram as variáveis binárias são independentes, mas não identicamente distribuídas, já que a probabilidade de sucesso depende das regressoras, assim como o parâmetro de cada agregado que depende do número de repetições.

Da mesma maneira como estudado em modelos lineares, o modelo de regressão Binomial tem o objetivo de modelar/estimar o valor esperado da variável resposta dadas as observações das regressoras. Nesse caso a esperança da variável binária se encontra entre os valores $[0, 1]$ e a variância depende da sua esperança, violando duas das suposições no modelo em (1) (o espaço do preditor linear é o conjunto dos números reais e homogeneidade de variâncias). Para que o preditor linear $\mathbf{X}\boldsymbol{\beta}$ componha o modelo para $E(Y)$, sem que os valores da estimação dos parâmetros sejam limitados, pode-se utilizar a função logística. A função logística transforma o preditor no intervalo requerido, dado que sua forma é sigmoidal com assíntotas em 0 e 1, conforme explicado em Giolo (2017). Todavia, existem outras funções que satisfazem tais necessidades para este tipo de dados, como a probito e complemento log-log (Collett, 2002; Hosmer Jr et al., 2013; Agresti, 2012; Paula, 2013). Como a função logística é a mais popular, apenas ela será considerada na sequência desta exposição.

Para uma variável aleatória Y, dado um conjunto de regressoras $\mathbf{x}$, tem-se que $E(Y|\mathbf{x}) = P(Y = 1|\mathbf{x}) = \pi(\mathbf{x})$. Na regressão logística, $\pi(\mathbf{x})$ é modelado por

$$\pi(\mathbf{x}) = \frac{exp(\boldsymbol{\beta}'\mathbf{x})}{1 + exp(\boldsymbol{\beta}'\mathbf{x})}. \tag{4}$$

Agora $\boldsymbol{\beta}$, o vetor de coeficientes da regressão, continua de dimensão p e $\mathbf{x}$ é um

vetor coluna cujos valores são 1 e demais valores observados para as k regressoras. Enquanto $\pi(\mathbf{x})$ fornece a probabilidade de ocorrer determinado evento $(Y = 1|\mathbf{x})$, $1 - \pi(\mathbf{x})$ torna-se seu complemento, ou seja, a probabilidade de não ocorrência de tal evento $(Y = 0|\mathbf{x})$. Consequentemente, a variância de $Y|\mathbf{x}$ é $\pi(\mathbf{x})(1 - \pi(\mathbf{x}))$, explicitando a dependência da variância em $\mathbf{x}$.

A razão entre as probabilidades de sucesso e fracasso é definida como chance. Ao aplicar o logaritmo na chance, obtém-se a função logito, que está diretamente associada ao preditor linear, ou seja,

$$\ln\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right] = \boldsymbol{\beta}'\mathbf{x}. \tag{5}$$

A equação (5), no contexto dos modelos lineares generalizados, é vista como uma função de ligação canônica associada ao modelo Binomial.

A estimação dos parâmetros do modelo de regressão Binomial é feita pelo método de máxima verossimilhança. A função de verossimilhança para o modelo Binomial é o produto de funções de probabilidades Binomiais dada por

$$L(\boldsymbol{\beta};\mathbf{y}) = \prod_{i=1}^{N} \begin{pmatrix} n_i \\ y_i \end{pmatrix} \pi_i^{y_i}(1-\pi_i)^{n_i - y_i}$$

em que $\pi_i = \pi(\mathbf{x}_i)$, N é o número de grupos ou agregados e n_i $(i = 1, \ldots, N)$ o número de repetições do agregado i. Nota-se que as probabilidades de sucesso π_i dependem dos $\boldsymbol{\beta}$'s pela equação (4), então a função de verossimilhança é reescrita em função do preditor linear $\boldsymbol{\beta}'\mathbf{x}_i$. Com a maximização de $L(\boldsymbol{\beta};\mathbf{y})$, ou equivalentemente $\ln L(\boldsymbol{\beta};\mathbf{y})$, obtém-se os estimadores dos parâmetros. Nesse sentido, tem-se que

$$\ln L(\boldsymbol{\beta};\mathbf{y}) = \sum_i \left\{ \ln \begin{pmatrix} n_i \\ y_i \end{pmatrix} + y_i \ \ln(\pi_i) + (n_i - y_i) \ \ln(1-\pi_i) \right\} =$$

$$= \sum_i \left\{ \ln \begin{pmatrix} n_i \\ y_i \end{pmatrix} + y_i \ \ln\left(\frac{\pi_i}{1-\pi_i}\right) + n_i \ \ln(1-\pi_i) \right\} =$$

$$= \sum_i \left\{ \ln \begin{pmatrix} n_i \\ y_i \end{pmatrix} + y_i \ \eta_i - n_i \ \ln(1+e^{\eta_i}) \right\}$$

com $\eta_i = \ln\left[\frac{\pi_i}{1-\pi_i}\right] \overset{(5)}{=} \boldsymbol{\beta}'\mathbf{x}_i = \sum_{j=0}^{k} \beta_j x_{ji}$ e $x_{0i} = 1$ para todo valor de i. A maximização ocorre quando as derivadas de $\ln L(\boldsymbol{\beta}; \mathbf{y})$ são igualadas a zero, em relação a cada β_j, fornecendo o estimador do parâmetro j. Devido à presença de covariáveis esse processo de maximização não tem solução fechada, sendo então necessário a utilização de um método iterativo, por exemplo, o algoritmo conhecido como método de *Score* de Fisher, para obter $\hat{\boldsymbol{\beta}}$ (Collett, 2002). Desenvolvendo essa metodologia, chega-se à expressão usual ao do método de mínimos quadrados reponderados. Para a função logística, os valores da variável dependente $z_i = \eta_i + (y_i - n_i\pi_i)/[n_i\pi_i(1-\pi_i)]$ são ajustadas pelas k regressoras em cada iteração, usando os pesos da diagonal principal da matriz $\mathbf{V}$ $(N \times N)$, expressos como

$$\upsilon_i = n_i\pi_i(1 - \pi_i), \tag{6}$$

com $i = 1, \ldots, N$. Nota-se que os elementos da diagonal de $\mathbf{V}$ correspondem à variância de uma variável com distribuição Binomial (n_i, π_i). Portanto, após um certo número de iterações, ao ocorrer a covergência, obtém-se os valores de $\hat{\boldsymbol{\beta}}$ (Collett, 2002).

Uma medida resumo da qualidade do ajuste de um determinado modelo M é dada pela função desvio (Giolo, 2017), definida por

$$D_M = 2\left\{\ln L_S - \ln L_M\right\}$$

na qual L_S é a verossimilhança do modelo saturado, ou seja, que possui tantos parâmetros quantas observações existirem e L_M é a verossimilhança do modelo M que apresenta menor número de parâmetros do que S, ambas avaliadas nos estimadores de máximas verossimilhanças. Fazendo uma relação com o modelo linear Normal, naquele caso, D_M coincide com a $SQ_{res}(M)$ (Soma dos Quadrados dos Resíduos segundo o modelo M).

Conforme Collett (2002) esclarece, o estimador da função desvio D_M, por ser obtida da estatística ln da razão de verossimilhanças, segue, assintoticamente, a distribuição qui-quadrado com $N - p_M$ graus de liberdade (p_M = número

de parâmetros do modelo M) no caso de dados agregados. Assintoticamente quer dizer que $n_i \to \infty$, ou seja, o número de repetições por agregado deve ser muito grande ($i = 1, 2, \ldots, N$) para satisfazer a condição. Assim, muitos pesquisadores utilizam D_M como uma medida da falta de ajuste do modelo M.

No caso de dados binários Collett (2002) mostra que a função desvio não é um critério adequado como medida de falta de ajuste já que ela depende das observações apenas através dos estimadores $\hat{\pi}(\mathbf{x}_i)$, e que não é possível associá-la à distribuição qui-quadrado.

Por meio da função desvio, é possível generalizar a tabela ANOVA (Tabela 1) para os GLM's que agora é chamada de tabela ANODEV (Tabela 2). Sua utilidade está na avaliação do(s) efeito(s) de uma (ou mais) regressora(s) na presença de outra(s), sequencialmente, ou seja, partindo-se do modelo mais simples para algum mais complicado. Ilustra-se, na Tabela 2, o caso no qual um efeito é avaliado dado as outras regressoras presentes no modelo anterior. Sua contribuição é medida pela diferença entre os valores da função desvio dos dois modelos. No caso de modelos encaixados, ou seja, o mais simples é um caso particular do mais complexo, pode-se usar o teste da razão de verossimilhanças (Collett, 2002).

Tabela 2. Exemplo da ANODEV com um modelo que contém duas regressoras contínuas e sua interação

Modelo	Diferença G.L.	Função desvio	Diferença
Nulo		D_N	
X_1	1	D_{X_1}	$D_N - D_{X_1}$
$X_2 \vert X_1$	1	D_{X_1,X_2}	$D_{X_1} - D_{X_1,X_2}$
$X_1 \times X_2 \vert X_1, X_2$	1	$D_{X_1,X_2,X_1\times X_2}$	$D_{X_1,X_2} - D_{X_1,X_2,X_1\times X_2}$

Assim, para testar as hipóteses $H_0 : \beta_j = 0$ contra $H_1 : \beta_j \neq 0$ usa-se o resultado de que, sob H_0, a diferença entre as duas funções desvio (Tabela 2) tem distribuição qui-quadrado com 1 grau de liberdade (χ^2_1). Logo, se o valor observado da diferença for maior do que o quantil $\chi^2_{1,1-\alpha}$, ao nível de significância α,

a hipótese nula é rejeitada. O teste pode ser aplicado para um grupo de parâmetros, simultaneamente, tal que o número de graus de liberdade é dado pela diferença no número de parâmetros dos dois modelos.

Para a construção de intervalos de confiança (IC's), seja o estimador da variância do estimador $\boldsymbol{\beta}$ dada por $\widehat{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}$. Para cada $\hat{\beta}_j$ ($j = 0, 1, \ldots, k$) em particular, seu erro-padrão é a raiz quadrada do $(j+1)$-ésimo elemento da diagonal principal de $\widehat{Var}(\hat{\boldsymbol{\beta}})$, denotado como $\widehat{EP}(\hat{\beta}_j)$. Assim, o IC para $\hat{\beta}_j$, baseado na estatística de Wald (Hosmer Jr et al., 2013), é expresso como

$$\hat{\beta}_j \pm z_{1-\alpha/2} \times \widehat{EP}(\hat{\beta}_j), \tag{7}$$

em que $z_{1-\alpha/2}$ é o quantil de ordem $1 - \alpha/2$ da distribuição Normal padrão. O IC para $\ln\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right] = \boldsymbol{\beta}'\mathbf{x}$ é obtido através de

$$\hat{\boldsymbol{\beta}}'\mathbf{x} \pm z_{1-\alpha/2} \times \widehat{EP}(\hat{\boldsymbol{\beta}}'\mathbf{x}),$$

em que $\widehat{EP}(\hat{\boldsymbol{\beta}}'\mathbf{x}) = [\mathbf{x}'(\mathbf{X}'\hat{\mathbf{V}}\mathbf{X})^{-1}\mathbf{x}]^{1/2}$ com $\mathbf{x}' = (1,\ x_1,\ \ldots,\ x_p)$ sendo o vetor de valores das regressoras fixados. Por conseguinte, os IC's aproximados para $\pi(\mathbf{x})$ são encontrados via transformação inversa.

Uma maneira de interpretar os parâmetros através do ajuste é com a utilização da razão de chances ajustadas. Esta medida de associação em análise de tabelas de contingência é chamada de razão de chances (*odds ratio*). Para cada regressora x_j pode-se calcular a razão de chances que mede o efeito do aumento de uma unidade sobre a chance de sucesso, dado que os valores das outras regressoras estão fixadas. Logo, o estimador de razão de chances ajustada ($\widehat{OR}$) é dado por

$$\widehat{OR} = exp(\hat{\boldsymbol{\beta}}).$$

Seja $\mathbf{x}_{\text{-}j}$ o vetor de valores das regressoras exceto a regressora x_j e $\boldsymbol{\beta}_{\text{-}j}$ o vetor $\boldsymbol{\beta}$ excluindo-se β_j. Da equação (5) tem-se que

$$\frac{\pi(x_j|\mathbf{x}_{\text{-}j})}{1 - \pi(x_j|\mathbf{x}_{\text{-}j})} = exp(\boldsymbol{\beta}'_{\text{-}j}\mathbf{x}_{\text{-}j} + \beta_j x_j)$$

e

$$\frac{\pi(x_j+1|\mathbf{x}_{-j})}{1-\pi(x_j+1|\mathbf{x}_{-j})} = exp(\boldsymbol{\beta}'_{-j}\mathbf{x}_{-j} + \beta_j + \beta_j x_j)$$

tal que

$$OR(x_j+1|\mathbf{x}_{-j}) = \frac{\frac{\pi(x_j+1|\mathbf{x}_{-j})}{1-\pi(x_j+1|\mathbf{x}_{-j})}}{\frac{\pi(x_j|\mathbf{x}_{-j})}{1-\pi(x_j|\mathbf{x}_{-j})}} = exp(\beta_j)$$

cujo estimador é

$$\widehat{OR}(x_j+1|\mathbf{x}_{-j}) = exp(\hat{\beta}_j).$$

Dado o IC para β_j (equação 7), e usando a transformação inversa é possível obter intervalos de confiança aproximados para OR. Como OR é uma razão de chances, se o valor de 1 estiver incluído no intervalo significa que x_j não tem efeito significante. Interpretações similares podem ser feitas para regressoras que são qualitativas. Para mais detalhes sobre estas e outras interpretações no ajuste de regressão Binomial e análises de tabelas de contingência veja Giolo (2017) e Hosmer Jr et al. (2013). Na Seção 2.3 apresentam-se os principais diagnósticos para o modelo de regressão logística e o linear.

2.3 Diagnósticos

A busca pela boa adequação do modelo gera diversas teorias e consequentemente ferramentas, muitas vezes gráficas, para avaliar e compreender qual o peso de cada observação ou um conjunto delas na inferência dos parâmetros. Nesse sentido, a identificação de pontos que se destacam, dado um componente avaliativo, é de grande importância para determinar seu impacto nos aspectos gerais do modelo. Na teoria dos MLG, cada função de probabilidade considerada em estudo, possui maneiras diferentes de mensurar o impacto das observações feitas e também avaliar a qualidade do ajuste, consequentemente, existem diferentes formas de construir medidas de avaliação dessas observações. Visto que as aplicações feitas neste trabalho foram construídas com base nos modelos de regressão linear e Binomial, apresenta-se os principais diagnósticos para os mesmos.

Ao considerar um modelo linear Normal deve-se, inicialmente, verificar se os pressupostos feitos para $\boldsymbol{\varepsilon}$ da equação (1) (basicamente normalidade, homocedasticidade e não correlação) são satisfeitos, com isso considera-se sua estimativa $\hat{\boldsymbol{\varepsilon}}$, denominada por resíduo bruto. Nesse sentido, definem-se dois outros tipos de resíduos com base nos resíduos brutos (Montgomery et al., 2012), o padronizado dado por

$$r_i = \frac{\hat{\varepsilon}_i}{S\sqrt{1-h_{ii}}},$$

com $S = \sqrt{\hat{\sigma}^2}$ e h_{ii} sendo o i-ésimo valor da diagonal da matriz $\mathbf{H}$ (equação 3) e o resíduo studentizado, definido como

$$\hat{\varepsilon}^*_{(-i)} = \frac{\hat{\varepsilon}_i}{S_{(-i)}\sqrt{1-h_{ii}}},$$

com $S_{(-i)}$ sendo S resultante do modelo ajustado sem a observação i. Uma vez que as pré-suposições do modelo estão de acordo, este resíduo segue uma distribuição t de Student com $n-p-1$ graus de liberdade.

Definidos os resíduos, pode-se verificar o pressuposto da normalidade com um gráfico dos resíduos padronizados contra os quantis teóricos de uma Normal padrão ou os resíduos studentizados contra os quantis teóricos da distribuição t. Sua construção é feita ao plotar os resíduos em ordem crescente contra os quantis teóricos de uma Normal ou de uma distribuição t de Student. Para melhor comparação entre os elementos que compõem o gráfico, constrói-se, a partir de simulações, uma banda de confiança denominada envelope (Atkinson, 1981). Nesse gráfico, deve-se esperar que os pontos estejam dentro dessa faixa criada, pois a presença de uma certa quantidade de pontos fora dessa banda pode indicar inadequação do modelo ou presença de observações estranhas/atípicas ou ainda de possível influência no ajuste.

Para verificação do pressuposto de variância constante (homocedasticidade), deve-se plotar os resíduos contra os valores ajustados ($\hat{y}_i$'s). Nele, deve-se buscar por pontos espalhados pelo gráfico sem observar tendências e com faixa de variação aproximadamente constante. Uma forma de amenizar tendências crescentes de

variabilidade é utilizando alguma transformação da variável resposta. Uma família de transformações bastante conhecida é a família Box-Cox (Box & Cox, 1964), mais detalhes sobre ela e outras transformações podem ser estudadas em Montgomery et al. (2012) e Faraway (2016). Falta de ajuste ou necessidade de inclusão de outros termos no modelo podem ser investigados plotando-se os resíduos em função de cada regressora. O padrão esperado é o mesmo descrito acima.

Três são os componentes de pontos do banco de dados que podem comprometer o modelo. São eles, os pontos atípicos/extremos (*outlier*), ponto de alavanca (Hoaglin & Welsch, 1978) e pontos de influência. Um ponto *outlier* é aquele que apresenta alto resíduo, provocado usualmente por valores extremos na resposta. Para a identificação desses pontos pode-se utilizar o gráfico envelope e identificar os valores fora da faixa simulada. Outra maneira é plotar os valores dos resíduos padronizados ou studentizados contra a ordem das observações, buscando por pontos fora do intervalo -2 e $+2$, visto que uma vez padronizados, espera-se que 95% deles estejam dentro da faixa referida, segundo as propriedades da densidade Normal padrão.

Um ponto de alavanca é aquele que se destaca em relação ao centro do espaço das regressoras. Para identificá-los plotam-se os pontos de h_{ii}, que são os valores da diagonal principal da matriz $\mathbf{H}$, contra a ordem das observações, buscando por valores maiores do que $2p/n$, conforme recomendado por Paula (2013) e Faraway (2016). Em geral, deve-se avaliar o motivo pelo qual os pontos se destacam e têm seu comportamento diferente dos demais, sendo assim, busca-se por alternativas que possam tratá-los ou até mesmo reavaliar o modelo construído.

Além de observações atípicas, certos conjuntos de dados não raramente podem apresentar pontos indesejáveis de alta influência no ajuste. Um ponto é dito influente quando o ajuste ou estimativa de um ou mais parâmetros da regressão sofre alteração destacada quando o ponto é removido da análise. Assim, a presença de pontos influentes podem levar a estimativas viciadas e pode, inclusive, comprometer a qualidade do ajuste e consequentemente afetar o processo de seleção de variáveis.

É importante esclarecer que um ponto considerado *outlier* não necessariamente tem influência no modelo. Nesse sentido, utiliza-se de ferramentas para a identificação de possíveis pontos influentes. A maneira mais simples é utilizando as medidas de h_{ii} (pontos de alavanca), já que as propriedades da matriz $\mathbf{H}$ estabelece que no modelo "ideal"os valores de h_{ii} deveriam ser todos iguais a p/n (Montgomery et al., 2012).

Entre as medidas mais conhecidas para identificar pontos influentes são os DFBETAS e DFFITS. O DFBETAS é uma medida que indica quanto cada coeficiente da regressão muda se a i-ésima observação for removida dos dados e o modelo reajustado, enquanto os DFFITS mede a influência dos valores ajustados. Detalhes sobre essas estatísticas podem ser consultadas em Montgomery et al. (2012) e Faraway (2016). No sentido de resumir a informação contida nos DFBETAS, pois a mesma reflete qual a influência do ponto em cada coeficiente, utiliza-se a distância de Cook (Cook, 1977), definida da forma

$$\frac{r_i^2}{p} \times \frac{h_{ii}}{1 - h_{ii}}.$$

com $i = 1, \ldots, n$. Em geral, valores que se distanciam dos demais devem ser analisados, mas como referência busca-se pelos pontos próximos de 1, logo, um gráfico dessas medidas contra cada observação pode auxiliar nesse procedimento.

Para um modelo de regressão logística, as medidas de diagnóstico propostas para o modelo de regressão linear Normal sofrem adaptações, e diversos tipos de resíduos são considerados. Tomando dados agregados e agora y_i sendo o número de sucessos no agregado i, o resíduo componente do desvio (Paula, 2013) é dado por:

$$d_i = sinal(y_i - \hat{y}_i) \left\{ 2y_i \ \ln\left(\frac{y_i}{\hat{y}_i}\right) + 2(n_i - y_i) \ \ln\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right) \right\}^{1/2},$$

em que $\hat{y}_i$ é a estimativa do número de sucessos e n_i o número de realizações repetidas no grupo i. O resíduo de Pearson é dado por

$$\delta_i = \frac{y_i \ - \ n_i \ \hat{\pi}_i}{\sqrt{n_i \ \hat{\pi}_i \ (1 \ - \ \hat{\pi}_i)}},$$

com $i = 1, .., N$ e $\hat{\pi}_i = \hat{\pi}_i(\mathbf{x})$ é a estimativa da proporção esperada em cada lote i.

Para padronizá-los, considera-se a matriz

$$\mathbf{H} = \mathbf{V}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{1/2} \tag{8}$$

sendo $\mathbf{V}$ $(N \times N)$ a matriz de pesos utilizados no ajuste do modelo (conforme equação 6). Sendo assim, ao dividir d_i e δ_i por $\sqrt{(1-h_{ii})}$, com h_{ii} sendo os elementos da diagonal principal da matriz $\mathbf{H}$, obtém-se os resíduos componente do desvio padronizados (r_{d_i}) e os resíduos de Pearson padronizados (r_{δ_i}), respectivamente. Outro resíduo que auxilia na identificação de pontos atípicos é chamado de resíduo da verossimilhança (Collett, 2002) e definido da forma

$$r_{L_i} = sinal(y_i - \hat{y}_i)\ \sqrt{h_{ii}\ r_{\delta_i}^2 + (1-h_{ii})\ r_{d_i}^2}.$$

Note que r_{L_i} é uma combinação de $r_{\delta_i}^2$ e $r_{d_i}^2$. Se h_{ii} for pequeno então r_{L_i} é similar a $r_{d_i}^2$.

Para a identificação de *outliers* no modelo de regressão Binomial, dado que os resíduos componente do desvio se distribuem normalmente (Giolo, 2017; Collett, 2002), um gráfico envelope pode ser construído. Logo, busca-se por pontos fora da faixa simulada. Outra maneira é plotar r_{d_i} e/ou r_{L_i} contra a ordem das observações que deve conter, em sua maioria, pontos no intervalo de -2 e $+2$, pelos mesmos motivos citados no modelo linear Normal.

Na busca por pontos influentes, uma das formas de medir, aproximadamente, a influência da i-ésima observação no ajuste geral do modelo, é avaliar o valor de $r_{L_i}^2$, eles medem a diminuição no valor do desvio do modelo quando a observação i é removida dos dados e, neste caso, é importante avaliar valores que se destacam dos demais. Outra medida que auxilia na busca desses pontos, é pelos elementos da diagonal principal da matriz $\mathbf{H}$ (equação (8)), os h_{ii}'s. São chamados de pontos de alavanca os valores que se destacarem no gráfico de h_{ii} contra os índices das observações. Neste caso, procura-se por pontos que sejam maiores do que $2p/N$ (Collett, 2002).

Da mesma forma vista em um modelo linear Normal, a distância de Cook também é uma ferramenta útil na identificação de pontos influentes para o

modelo de regressão Binomial (Collett, 2002), sendo definida da forma

$$\psi_i = \frac{r_{\delta_i}^2 h_{ii}}{p(1-h_{ii})}, \quad i = 1, \ldots, N.$$

Novamente, busca-se por pontos que se destacam em um gráfico de ψ_i contra a ordem das observações. Para efeitos de comparação os mesmos devem ter valores pequenos, enquanto valores próximos de 1 devem ser avaliados. De forma geral, é importante tentar descobrir os motivos que levam pontos dos dados possuírem valores fora do padrão, embora na prática isso nem sempre é viável, para que se possa tratá-los. Mudanças na distribuição da variável resposta, inclusão de termos das regressoras, entre outras possibilidades, podem ser utilizadas com o intuito de diminuir os problemas diagnosticados no ajuste. Outras maneiras de medir a influência, bem como análise da adequação e procedimentos para possíveis transformações nas variáveis, podem ser estudados com detalhes em Collett (2002) e Paula (2013).

2.4 Seleção de Variáveis

2.4.1 Métodos e critérios usuais

Os princípios da modelagem estatística preconizam o ajuste de um modelo parcimonioso aos dados observados (Montgomery et al., 2012; Draper & Smith, 2014), ou seja, um modelo com número reduzido de parâmetros de forma que apenas as variáveis preditoras de importância sejam incluídas, porém mantendo-se a utilidade prática do modelo. Com isso, a interpretação e o uso posterior do modelo parcimonioso torna-se relevante pela sua simplicidade, exigindo menos tempo e gasto financeiro com possíveis predições, aspectos muitas vezes levados em consideração por pesquisadores e proprietários dos dados. A busca por um modelo que se adéque às exigências mencionadas pode ser realizado via teoria de seleção de variáveis. A bibliografia especializada apresenta várias técnicas e critérios clássicos que podem ser utilizados para selecionar variáveis na modelagem estatística, seja no modelo linear Normal, generalizado ou outros envolvendo diversas regressoras.

A dificuldade encontrada no processo de seleção deve-se à existência, na grande maioria, de colinearidade entre as regressoras causando instabilidade no processo de seleção. Esse problema é mais frequente nos estudos observacionais já que nos experimentos a ortogonalidade entre as regressoras pode ser obtida via planejamento, o que simplifica a seleção dos efeitos significantes devido à independência entre as estimativas dos diversos efeitos. Os métodos revisados neste trabalho são aplicáveis apenas aos casos nos quais o número de observações (n) é maior do que o número de parâmetros possíveis no modelo (p). No entanto, o desenvolvimento tecnológico da atualidade permite mensuração de um número muito grande de variáveis numa mesma unidade ou indivíduo, gerando o caso de $n << p$, como é o caso de estudos epidemiológicos genéticos (Wasserman & Roeder, 2009). Nesse contexto, o problema de seleção de variáveis é muito complexo, exigindo alternativas de estimação que incorporam penalidades às estimativas dos parâmetros, forçando-as a se aproximarem do valor nulo (Tibshirani, 1996).

A busca por um modelo parcimonioso leva à seguinte problemática, poucas regressoras no ajuste pode conduzir a estimativas e predições tendenciosas e um grande número delas conduz à instabilidade e consequente aumento da imprecisão na estimação e na predição de novas observações, além de poder resultar em modelos superajustados. Um modelo superajustado representa os dados quase que fielmente, explicando até parte da variabilidade aleatória intrínseca, o que não é desejável desde que dados amostrais ou experimentais, pela sua natureza, apresentam imperfeições. Assim, modelos superajustados são perigosos para novas previsões.

A colinearidade também atrapalha o processo de seleção de variáveis levando à instabilidade na escolha do modelo final. Além do mais, existe uma variedade de métodos e critérios para a seleção. É sabido, por exemplo, que variações dos métodos do tipo *stepwise* nem sempre convergirão para o mesmo modelo, e, mesmo fixando-se o método, pequenas perturbações nos dados também podem conduzir a modelos distintos. Outro aspecto é a existência de diferentes critérios de ordenação da qualidade do ajuste que podem levar a certa insegurança e dificuldade

para escolher o melhor modelo num grande conjunto de possibilidades.

Para introdução dos critérios de seleção é importante estabelecer a notação que será utilizada. Assume-se n observações independentes da variável $\mathbf{Y}$ e uma matriz $\mathbf{X}$ $n \times p$ $(p < n)$ de posto completo p cujas colunas estão indexadas pelos elementos do conjunto $\alpha^* = \{1, 2, .., p\}$. Seja α um subconjunto qualquer de α^* tal que contrói-se a matriz $\mathbf{X}_\alpha$ de dimensão $n \times p_\alpha$ e de posto p_α com $p_\alpha \leq p$. A matriz $\mathbf{X}_\alpha$ é a parte preditora do modelo α e $\boldsymbol{\beta}_\alpha$ são os parâmetros de regressão. O problema de seleção é o de descobrir α que resulta no modelo parcimonioso que explica $\mathbf{Y}$ em função de $\mathbf{X}_\alpha$. Os critérios de seleção recorrentes são o critério de informação de *Akaike* (AIC), proposto por Akaike (1973), e o critério de informação Bayesiano (BIC), proposto por Schwarz et al. (1978). Visto que AIC e BIC se baseiam na mesma equação para impor alguma penalidade à complexidade do modelo, Konishi & Kitagawa (1996) ampliou essa teoria introduzindo o critério de informação generalizado, dado pela equação

$$GIC(\alpha, \lambda) = \hat{Q}(\alpha) + \lambda \; p_\alpha \tag{9}$$

sendo $\hat{Q}(\alpha)$ o componente que mede a falta de ajuste do modelo α, por exemplo, soma de quadrados dos resíduos ou $-2 \; \ln(\boldsymbol{\beta}_\alpha, \phi; \mathbf{y})$, em que ϕ é o parâmetro de dispersão ($\phi = 1$ para binomial), cuja complexidade é medida por p_α (posto do modelo α) e λ é o multiplicador da penalidade. Quando $\hat{Q}(\alpha) = -2 \; \ln(\boldsymbol{\beta}_\alpha, \phi; \mathbf{y})$ e $\lambda = 2$ ou $\lambda = \ln(n)$, obtém-se o AIC e BIC, respectivamente. Na comparação de dois ou mais modelos, sob os critérios de informação, aquele com valor menor é preferível já que $\hat{Q}(\alpha)$, de certa forma, mede o distanciamento do modelo α ao "verdadeiro" modelo. Para definições de outros critérios veja Faraway (2016).

Para proceder à uma seleção utilizando as abordagens usuais é necessário decidir pelo critério e pelo procedimento de visitação dos diversos modelos. O procedimento de visitação dos modelos é referido como método de seleção. Uma classe de métodos para seleção de regressoras bastante usual é chamada de *stepwise*, que inclui três abordagens: *forward*, *backward* e *stepwise* (Montgomery et al., 2012). A *forward* adiciona regressoras ao modelo uma a uma, partindo-se de um modelo

nulo, a *backward* exclui regressoras uma a uma ao considerar, inicialmente, modelo completo, ou seja, aquele com número máximo de regressoras disponíveis, enquanto a *stepwise* combina os dois procedimentos, começando pelo modelo completo. No procedimento *forward*, uma vez que determinada variável é incluída, ela segue no modelo até o final do processo. Já no *backward*, uma vez que a variável é excluída do modelo, ela não volta a ser testada.

A inclusão ou exclusão das variáveis é feita seguindo algum critério, que informa se a regressora é importante no modelo que está sendo construído. O critério pode ser um teste de significância do efeito da regressora, por exemplo t, F, Wald, ou R^2 (Montgomery et al., 2012; Charnet et al., 2015), ou critérios de informação como AIC e BIC (Faraway, 2016), entre outros. No caso do critério ser um teste, recomenda-se que ao nível de significância estipulado não seja muito pequeno para dar maior chance de investigação de efeito de regressoras na presença de outras (veja Box & Draper (1987)). Caso o nível de significância for muito exigente (pequeno) a entrada de variíveis (*forward*) pode ser rara, assim como a saída (*backward*) muito frequente. Combinando, por exemplo, o método *forward* e o critério AIC, é feito a comparação dos valores AIC para cada modelo possível a cada passo. Dessa maneira, ao iniciar, a variável selecionada será aquela cujo ajuste produzir o menor AIC entre todos, inclusive do modelo nulo. O procedimento se repete até que o menor valor de AIC seja maior do que o valor do modelo construído até aquele passo.

Além dos métodos tipo *stepwise* para seleção de modelos pode-se também proceder uma busca exaustiva na qual todos os subconjuntos de regressoras possíveis de um modelo completo são visitados e comparados conforme um critério pré estipulado, por exemplo, o valor de AIC. Considerando que β_0 esteja contido em todos os modelos, para um total de $p-1$ regressoras, existem 2^{p-1} modelos possíveis para estimar e examinar. Portanto, o método é viável nos estudos com poucas variáveis independentes mas torna-se custoso conforme p aumenta, embora, com a evolução dos métodos e equipamentos computacionais tal custo tende a descrescer.

2.4.2 O método *fence*

Jiang et al. (2008) introduziram o método *fence*, ou cerca, para auxiliar na seleção de modelos dentro da classe de modelos mistos. Um modelo é dito ser de efeitos mistos quando, além da parte fixa que compõe o modelo de regressão usual, ele inclui também outros efeitos, além dos erros, que são supostos variarem aleatoriamente. Tal classe de modelos encontra aplicações nos problemas com dados de medidas repetidas ou que exibem algum tipo de agrupamento entre as observações, devido à forma de amostragem. Referências clássicas sobre esses modelos são, por exemplo, Pinheiro & Bates (2000) e Littell et al. (2006). Motivados pela dificuldade da definição ou interpretação de critérios baseados na equação (9) no contexto de modelos mistos, juntamente com a evolução computacional, Jiang et al. (2008) introduziram o método *fence* ou cerca, posteriormente simplificado em Jiang et al. (2009), com variações em Jiang (2014).

O método *fence* aplica a ideia de cercar subgrupos de modelos com propriedades similiares. O processo é baseado na condição

$$\hat{Q}(\alpha) - \hat{Q}(\alpha^*) \leq c, \tag{10}$$

no qual $\hat{Q}(\alpha)$ é uma medida de falta de ajuste do modelo α em análise, $\hat{Q}(\alpha^*)$ é uma medida de falta de ajuste do modelo completo α^*, que contém o efeito de todas as regressoras e c é uma constante que determina um ponto de separação ou de corte. Vale ressaltar que os modelos candidatos são aqueles encaixados em α^*, ou seja, seus parâmetros formam um subconjunto dos parâmetros de α^*. Jiang et al. (2009) apresenta o método em detalhes e Jiang (2014) faz uma revisão indicando extensões e variações do método, inclusive para problemas com $p >> n$, nos quais o método *fence* foi aplicado previamente para reduzir a dimensionalidade previamente às técnicas de regularização.

2.4.3 Ferramentas gráficas

Recentemente Tarr et al. (2018) desenvolveram uma ferramenta computacional e gráfica para assessorar o pesquisador no processo de seleção de variáveis, influenciado fortemente nos trabalhos de Murray et al. (2013), Müller & Welsh (2010) e Jiang et al. (2009), disponibilizada no pacote `mplot` (Tarr et al., 2018) para o *software* R (R Core Team, 2017). Esse pacote permite a construção de gráficos que se baseiam na teoria do critério de informação generalizado, no método *fence* e estudos de estabilidade de modelos. Para estabilidade é utilizado o método de reamostragem de *bootstrap* ponderado exponencialmente que consiste na reponderação das observações usando pesos simulados de uma função de densidade exponencial com média 1 (Murray et al., 2013). Existem diversas variações do método *bootstrap*, inclusive com pesos, conforme revistos em Gotwalt et al. (2018). Com base nos resultados de simulações de Jin et al. (2001) conclui-se que o uso de pesos dados por realizações de uma variável aleatória com média e variância iguais a um oferece estimadores com boas propriedades. Mais detalhes do método são dados em Murray et al. (2013). Em Müller & Welsh (2010) define-se que um procedimento/método de seleção de modelo é considerado instável quando, para uma pequena variação na penalidade (λ) na equação (9), pode-se selecionar modelos de diferentes dimensões.

O `mplot` oferece ferramentas para o auxílio no processo de seleção de modelos no caso linear e linear generalizado. Dois comandos principais calculam as medidas para a investigação da contribuição das variáveis e estabilidade. Uma vez criado os objetos, pelas informações obtidas nos dois comandos, é possível a construção de cinco tipos de gráficos: o gráfico para exploração do espaço dos modelos, o gráfico de estabilidade, o gráfico de inclusão de variável e os dois gráficos baseados no método *fence*.

O gráfico para exploração do espaço dos modelos é um diagrama do componente de perda, $-2\ln(\boldsymbol{\beta}_\alpha, \phi; \mathbf{y})$, em função da dimensão do modelo. Para cada dimensão, todos os modelos possíveis, incluindo o intercepto, são ajustados. No gráfico é permitido o uso de uma legenda para destacar os modelos que in-

cluem/excluem determinada covariável, proporcionando uma visualização informativa da contribuição da covariável ou ajuste, ou seja, na diminuição de $-2\ln(\boldsymbol{\beta}_\alpha, \phi; \mathbf{y})$. Uma outra versão desse gráfico permite o estudo de estabilidade de modelos realizado por reamostragens *bootstrap*. Ele concede a proporção de vezes que o modelo, para dada dimensão, apresentou o valor mínimo do componente de perda entre as repetições realizadas. Cada ponto no gráfico, que plota $-2\ln(\boldsymbol{\beta}_\alpha, \phi; \mathbf{y})$ versus a dimensão do modelo, é representado por um círculo com diâmetro proporcional à frequência de vezes que o modelo foi selecionado como sendo o melhor nas reamostras da dimensão considerada.

O gráfico de inclusão de variáveis (VIP) também é construído com base nas informações do processo de reamostragem. Para isso, toma-se B reamostras considerando um valor de λ, que varia em pequenos intervalos, e calcula-se a proporção de vezes que esta variável estava presente no modelo final, sendo esse o que tem menor valor no critério de informação generalizado (equação (9)) para a penalidade λ definida inicialmente. O gráfico VIP é um diagrama de linhas traçadas pelas probabilidades em função de λ. Espera-se que para valores pequenos, a maioria das regressoras estejam presentes nos ajustes e conforme há o aumento da penalidade, poucas pertençam a ele. Nesse processo é possível incluir uma variável redundante (RV), simulada a partir de uma distribuição Normal padrão. Seu objetivo é fornecer uma curva de base em VIP que serve de base de comparação para as outras variáveis. Curvas próximas ou abaixo da RV podem ter sido incluídas por acaso e indicam regressoras de baixa relevância no modelo. Ao considerar os dois primeiros gráficos mencionados, os modelos que contêm RV não são válidos.

Outras alternativas para o estudo de estabilidade de modelos são oferecidos pelo `mplot`, através da aplicação do método *fence* simplificado (Jiang et al., 2009), que foi proposto inicialmente para os modelos mistos, mas adaptado para GLM's por Tarr et al. (2018). Ilustrando o caso do modelo linear Normal, primeiramente escolha o conjunto de valores de c e ajuste o modelo completo α^*, obtendo $\hat{Q}(\alpha^*)$. Para cada valor de c segue-se os passos:

1. Reamostre (B vezes) parametricamente $\epsilon \sim N(0, \hat{\sigma}^2_{\alpha^*})$, em que $\hat{\sigma}^2_{\alpha^*}$ é o quadrado médio dos resíduos uma vez ajustado o modelo α^* .

   Para cada reamostra, identifique o modelo de menor dimensão $\hat{\alpha}(c)$ que satisfaz a condição na equação (10). Caso exista mais de um modelo do mesmo tamanho dentro da cerca, utilize o que tem menor $\hat{Q}(\hat{\alpha}(c))$.

   Calcule a probabilidade empírica de selecionar o modelo α, dada por $p^*(\alpha, c) = P^*(\hat{\alpha}(c) = \alpha) = \frac{\#(\hat{\alpha}(c)=\alpha)}{B}$;

2. Obtenha a probabilidade de seleção global $p^*(c)$ para dado c, que é dada por $p^*(c, \alpha)$ máximo. Nesse passo, selecione α que apresenta a maior probabilidade para dado c;

Após esse processo, faça o gráfico de $p^*(c)$ em relação a c e procure por picos e/ou regiões de estabilidade, ou seja, que mantêm o mesmo modelo numa faixa de valores de c.

Uma variação do procedimento é, no passo 1, ao invés de usar apenas o "melhor"modelo, considerar todos do mesmo tamanho e ponderar sua probabilidade por $1/k$, tal que k é o número de modelos de tamanho α dentro da cerca. Ao avaliar os dois possíveis gráficos construídos, a preferência é dada ao modelo que pertencer ao pico com menor c, ou seja, prefere-se o modelo cuja probabilidade de inclusão dentro da cerca é alta e cuja distância ao modelo completo é pequena. Nessa variação, serão encontradas probabilidades menores, entretanto pode-se destacar novos modelos que sofreram pouco com a aplicação da ponderação. Logo, ao construir esses gráficos, aplica-se o conceito de estabilidade ao buscar por intervalos na abscissa que contêm um único ajuste selecionado. Sendo assim, é de interesse verificar a existência de picos e intervalos de c que contêm um único ajuste destacado.

Na geração de todos os gráficos é necessário informar o número de reamostras (B) que devem ser feitas. Murray et al. (2013) optaram por B = 1000. Tarr et al. (2018) mostraram uma previsão do tempo gasto em cada função principal ao fixar B = 50 e variar o número de regressoras e recomendaram o uso de B = 150

e B = 200. Neste trabalho, optou-se pela variação de B entre 150 e 1000.

Para aumentar a praticidade das ferramentas propostas, uma das funções do `mplot` conduz a uma página no navegador de internet que possibilita a interação do usuário com todos os gráficos gerados pelas funções. Portanto, ao utilizar essas ferramentas, o pesquisador será capaz de identificar modelos e variáveis que possam contribuir significativamente na explicação da resposta. Para exemplos desse recurso, veja Tarr et al. (2018).

3 MATERIAL E MÉTODOS

Para a aplicação das ferramentas que assessoram o processo de seleção de variáveis disponíveis no `mplot`, utilizaram-se dois conjuntos de dados, o primeiro ilustra a seleção de variáveis no modelo linear Normal e o segundo no modelo de regressão logística. Em ambas as aplicações, a necessidade de inclusão de interações entre regressoras obedece o princípio da hierarquia entre efeitos, muito utilizado em análise de dados e planejamento de experimentos e bem explicado em Wu & Hamada (2009). Assim, primeiramente fez-se a busca pelos efeitos principais e posteriormente para as interações entre os mesmos. Decidiu-se pelo uso desse princípio uma vez que o `mplot` não distingue entre efeito principal e interação, ou seja, a função trata, por exemplo, o termo de interação entre duas regressoras como se fosse outra regressora, o que não faz sentido na modelagem. Outro motivo para o uso deste princípio é que, para dados observacionais com muitas regressoras, a inclusão de por exemplo todos efeitos principais e interações pode levar ao problema de colinearidade entre colunas da matriz $\mathbf{X}$ e até ao problema $n < p$. Nas seções 3.1 e 3.2 são apresentadas algumas informações sobre os dados para as duas aplicações. Todas as análises e cálculos foram realizados no programa R (R Core Team, 2017).

3.1 Aplicação 1: estudo sobre peso de recém-nascidos (RN) prematuros

O primeiro conjunto foi obtido de colaborações entre pesquisadores do Departamento de Bioestatística (IBB, Unesp) e Faculdade de Medicina de Botucatu (Unesp) e está relacionado ao estudo de Prigenzi et al. (2008), que investiga fato-

res de risco associados à mortalidade de recém-nascidos. Uma parcela dos dados está exposta no Anexo 3. Neste trabalho, estudou-se a relação do peso de recém-nascidos prematuros e outras características da mãe, do período gestacional e do recém-nascido. A amostra contém 90 observações e treze variáveis com todos os dados completos, sendo cinco quantitativas (idade da mãe, número de consultas, número de gestações, número de partos, idade gestacional e peso do recém-nascido) e sete qualitativas binárias (hábito de fumar da mãe - sim ou não; realização do pré-natal - sim ou não; mãe com hipertensão arterial sistêmica - sim ou não; utilização de corticoide pela mãe - sim ou não; gestação única - sim ou não; alguma outra doença durante a gestação - sim ou não; sexo do recém-nascido - masculino ou feminino). Nas Tabelas 3 e 4 são apresentadas algumas medidas descritivas de cada característica.

Tabela 3. Medidas descritivas das variáveis quantitativas, dados da Aplicação 1 ($n = 90$)

Variável	Código	Mín.	Mediana	Média	Máx.	D.P.
Peso dos recém-nascidos (g)	peso	635	1290	1318	2450	322
Idade da mãe (anos)	idade_mae	14	26	26,47	44	6,6
Número de consultas	n_consultas	0	4	4,4	14	2,5
Número de gestações	n_gest	1	2	2,36	7	1,4
Número de partos	n_partos	0	1	0,97	5	1,2
Idade gestacional (semanas)	IG	24	30	30	36	2,4

Tabela 4. Medidas descritivas para as variáveis binárias, dados da Aplicação 1 ($n = 90$)

Variável	Código	Categoria	Número	%
Hábito de fumar	fumo	Sim	21	23,33
Realização do pré-natal	pre_natal	Sim	81	90,00
Hipertensão	HAS	Sim	13	14,44
Uso de corticoide	corticoide	Sim	60	66,66
Gestação única	GU	Sim	75	83,33
Doença na gestação	doenca_na_ges	Sim	64	71,11
Sexo do recém-nascido	GR	Feminino	48	53,33

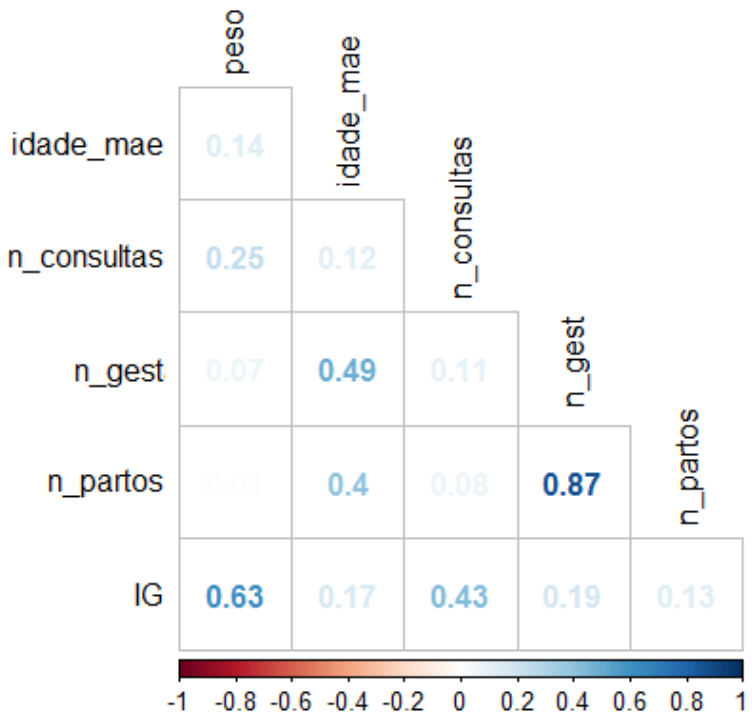

Figura 1 - Valores da correlação entre as regressoras quantitativas e a resposta peso, dados da Aplicação 1 ($n = 90$)

Nota-se pela Figura 1 que, analisando as correlações lineares (Pearson) entre pares de variáveis, a regressora mais correlacionada com o peso de RN é `IG` (idade gestacional), enquanto que as demais regressoras apresentam baixíssimo coeficiente de correlação com a resposta. Há alta colinearidade entre as covariáveis `n_gest` e `n_partos`, o que é esperado, pois para ocorrência de um parto é necessário que ocorra a gestação (número de partos é sempre menor ou igual ao número de gestações).

O objetivo nesse estudo é investigar as relações entre o peso e as demais variáveis, explorando alguns métodos de seleção de variáveis e recursos oferecidos no pacote `mplot`. Para isso, iniciou-se com um ajuste de um modelo de regressão múltipla com erros Normais aplicando a função `lm` e seguido de algumas análises de diagnóstico preliminares. Para verificar o pressuposto da normalidade, utilizou-se a função `qqPlot` do pacote `car` (Fox & Weisberg, 2011), que resultou na Figura 3 e 4. Dessas análises, concluiu-se a necessidade de transformar a variável resposta, o que foi feito seguindo a metodologia de Box-Cox (Box & Cox, 1964) e aplicando-se a função `box.Cox` do pacote `car`. Para explorar os métodos de seleção de variáveis aplicou-se os procedimentos usuais de *stepwise* utilizando os critérios de AIC e BIC pela função `step` do `R`.

Utilizando os recursos do `mplot`, foram identificados alguns modelos potenciais para o problema, sendo que em alguns deles abriu-se a possibilidade de investigação de termos de interação entre variáveis regressoras. Seguindo o princípio de hierarquia entre os efeitos, no `mplot`, os termos de interações foram avaliados aplicando-se o processo de seleção em um modelo que contém em sua resposta os resíduos de um ajuste feito com os efeitos principais selecionados e, na parte preditiva, as possíveis interações entre esses efeitos. Nota-se que esse procedimento é feito para todos os modelos potenciais.

Em suma, os passos tomados até a obtenção dos modelos potenciais foram:

1) ajuste um modelo linear Normal com todas as regressoras;

2) análise de diagnóstico indicou heterogeneidade de variâncias, o que foi resolvido com a transformação Box-Cox;

3) aplicação dos procedimentos de seleção de variáveis para a variável resposta peso transformada e efeitos principais das regressoras;

4) aplicação dos procedimentos de seleção de variáveis nos efeitos de interação das variáveis selecionadas no item 3.

Nos passos 3 e 4 tantos os procedimentos automáticos quanto o `mplot` foram aplicados. No Anexo 1 são apresentados os códigos utilizados para construção e geração das ferramentas gráficas.

3.2 Aplicação 2: probabilidade de prenhez em inseminação artificial (IA) de vacas

O segundo conjunto de dados, se refere a um estudo das relações entre a probabilidade de prenhez de vacas inseminadas artificialmente (IA) e as diversas características do sêmen utilizado, obtido a partir das colaborações do Departamento de Bioestatística (IBB, Unesp) e pesquisadores da área de veterinária. No estudo um total de 3657 vacas foram IA com sêmen de touros também distintos, e classificadas em prenhez (sucesso) ou não (fracasso). Além das características do sêmen que englobam treze covariáveis, com observações completas, o estudo apresenta outras variáveis ambientais e de manejo que devem ser levadas em consideração como as fazendas, os inseminadores, as datas da realização da IA e outros fatores. Essas informações foram condensadas em uma variável categórica com quatro grupos, incluída na modelagem, como uma variável de blocagem. Em geral, mais de uma vaca foi inseminada com sêmen da mesma amostra, havendo repetições de condições. Cada condição é definida como uma combinação específica dos valores das características do sêmen e do fator bloco, que será referida como lote. Assim, a modelagem pode proceder a partir de dados agregados ou binários. Na forma agregada, o conjunto

de dados ficou composto por 72 lotes cujas respostas são em termos de números de vacas no lote e número de sucessos. Os lotes são desbalanceados quanto ao número de vacas. Na forma binária, a resposta de cada vaca é 0 (fracasso) ou 1 (sucesso).

O tipo de variável resposta leva ao modelo de regressão Binomial capaz de modelar a probabilidade de prenhez em função das diversas covariáveis. Tanto a versão para dados agregados, que supõe variáveis binomiais independentes aos lotes, cada uma com parâmetro n específico, quanto a versão para dados binários, que supõe variáveis Bernoulli independentes, levam às mesmas estimativas de efeitos, uma vez que as covariáveis são as mesmas. As probabilidades de sucesso são vistas como funções dos efeitos das covariáveis, inclusive a de blocagem. Entretanto, para fins de diagnóstico do ajuste, a forma agregada se apresenta mais vantajosa (Collett, 2002). Neste trabalho, as duas formas foram utilizadas, cada qual escolhida para resolver os problemas de implementação computacional enfrentados nas diversas fases, conforme será esclarecido no decorrer da descrição dos métodos.

Para a construção do modelo a estratégia foi, inicialmente, a partir de uma análise preliminar, explorar a presença de observações influentes e/ou extremas e, num primeiro momento, remover aquelas observações com indicações de altamente espúrias ou influentes, antes da aplicação da metodologia de seleção de variáveis. Após a seleção do modelo (ou dos modelos que se mostraram mais promissores), procedeu-se às técnicas de diagnóstico do ajuste e possíveis interpretações dos resultados.

Assim, inicialmente foi realizada uma análise preliminar ajustando um modelo com todos os efeitos principais das covariáveis, na versão de dados agregados (72 lotes), com função de ligação logito, caracterizando então o modelo de regressão logística. Para o diagnóstico do ajuste utilizaram-se os recursos do pacote `hnp` (Moral et al., 2017). Essa inspeção mostrou 7 lotes extremamente influentes no ajuste e decidiu-se trabalhar com apenas N = 65 lotes, totalizando $nv = 2399$ vacas inseminadas para a aplicação da teoria de seleção de variáveis, visto que o ajuste com esses lotes não destaca pontos influentes e parece satisfazer às pressu-

posições do modelo logístico, conforme brevemente descritos no Cápitulo 2, Seção 2.3. Poderia ser também de interesse verificar o quanto a presença desses pontos, considerados atípicos e/ou influentes, afetam ou dificultam os processos de seleção, porém esse procedimento não foi realizado neste trabalho. Análises descritivas das características do sêmen dos 65 lotes estão apresentadas nas Tabelas 5 e 6.

No caso da regressão Binomial, devido à utilização do método *bootstrap* para o cálculo das várias medidas pelo pacote `mplot` é necessário que a variável resposta seja limitada em $[0; 1]$, podendo ser a proporção de sucessos e declarando-se n_i (número total de vacas no lote i) como o peso da observação i ($i = 1, 2, \ldots, 65$) ou o uso dos dados na forma expandida. Esta é uma limitação do pacote que, felizmente, pode ser contornada devido à equivalência entre os ajustes.

Para tal, declara-se na função `glm` do `R` a opção `weight` $= \mathbf{w}$, na qual $\mathbf{w}$ é o vetor coluna cujos elementos são os n_i's. É possível também trabalhar com os dados expandidos, ou seja, com uma observação para cada vaca inseminada com resposta binária (0 ou 1) pois, como cada lote recebeu apenas um tipo de sêmem e, é sabido, o total de vacas inseminadas e prenhas, é possível replicar as linhas necessárias para as covariáveis. Outra obrigatoriedade do pacote utilizado é que a função de ligação utilizada, quando considerado um modelo Binomial, deve ser a logística, visto que existem outras possibilidades para a mesma (Collett, 2002).

Nesta aplicação, existe um fator de delineamento do estudo, o fator de blocagem que agrega os efeitos de fazenda, inseminadores, época e outros, controlando a heterogeneidade externa que não deve sofrer o processo de seleção. O `mplot` também não é capaz de diferenciar esse tipo de variável das demais sujeitas ao processo de seleção. Assim, foi necessário buscar uma alternativa para aplicar o procedimento `mplot`. Ao invés de utilizar a variável prenhez (binária) ou a proporção de sucessos, utilizaram-se os resíduos de um modelo logístico ajustado apenas para o fator "bloco". Para isso, foi utilizado a estratégia de linearização do modelo logístico e aplicação do método de mínimos quadrados ponderados. Esse método produz estimativas equivalentes às do modelo logístico usual (Hosmer et al., 1989). Os passos

são:

1. ajuste uma regressão logística, com todos os efeitos principais, para dados binários ($y_i = 0; 1$ com $i = 1, 2, \ldots, nv$), em que nv é o número de vacas inseminadas;

2. extraia as probabilidades estimadas, $\hat{\pi}_i$;

3. crie uma nova variável dependente, dada por

$$z_i = \ln\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) + \frac{y_i - \hat{\pi}_i}{\hat{\pi}_i\ (1-\hat{\pi}_i)},$$

   com a segunda parcela definida como os resíduos tipo *working*;

4. considere os pesos como

$$v_i = \hat{\pi}_i\ (1-\hat{\pi}_i);$$

5. ajuste um modelo linear Normal com ponderação no qual o vetor resposta é $\mathbf{z} = (z_1, z_2, \ldots, z_{nv})$, as regressoras são as indicadoras de blocos e os pesos dado pelo vetor $\mathbf{v} = (v_1, v_2, \ldots, v_{nv})$. Coloque os resíduos desse ajuste no vetor $\mathbf{r}$;

6. ajuste um modelo linear Normal similar ao passo 5 porém com cada regressora, digamos `x`, como resposta e coloque os resíduos desse ajuste no vetor `xr`. Repita para todas as regressoras que deseja explorar, nomeando o vetor de resíduos com o nome da regressora acrescido da letra `r`. A ideia nesse passo é remover de cada regressora aquele componente que é explicado pelo fator bloco. Note que, a ordem da regressora utilizada nesse passo não afetará o resultado final, pois tem-se um ajuste separado para cada regressora;

7. ajuste um modelo linear Normal ponderado de $\mathbf{r}$ em função dos resíduos obtidos no passo 6;

8. aplique os comandos do `mplot` para gerar os gráficos.

No caso da aplicação, as regressoras que sofrerão os ajustes no passo 6 são aquelas apresentadas na Tabela 5.

Após encontrar alguns modelos candidatos potenciais para explicar a probabilidade de sucesso na inseminação, explorou-se a inclusão de termos de interação de primeira ordem entre as regressoras pertinentes para cada um deles. Novamente, a estratégia é aplicando-se o mesmo procedimento utilizado na Aplicação 1, baseado no princípio da hierarquia entre efeitos.

Tabela 5. Medidas descritivas das variáveis quantitativas (características do sêmem), dados da Aplicação 2 ($N = 65$)

Variável	Código	Mín.	Mediana	Média	Máx.	D.P.
Membrana íntegra (%)	mint	4,66	50,96	47,61	74.98	16,23
Atividade mitocondrial (%)	mitp	19,81	47,39	46,66	72,61	12,05
Motilidade do sêmem (%)	mot	4,30	39,20	35,95	67,90	14,59
Defeitos totais (%)	deft	0,06	0,15	0,16	0,40	0,08
Defeitos maiores (%)	defm	0,02	0,08	0,10	0,28	0,07
Velocidade linear ($\mu m/s$)	vsl	54,93	75,11	77,30	105,72	10,57
Velocidade curvilínea (ou total)	vcl	101,50	147,20	152,20	228,50	26
Velocidade média da trajetória do deslocamento ($\mu m/s$)	vap	64,13	89,04	91,28	128,03	12,80
Amplitude do deslocamento lateral da cabeça (μm)	alh	4,33	6,52	6,84	9,74	1,31
Motilidade progressiva	progr	2,60	29,10	26,74	48,20	11,07
Linearidade =(vsl/vcl)×100	lin	37,83	51,42	51,35	64,41	5,56
Retilinearidade =(vsl/vap)×100	str	73,23	84,05	83,51	90,56	4,36

Tabela 6. Medidas descritivas da variável resposta por lote, dados da Aplicação 2 ($N = 65$)

Variável (por lote)	Mín.	Mediana	Média	Máx.	D.P.
Número de vacas (n_i)	10	20	36,91	138	32,08
Número de vacas prenhas (y_i)	2	11	17,26	65	15,32
Proporção de vacas prenhas ($\tilde{p}_i$)	0,11	0,47	0,49	0,92	0,14

Em suma, o roteiro de análise dos dados da Aplicação 2 foi:

1) ajuste de um modelo de regressão binomial com função de ligação logito (N=72);

2) análise diagnóstico indicou 7 lotes altamente influentes optando-se pela exclusão destes;

3) construção do modelo linearizado incluindo o efeito da regressora de blocos;

4) aplicação dos procedimentos de seleção de variáveis no modelo de efeitos principais (N=65);

5) aplicação dos procedimentos de seleção de variáveis nos efeitos de interação das variáveis selecionadas no item 3.

Nos passos 4 e 5 tantos os procedimentos automáticos quanto o `mplot` foram aplicados. No Anexo 2 são apresentados os códigos utilizados para contrução e geração das ferramentas gráficas.

Nota-se pela Figura 2 que, analisando as correlações lineares (Pearson) entre pares de variáveis, as únicas regressoras que apontam algum destaque com a proporção de prenhes são `lin` e `str`, com estimativas negativas, embora os coeficientes estimados sejam baixos em valor absoluto. Essas duas covariáveis mostram altíssima correlação entre si. Vários outros coeficientes são altos, tanto positivos quanto negativos, indicando problemas de colinearidade entre as regressoras, o que torna a seleção de variáveis mais complicada.

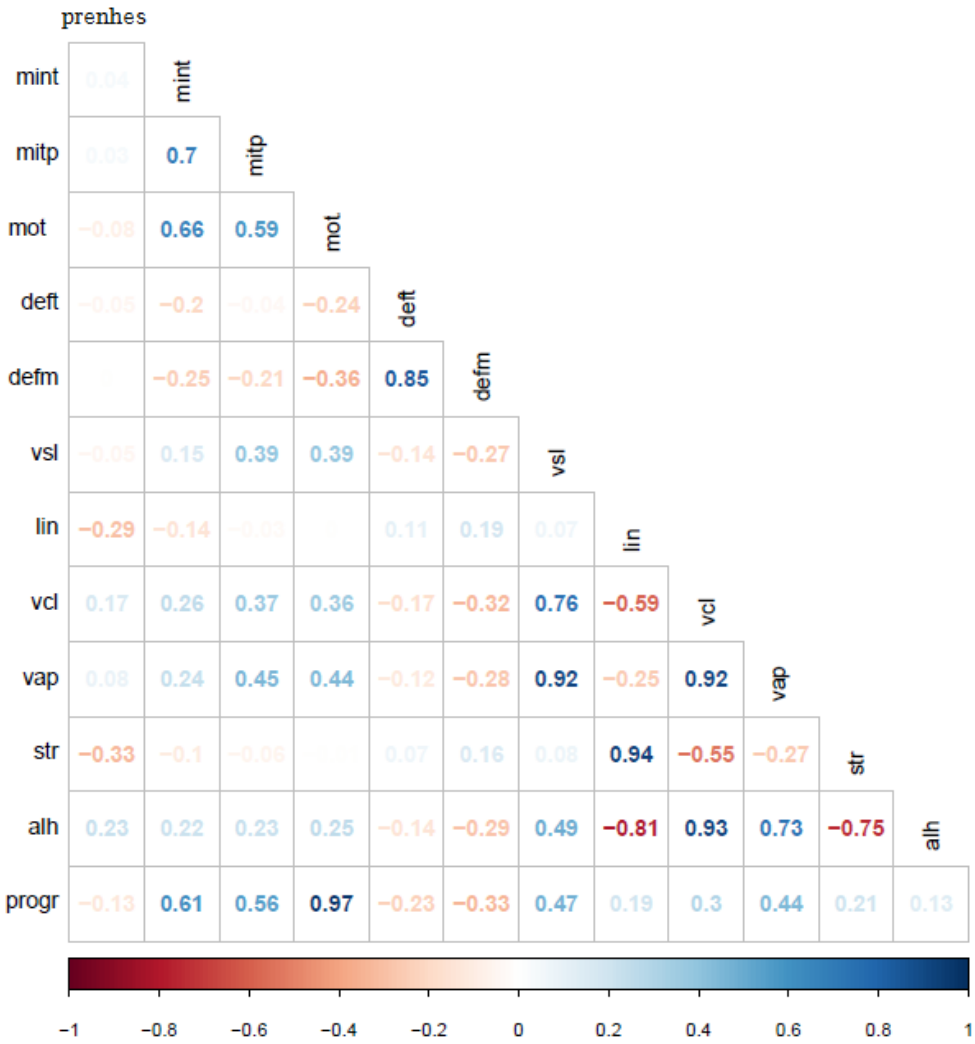

Figura 2 - Valores da correlação entre as regressoras quantitativas e da resposta por lote, dados da Aplicação 2 ($N = 65$)

4 RESULTADOS

4.1 Aplicação 1: estudo sobre o peso de recém-nascidos (RN) prematuros

Nessa aplicação pretende-se explorar as técnicas de seleção de variáveis para modelos explicativos do peso de recém-nascidos (RN) em função de variáveis relacionadas à gestação. Considerou-se, inicialmente, um modelo Gaussiano com apenas os efeitos principais das possíveis regressoras. Embora a análise de variância (Tabela 7) mostre a significância da regressão, ou seja, traga evidência de que ao menos uma regressora possui relação linear com a resposta ($F_{0,95;12;77}$ =1,9), nota-se graficamente, na Figura 3, à direita, que o pressuposto de variância constante não é satisfeito, pois verifica-se um aumento da dispersão dos resíduos conforme o peso estimado aumenta. O gráfico envelope (à esquerda) também confirma a existência de resíduos positivos ultrapassando a banda. Consequentemente, utilizou-se a transformação Box-Cox, que indicou a transformação logarítmica para o peso, ou seja, $y = \ln(peso)$. Propõe-se então um novo modelo, cujas informações sobre o ajuste são apresentadas na Tabela 8 e na Figura 4.

Tabela 7. ANOVA do modelo de regressão ajustado para o peso de RN

Fonte de variação	S.Q.	G.L.	Q.M.	F_0
Regressão	4562050	12	380171	6,3
Resíduo	4669544	77	60643	
Total	9231594	89		

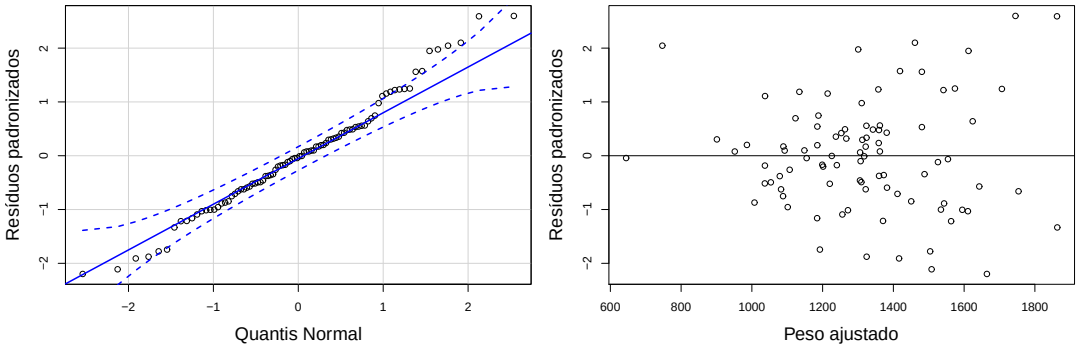

Figura 3 - Gráfico de resíduos do modelo de efeitos principais ajustado para a variável - peso de RN: envelope quantil-quantil Normal (à esquerda) e resíduos em função dos valores ajustados (à direita).

Tabela 8. ANOVA do modelo de regressão ajustado com $y = \ln(peso)$ de RN

Fonte de variação	S.Q.	G.L.	Q.M.	F_0
Regressão	0,0509	12	0,0042	6,8
Resíduo	0,0479	77	0,0006	
Total	0,0988	89		

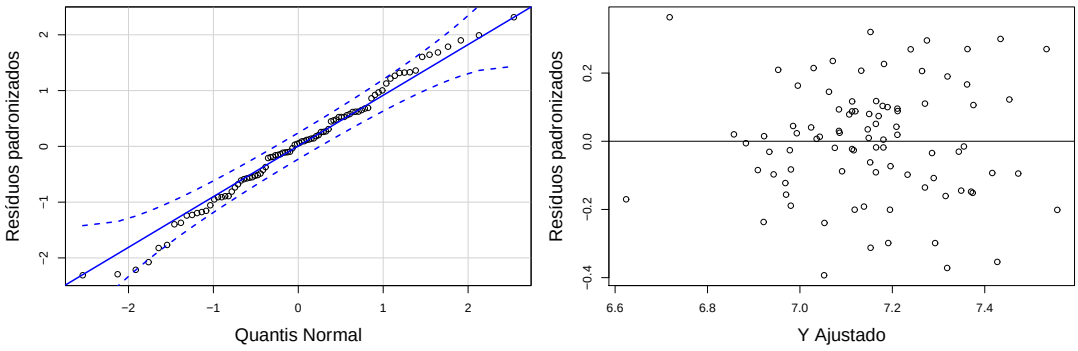

Figura 4 - Gráfico de resíduos do modelo de efeitos principais ajustado para a variável - $\ln(peso)$ de RN: envelope quantil-quantil Normal (à esquerda) e resíduos em função dos valores ajustados (à direita).

O gráfico na Figura 5 relaciona os resíduos com os valores indicadores de alavanca, os h_{ii}'s, elementos da diagonal da matriz **H**. A área dos círculos é proporcional aos valores das distâncias de Cook, permitindo a exploração de possíveis pontos influentes no ajuste. Dois pontos, 22 e 67, apresentam $h_{ii} > 2p/n \approx$0,29. Em particular, a observação 22, é de um bebê cuja mãe não realizou o pré-natal e, dentre essas mães, foi a que apresentou o maior número de gestações, 5, e apenas 3 partos. Já a observação 67 apresenta o maior número de consultas entre todas, 14. Entretanto, em termos da distância de Cook essas observações não se destacam. Verifica-se que três outros pontos, 29, 12 e 87, apresentam destaque em termos de distância de Cook, porém seus valores são 0,120; 0,123 e 0,135, respectivamente, considerados ainda baixos e provavelmente não muito influentes no ajuste. Os outros pontos não se destacaram o suficiente para que se faça um possível tratamento.

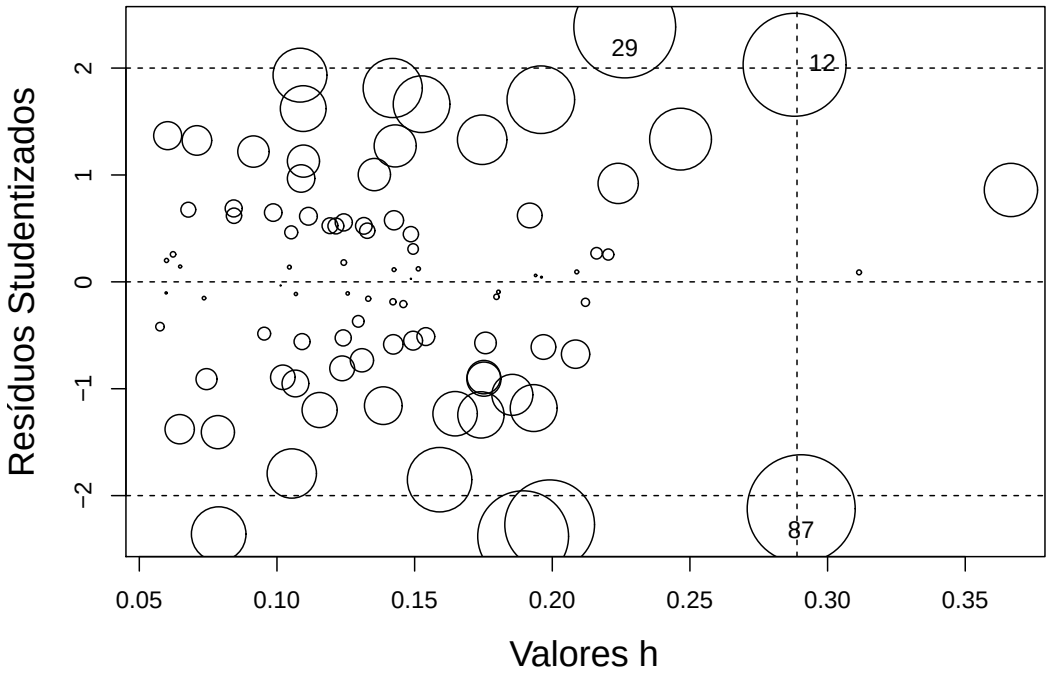

Figura 5 - Gráfico para diagnóstico de influência do ajuste do modelo de efeitos principais para ln($peso$) de RN. O raio dos círculos são proporcionais à distância de Cook.

Aplicando os métodos *stepwise*, *backward* e *forward* para seleção de variáveis, obteve-se os mesmos modelos quando o critério foi mantido. Os modelos selecionados são indicados na Tabela 9, mostrando que, ao considerar um valor de λ maior, pois $\ln\ 90 > 2$, o ajuste encontrado contém uma regressora a menos.

Tabela 9. Modelos selecionados pelos métodos usuais segundo o critério, para o $\ln(peso)$ de RN

Critério	Modelos
AIC ($\lambda = 2$)	$E(Y) = \beta_0 + \beta_1\ IG + \beta_2\ GR1 + \beta_3\ pre_natal1$
BIC ($\lambda = \ln\ 90$)	$E(Y) = \beta_0 + \beta_1\ IG + \beta_2\ GR1$

Considerando a metodologia do pacote `mplot`, verificam-se diferentes possibilidades de modelos ao analisar estabilidade e observar quais covariáveis se destacam. A Figura 6 apresenta o gráfico de inclusão de variáveis que mostra a regressora idade gestacional (`IG`) sendo incluída com probabilidade 1, independentemente da penalização (λ). A variável sexo do RN (`GR`) também se destaca com altas probabilidades de inclusão para uma grande faixa de valores de λ. Com probabilidades que decrescem com rapidez, porém bem maiores do que as da variável de referência (RV), aparece a regressora binária `pre_natal`. Duas outras variáveis que apresentam probabilidades altas para uma penalidade baixa e que são ligeiramente maiores que as de RV, são as binárias gestação única (`GU`) e doença hipertensiva (`HAS`), porém as mesmas não seriam selecionadas em favor da simplicidade do modelo.

É possível destacar no gráfico da Figura 6 qualquer regressora que o pesquisador desejar, neste caso optou-se pela RV que auxilia na escolha de variáveis, como mencionado. Outro destaque, é a indicação feita para as penalidades quando considerados os critéros AIC ($\lambda = 2$) e BIC ($\lambda = \ln 90$).

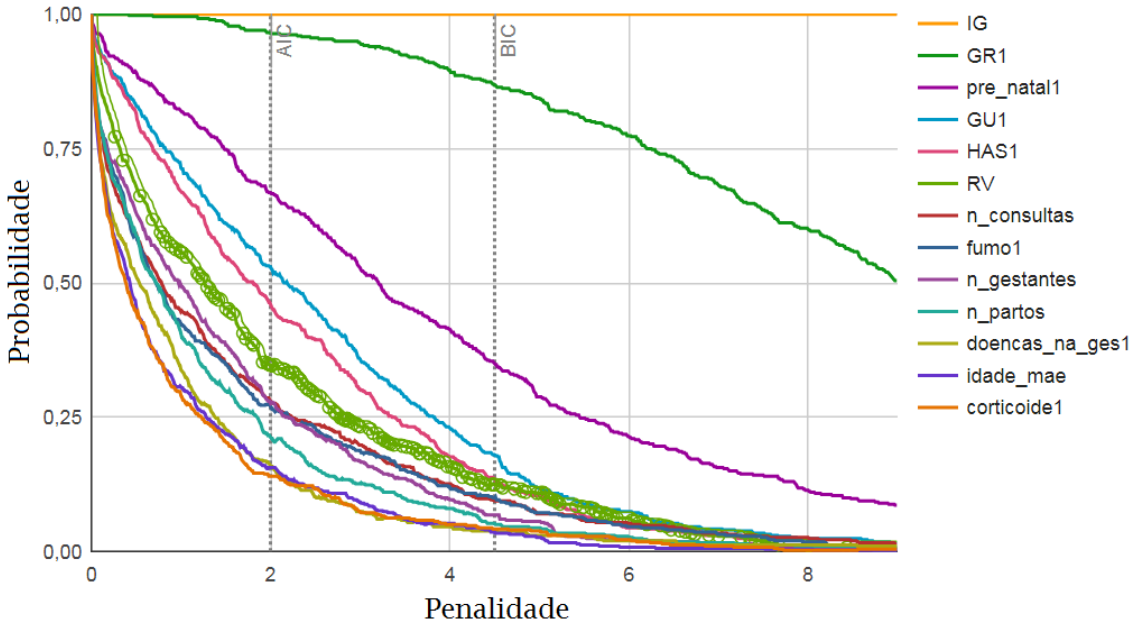

Figura 6 - Gráfico de inclusão de variáveis (VIP) em modelos de efeitos principais para ln($peso$) de RN, com destaque da curva de RV.

A Figura 7 apresenta gráficos de estabilidade dos modelos nos quais para cada tamanho de modelo (número de regressoras mais intercepto) uma medida de perda ou falta de ajuste é apresentada. Enquanto o ponto azul, para 1 parâmetro, apresenta o modelo apenas com intercepto, o ponto vermelho, com 14 parâmetros, mostra um ajuste composto pelos efeitos principais das 12 regressoras mais o intercepto e a variável redundante (RV). Destacam-se, nos gráficos, os modelos que contêm, ou não, determinada variável de interesse, neste caso analisou-se com três regressoras que possuem maiores probabilidades na Figura 6. Todavia, é possível destacar qualquer covariável de interesse. Nota-se no primeiro gráfico que a inclusão da regressora `IG` tem grande impacto nos valores do componente de perda, mesmo no modelo em que ela se apresenta sozinha. Verifica-se também que, dado `IG`, ao incluir a covariável `GR`, ocorre a diminuição substancial dos valores no segundo gráfico. Já para a regressora `pre_natal` não se observa clara separação entre os modelos, logo deve-se buscar pelo auxílio de outras ferramentas para investigar sua importância.

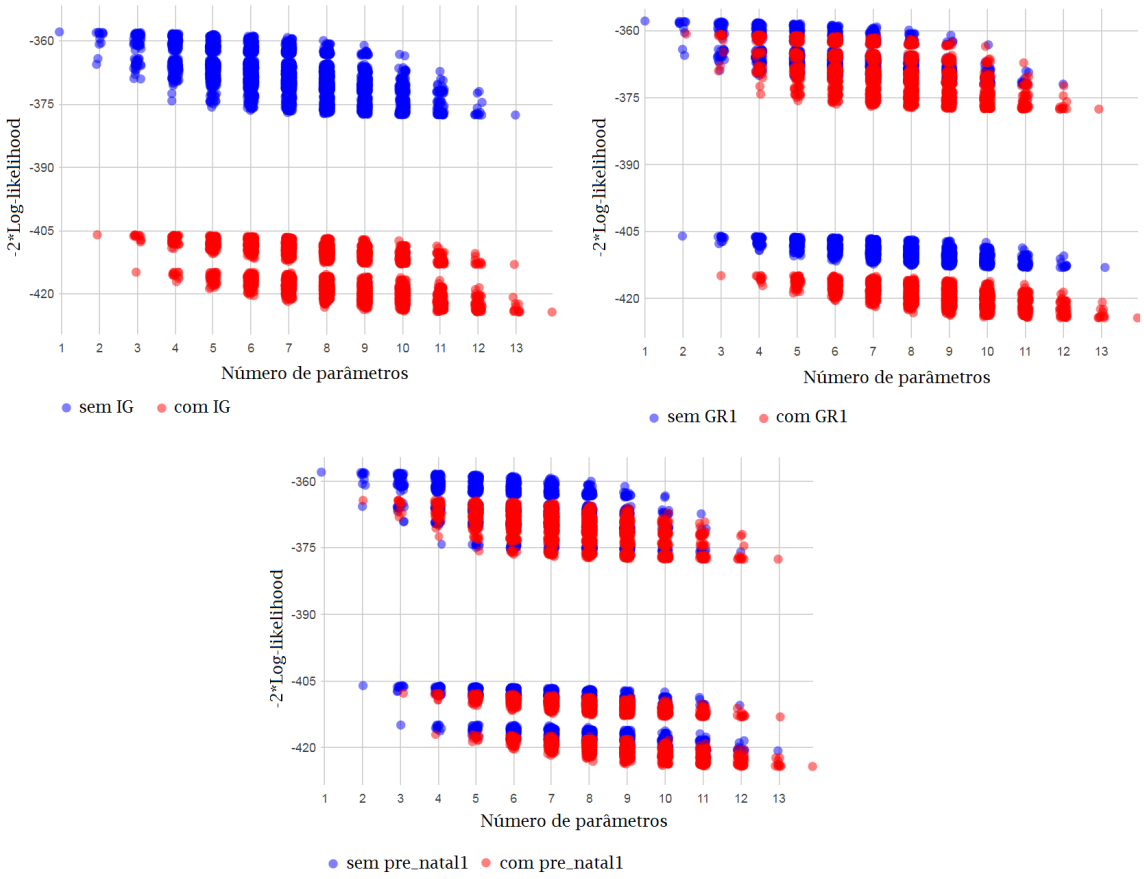

Figura 7 - Gráficos do valor do componente de perda contra a dimensão do modelo. Modelos de efeitos principais para ln(*peso*) de RN. Destaque para as regressoras `IG` (à esquerda), `GR` (à direita) e `pré_natal` (inferior).

A Figura 8 complementa a Figura 7 ao relacionar o valor do componente de perda ao número de parâmetros do modelo, com a informação adicional sobre a probabilidade de seleção dos modelos com determinada dimensão representado pelo tamanho do círculo. Na figura também pode-se destacar uma variável que, no caso mostrado, foi a variável `pré_natal`. Assim, o modelo nulo e o completo apresentam probabilidade 1 pois, por conterem o número mínimo e o máximo de parâmetros, respectivamente, eles sempre serão selecionados na reamostragem de seu tamanho. Note que modelos sem o intercepto não são permitidos. Modelos de destaque são aqueles com probabilidades altas e valores baixos da função de perda e número de parâmetros. As probabilidades dos três modelos em destaque, no canto inferior esquerdo do gráfico, estão apresentadas na Tabela 10. Nota-se que mesmo com pouco destaque, o modelo que contém pré-natal foi selecionado em 28% das vezes, dado a dimensão quatro, tornando sua inclusão uma nova opção para a modelagem. No tipo de gráfico desta figura, pode-se destacar qualquer regressora de interesse uma a uma, entretando, decidiu-se mostrar apenas a `pré_natal`, pois ela não apresentou de forma evidente sua importância nas outras figuras analisadas.

Por se tratar de uma ferramenta interativa, quando um círculo é selecionado, como ilustrado no gráfico da Figura 8, uma caixa com algumas informações sobre o modelo apontado é disponibilizada. Dentre as informações tem-se o modelo, o valor de $-2*$Log-likelihood (LL), a variável de destaque (var.ident) e a probabilidade de seleção feita pelo método de reamostragem (prob). A letra k indica o número de parâmetros do modelo, mas não apresenta com exatidão (devido ao efeito *jitter* usado no gráfico para possibilitar a identificação de modelos do mesmo tamanho), logo deve ser aproximado para um número inteiro mais próximo.

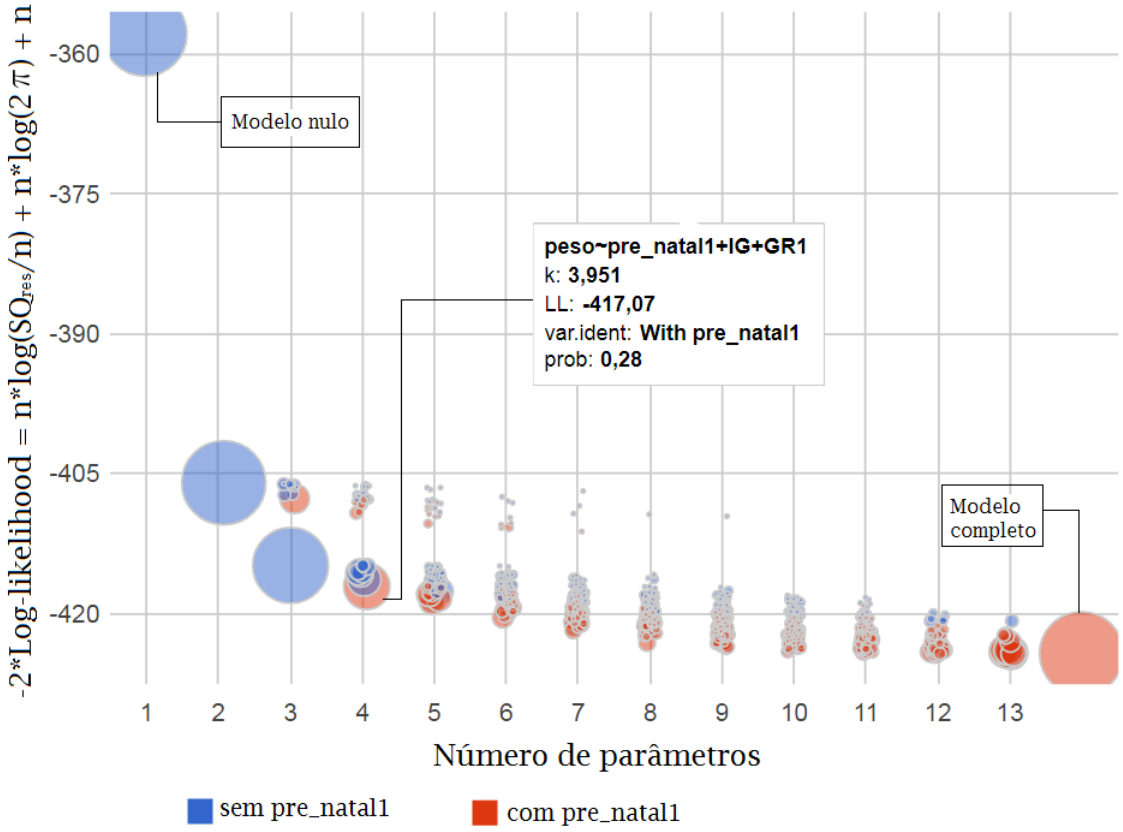

Figura 8 - Gráfico de probabilidades de seleção de modelos em função do número de parâmetros com destaque para a regressora pré-natal. Modelos de efeitos principais para ln(*peso*) de RN, utilizando *bootstrap* ponderado.

Tabela 10. Modelos com probabilidade de seleção superior a $0,20$ visualizados na Figura 8. Modelos de efeitos principais para ln(*peso*) de RN

Variável presente	Probabilidade
IG	1,0
IG e GR	0,77
IG, GR e pre_natal	0,28

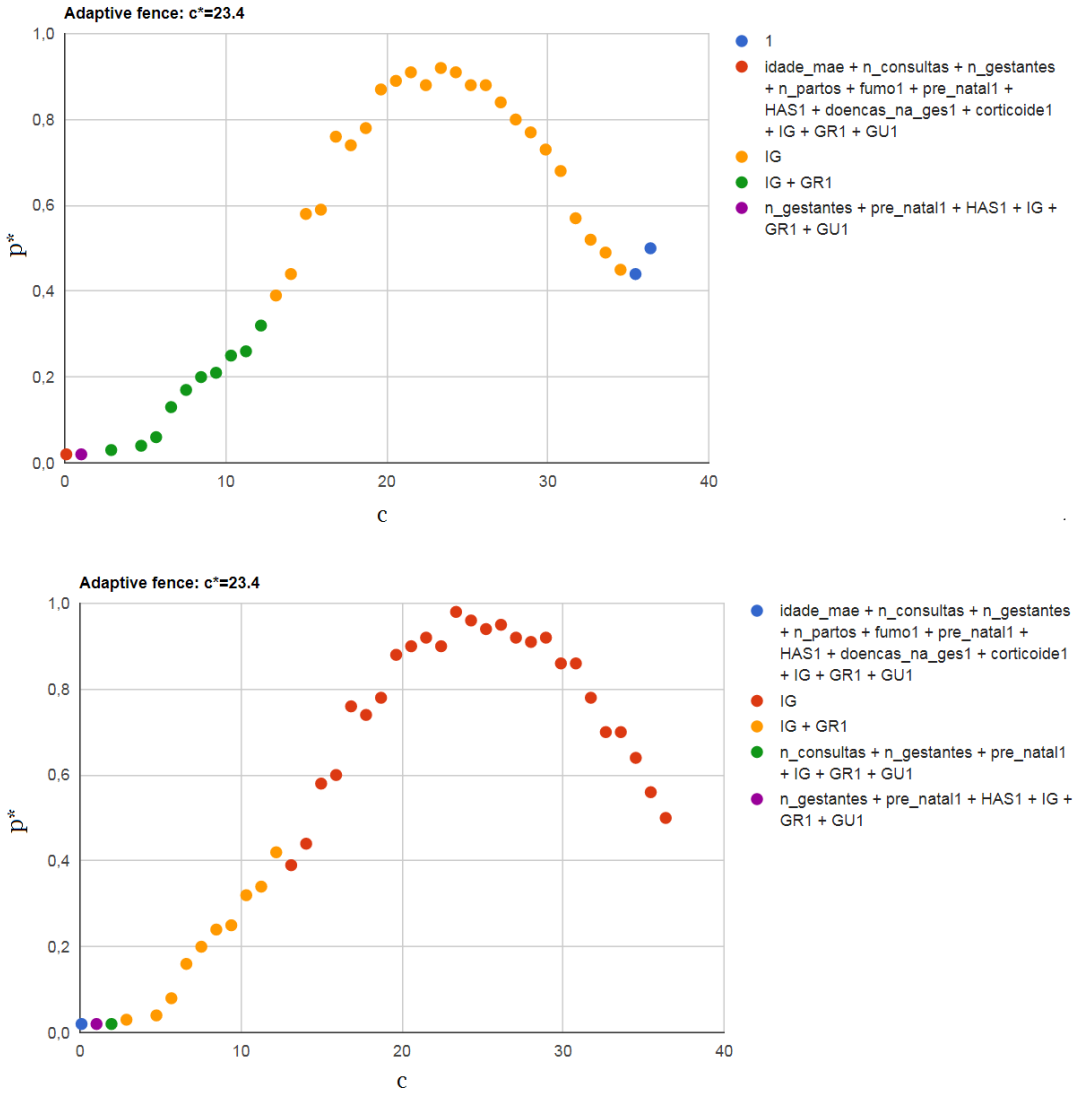

Figura 9 - Gráficos de probabilidades de seleção de modelos usando o método *fence*. Modelos de efeitos principais para o ln(*peso*) de RN.

A Figura 9 mostra o gráfico para avaliar a estabilidade de modelos usando o método *fence*. Nesse gráfico, deve-se procurar regiões em c que destaquem apenas um modelo e pontos com altas probabilidades (picos). O gráfico superior é a versão quando apenas o melhor modelo é considerado no cálculo de $p^*(\alpha, c)$ e o inferior quando todos são considerados e as probabilidades são ponderadas. Ambos apresentam padrão similar, pois as maiores probabilidades indicam o ajuste que

contem apenas a regressora `IG`. Outro modelo de destaque é o que contem as regressoras `IG` e `GR`, pois se destacam num intervalo de c com probabilidades consideráveis. Nos mesmos gráficos, os outros modelos que são visualizados não se destacaram em nenhum intervalo e suas probabilidades são pequenas, logo são descartados.

A Tabela 11 apresenta as regressoras dos modelos selecionados para investigação mais detalhada. Para os modelos 2 e 3 existe a possibilidade de investigação de inclusão de termos de interação. Para esta investigação, ajusta-se um modelo cuja variável resposta é formada pelos resíduos do ajuste do modelo com os efeitos principais, das regressoras de interesse, e a parte preditiva contém somente as interações, conforme indicado pelos próprios autores do pacote, Tarr et al. (2018).

Tabela 11. Modelos selecionados conforme aplicação do `mplot` para ln(*peso*) de RN

Modelo	Variável presente
1	IG
2	IG e GR
3	IG, GR e pré-natal

Os gráficos de inclusão de variáveis para as interações nos modelos 2 e 3 estão apresentados na Figura 10, indicando que a inclusão de interações não é importante em nenhum deles. No gráfico, nota-se que para os dois modelos, as interações apresentam curvas de probabilidades baixas que caem rapidamente com a penalidade estando próximas ou abaixo da curva de RV. A exploração das demais ferramentas, para estudar a relevância das interações, não revelou resultados diferentes. Portanto, busca-se pelas adequações dos modelos que contêm apenas os efeitos principais das regressoras indicadas na Tabela 11. Estimativas pontuais e intervalos com 95% de confiança são apresentados na Tabela 12.

Para fins de exploração, aplicou-se também a seleção de termos de interação de primeira ordem nos modelos selecionados pelos métodos automáticos (*stepwise*) usando os critérios AIC e BIC (Tabela 9). Em nenhum dos casos, termos de interação foram selecionados, indicando como modelos finais os mesmos modelos

2 e 3 da Tabela 11.

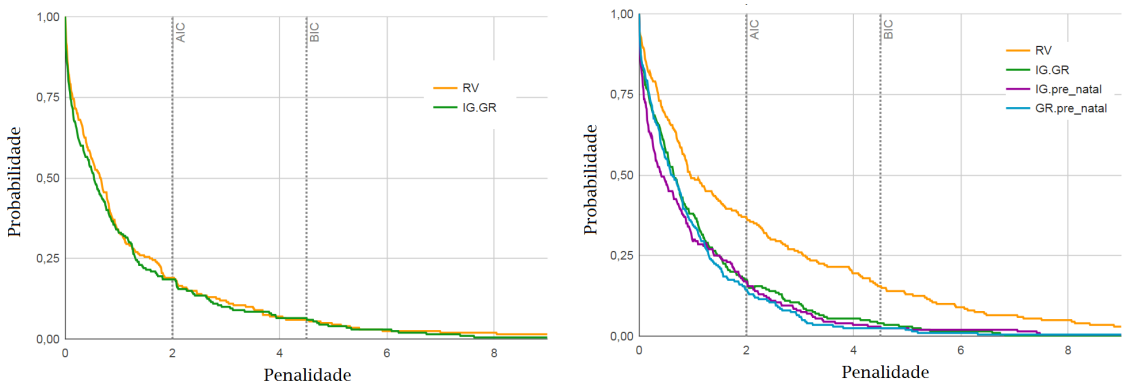

Figura 10 - Gráficos VIP dos termos de interação para os modelos 2 (à esquerda) e 3 (à direita) para o ln(*peso*) de RN.

Tabela 12. Estimativas dos parâmetros e intervalos de confiança (95%) para os ajustes dos modelos selecionados (Tabela 11) para ln(*peso*) de RN

	Estimativa (I.C.)		
Regressora	Modelo 1	Modelo 2	Modelo 3
Intercepto	5,228 (4,739; 5,717)	5,228 (4,759; 5,695)	5,231 (4,765; 5,696)
IG	0,064 (0,048; 0,080)	0,066 (0,050; 0,082)	0,064 (0,048; 0,079)
GR (feminino)	-	-0,112 (-0,186; -0,038)	-0,114 (-0,188; -0,041)
pre_natal (sim)	-	-	0,089 (-0,037; 0,215)

Análises de diagnóstico dos ajustes dos três modelos não evidenciaram problemas com os pressupostos e não foram encontrados pontos de influência consideráveis. Logo, obteve-se as estimativas dos parâmetros dos modelos finais e os respectivos intervalos de confiança, feitos pela função `confint` no `R`, apresentados na Tabela 12. Dessa forma, mantendo-se fixas as demais variáveis do modelo, conforme a idade gestacional cresce há um aumento esperado do peso. O efeito esperado de sexo é negativo, indicando que em média bebês do sexo feminino apresentam menor peso ao nascer. Avaliando o efeito de pré-natal, tem-se que, ao manter as outras

regressoras fixas, espera-se um aumento do peso. Deve-se notar, entretanto, que o IC para o efeito dessa variável, engloba o valor 0, indicando evidência fraca de efeito dessa variável. Deve-se destacar também que, das 90 observações, apenas 9 não fizeram o pré-natal, uma possível explicação para a falta de poder na detecção do efeito desse fator. Do ponto de vista prático, essa evidência fraca é importante já que alerta aos benefícios do pré-natal, principalmente em gestações de risco.

Avaliando o modelo 1, nota-se que para cada semana a mais de gestação, o peso estimado do RN ficará $6,6\%$ maior do que o peso da semana anterior. Para os outros modelos essa porcentagem é similar, já que suas estimativas são muito próximas.

Estimando a diferença entre sexos de RN em que $y = \ln(peso)$ no modelo 2, tem-se que

$$\widehat{y}(GR = 0) - \widehat{y}(GR = 1) =$$

$$= (5,228 + 0,066 \times IG - 0,112 \times 0) - (5,228 + 0,066 \times IG - 0,112 \times 1) = 0,112.$$

Como $\hat{y} = \ln(\widehat{peso})$, fazendo-se a tranformação tem-se $\widehat{peso} = 1,119$. Logo, a estimativa da diferença esperada entre os pesos dos RN do sexo masculino e feminino é constante, não importando os valores de `IG`, indicando que o RN do sexo masculino tem peso esperado de $11,9\%$ maior do que o sexo feminino.

Para o modelo 3 a estimativa da diferença da resposta entre o sexo masculino e feminino, considerando fixas as covariáveis `IG` e `pré_natal`, tem-se que

$$\hat{y}(GR = 0) - \hat{y}(GR = 1) =$$

$$= (5,231 + 0,064 \times IG - 0,114 \times 0 + 0,089 \times pre_natal) -$$

$$(5,231 + 0,064 \times IG - 0,114 \times 1 + 0,089 \times pre_natal) = 0,114.$$

Aplicando-se a transformação inversa tem-se $\widehat{peso} = 1,121$. Logo, estima-se uma diferença constante em que o valor esperado do peso do RN do sexo masculino é $12,1\%$ maior do que o sexo feminino. Considerando a realização ou não do pré-natal dado as outras covariáveis fixas, a diferença entre $\hat{y}$ é dado por

$$\hat{y}(pre_natal = 1) - \hat{y}(pre_natal = 0) =$$

$$= (5,231 + 0,064 \times IG - 0,114 \times GR + 0,089 \times 1) -$$

$$(5,231 + 0,064 \times IG - 0,114 \times GR + 0,089 \times 0) = 0,089.$$

Novamente, aplicando a inversa, tem-se que $\widehat{peso} = 1,93$. Logo, a estimativa da diferença dos pesos é constante, sendo que o RN cuja mãe realizou o pré-natal, apresenta uma estimativa de peso esperado de $9,3\%$ maior do que os que não realizaram.

A Figura 11, apresenta os valores de $\ln(peso)$ contra os valores de IG e as retas ajustadas segundo o modelo 2.

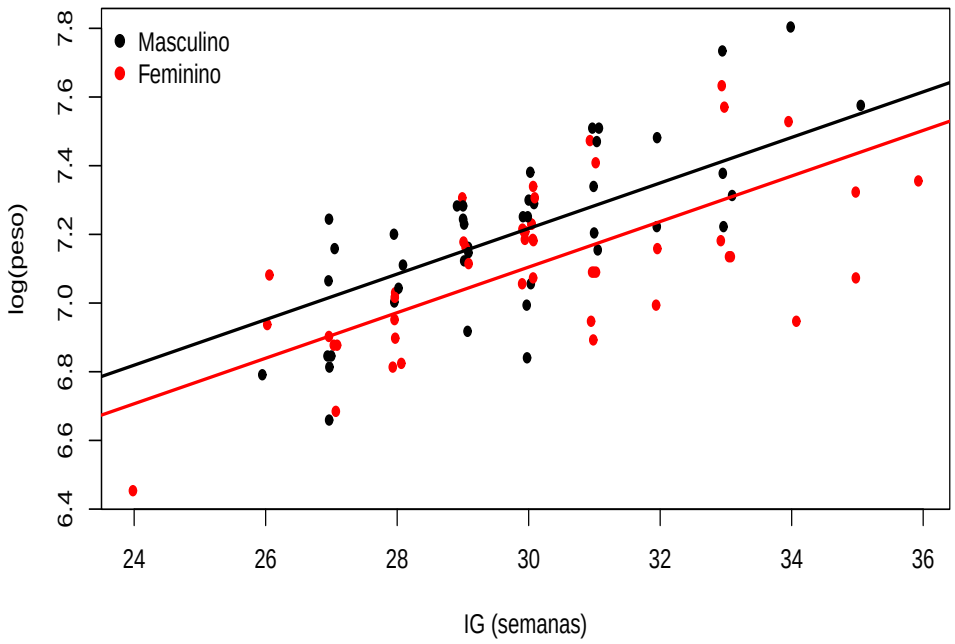

Figura 11 - Retas ajustadas para o $\ln(peso)$ de RN segundo o modelo 2.

Na Figura 12, tem-se os valores do $\ln(peso)$ contra a idade gestacional, com as retas ajustadas segundo o modelo 3. Nota-se que um recém-nascido do sexo masculino ao não realizar o pré-natal, tem seu $\ln(peso)$ esperado, e consequentemente, seu peso esperado, próximo de um recém-nascido do sexo feminino que realizou o pré-natal.

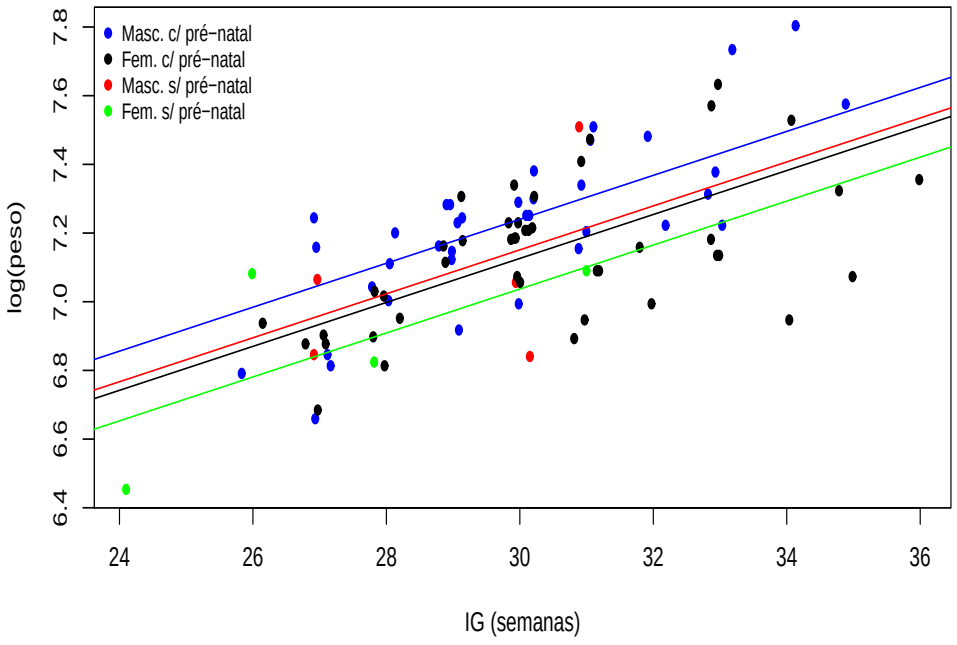

Figura 12 - Retas ajustadas para o ln($peso$) de RN segundo o modelo 3.

4.2 Aplicação 2: probabilidade de prenhez em inseminação artificial (IA) de vacas

Para estudar a relação entre a probabilidade de prenhez de vacas e as características do sêmen do touro, considerou-se, inicialmente, um modelo de regressão Binomial com função de ligação logística, e apenas os efeitos principais no preditor linear. Aqui destaca-se que o estudo apresenta um número razoável de covariáveis (12), além do fator de blocos, não sendo viável a exploração de um modelo muito complexo, por exemplo, incluindo termos de interação entre as covariáveis. Assim, antes de proceder com os métodos de seleção, realizou-se uma análise de diagnóstico preliminar na tentativa de explorar a presença de pontos problemáticos tanto no ajuste quanto no processo de seleção já que estes podem influenciar na seleção. Segundo Faraway (2016), na existência de observações influentes, estas devem ser, pelo menos temporariamente, removidas do conjunto de dados antes de aplicar qualquer estratégia de seleção de variáveis. Outro ponto a se destacar é que, para regressão Binomial, apenas a função de ligação logística está implementada no `mplot`.

A Figura 13 apresenta quatro gráficos para diagnóstico do ajuste do modelo preliminar, indicando alguns pontos de influência e razoável sobredispersão. Os pontos de maior destaque foram removidos um a um e os ajustes reavaliados até que um padrão relativamente aceitável foi obtido. Com isso 7 lotes (agregados) foram removidos, um número considerado grande, mas na ausência de outras alternativas, como por exemplo, coleta de observações de outras variáveis de relevância ou aumento do número de repetições, optou-se, a fim de ilustração da metodologia de seleção, prosseguir com o conjunto de dados reduzido, ou seja, com 65 lotes (agregados).

Os gráficos de diagnósticos que auxiliaram na retirada dos 7 lotes, são baseados nos mesmos apresentados nas Figuras 13 e 14. Entre os pontos retirados, cinco deles continham o maior número de vacas inseminadas de todo o experimento, enquanto os outros dois se destacam em algumas características de seu respectivo

bloco. Os quatro diagnósticos para o ajuste com 65 lotes são apresentados na Figura 14. Nela nota-se que alguns pontos do gráfico envelope estão fora da banda simulada, enquanto no gráfico à direita nenhum ponto se destacou. Nos gráficos de influência (parte inferior da figura), embora algumas observações se mostrem um pouco distantes das demais, não se considerou que fossem de influência grave no ajuste, mesmo com o gráfico à direita mostrando alguns pontos acima da linha de corte. O ajuste do modelo com $N = 65$ apresentou valor da função desvio igual a 72,19 com 49 graus de liberdade. Visto que, $\chi^2_{0,95;49} = 66{,}3$ e ao comparará-la com o valor do desvio, tem-se o indício de certo problema de falta de ajuste, embora não muito preocupante já que o modelo ainda não inclui nenhum termo de interação.

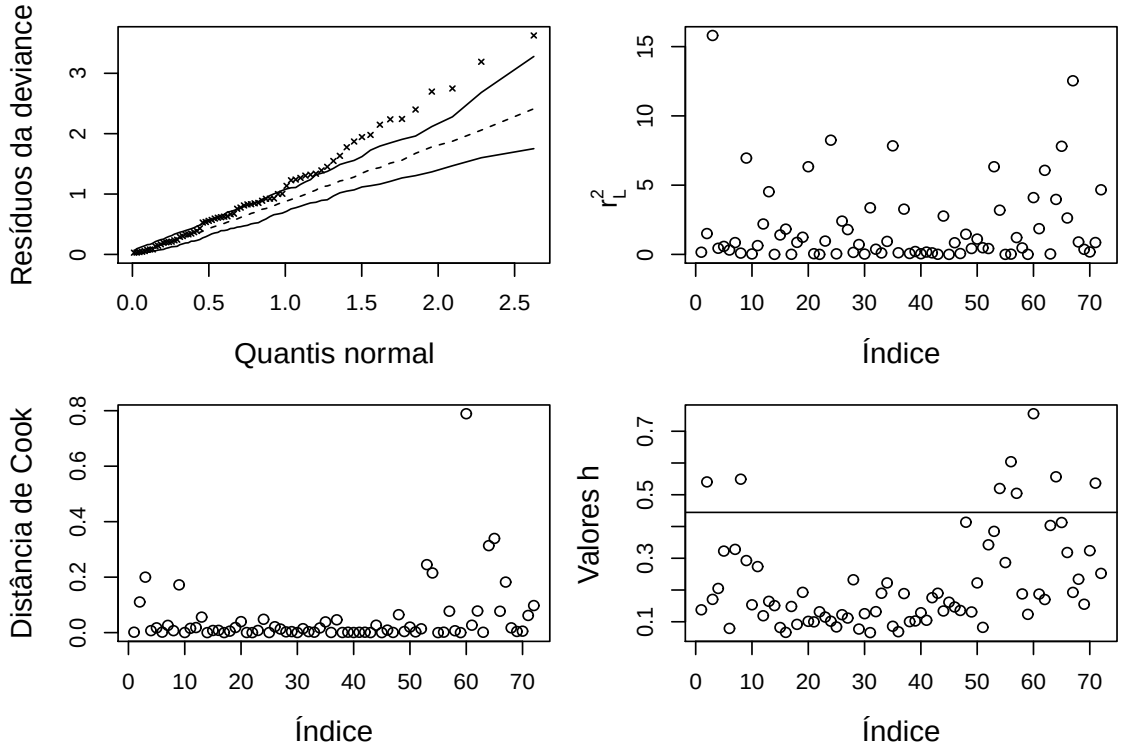

Figura 13 - Gráficos de diagnóstico para o ajuste do modelo de efeitos principais para a probabilidade de prenhez sob IA, dados agregados completos (N= 72).

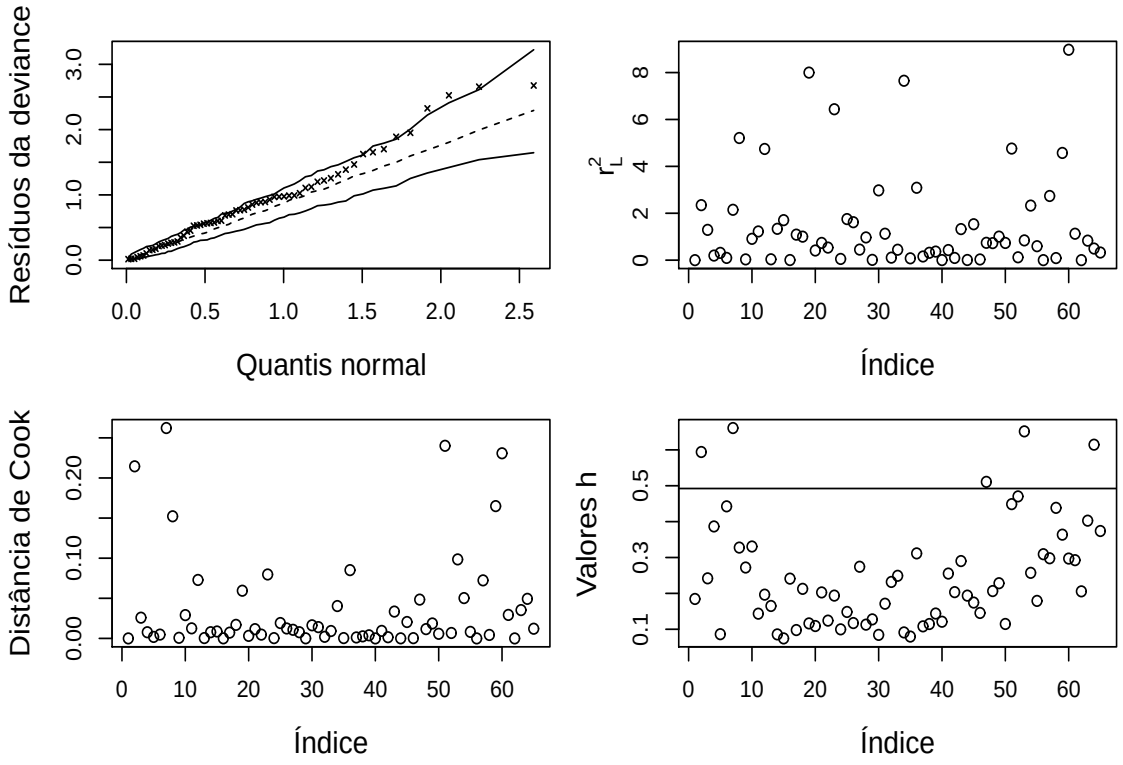

Figura 14 - Gráficos de diagnóstico para o ajuste do modelo de efeitos principais para a probabilidade de prenhez sob IA, dados agregados reduzidos (N= 65).

Como os dados apresentam uma variável de blocos, que carrega informações das condições e manejo das vacas sujeitas à inseminação, seu efeito deve ser considerado no modelo não estando sujeito à inclusão ou exclusão no processo de seleção de variáveis. Nos métodos *stepwise*, isso pode ser alcançado, declarando que o menor modelo a ser considerado inclui essa variável, sendo esse procedimento feito pela função `step`. Nesse sentido, aplicando essa metodologia ao modelo com os efeitos principais, dado que a variável de blocagem já pertence ao modelo, obteve-se os modelos com as regressoras apresentadas na Tabela 13. Nela verifica-se que para o critério BIC ($\lambda = \ln(65)$) não há mudanças nas variáveis presentes no modelo final, independentemente do método utilizado. Entretanto, ao usar o critério AIC ($\lambda = 2$), nota-se que os métodos *backward* e *stepwise* resultam em modelos maiores do que

o método *forward*. Sendo assim, tem-se três possíveis modelos de efeitos principais selecionados pelos métodos e critérios usuais.

Tabela 13. Regressoras presentes nos modelos selecionados segundo os métodos usuais de seleção de variíveis, dados de prenhez sob IA agregados (N= 65)

	Critério	
Método	AIC	BIC
forward	bloco, mot, lin e str	bloco, mot e str
backward	bloco, mitp, mot, vsl, vcl, progr, lin e str	bloco, mot e str
stepwise	bloco, mitp, mot, vsl, vcl, progr, lin e str	bloco, mot e str

O `mplot` não tem a opção para inclusão obrigatória de uma regressora no modelo. Então, a solução encontrada para o fator bloco foi trabalhar com os resíduos de um modelo com essa variável ajustada. Sabe-se que nos modelos lineares, quando se ajusta os resíduos do modelo sem a regressora X_k, ou seja, X_k temporariamente removida, em função dos resíduos de X_k ajustada pelas outras regressoras, recupera-se o efeito de X_k sobre a resposta. Porém, nos modelos lineares generalizados, essa equivalência não ocorre, já que o componente aleatório nesses modelos está implícito na distribuição de probabilidade da variável resposta, não aparecendo aditivamente na equação do preditor, como acontece nos modelos lineares. Não foi encontrado referência bibliográfica sobre esse assunto, mas reflexão de resolução desse problema nessa pesquisa levou ao argumento da não aplicabilidade desse princípio. Destaca-se que, uma vez ajustado o modelo de regressão logística (resposta sendo assumida com distribuição Binomial ou Bernoulli), não há suporte para a suposição de Normalidade aos resíduos e portanto, o princípio falha.

Assim, foi necessário utilizar os dados na forma expandida (resposta binária para cada vaca), ajustar a regressão logística, obter as probabilidades estimadas e construir a variável de linearização do modelo (z), conforme o algoritmo apresentado na Seção 3.2. Uma vez com a variável linearizada, foi possível a aplicação do modelo linear e, então, eliminou-se o efeito de blocos e ajustou-se o modelo com

os resíduos das demais regressoras, também eliminando-se o efeito de blocos. Ainda, para satisfazer o modelo logístico, usando a linearização, foi necessário a inclusão de um peso para cada observação definido como $v_i = \hat{\pi}_i(1 - \hat{\pi}_i)$. Pesos são utilizados para lidar com a heterogeneidade de variâncias implícita nas distribuições Bernoulli com parâmetros π_i. Já a variável z é dada por $z_i = \ln(\hat{\pi}_i/(1-\hat{\pi}_i)) + r_w$, para qual $\hat{\pi}_i$ é a probabilidade de sucesso estimada na regressão logística com dados binários e r_w os seus resíduos de trabalho (*working residuals*). O vetor de resíduos para o ajuste de cada regressora em função do fator bloco entrou como regressora no modelo ao qual os processos de seleção de variáveis foram aplicados. Assim, esses vetores foram renomeados pelo nome da covariável acrescido pela letra `r`. As novas regressoras são, portanto, `mintr`, `mitpr`, `motr`, `deftr`, `defmr`, `vslr`, `vclr`, `vapr`, `alhr`, `progrr`, `linr` e `strr` e a resposta definida como `r`, para explicitar que o efeito de blocos foi removido.

Aplicando a metodologia usual para o processo de seleção de variáveis no modelo linearizado, obteve-se a Tabela 14 que mostra as regressoras presentes segundo o método e critério utilizados. Obteve-se o mesmo modelo usando o critério BIC ($\lambda = \ln(2399)$) independente do método. Devido à altíssima penalização, o modelo apresenta apenas uma regressora (`strr`). Já para o critério AIC ($\lambda = 2$), obtém-se, obviamente, os mesmos modelos finais da Tabela 13, já que os modelos ajustados são equivalentes, independentemente da estratégia utilizada (linearização ou `glm`).

Tabela 14. Regressoras presentes no modelo linearizado segundo os métodos usuais para seleção de variáveis, dados de prenhez sob IA

	Critério	
Método	AIC	BIC
forward	motr, linr e strr	strr
backward	mitpr, motr, vslr, vclr, progrr, linr e strr	strr
stepwise	mitpr, motr, vslr, vclr, progrr, linr e strr	strr

Utilizando as ferramentas concedidas pelo `mplot` com a ideia de auxiliar no processo de seleção de variáveis, observa-se, primeiramente, o gráfico de inclusão de variáveis, apresentado na Figura 15. Nota-se que para penalidades baixas, seis variáveis se destacam, com `strr` se mantendo com as probabilidades mais altas em todo o intervalo considerado. Das outras cinco, quatro têm probabilidades que decrescem rapidamente com o aumento da penalidade, sendo elas, `motr`, `linr`, `vslr` e `vclr`, com as duas últimas chegando bem próximo da linha de RV. Já a variável `progrr`, que para baixas probabilidades se destacava menos que `motr` e `linr`, passa a ter destaque para penalidades acima de 5,5. Esse comportamento é característico quando existe colinearidade entre grupo de regressoras.

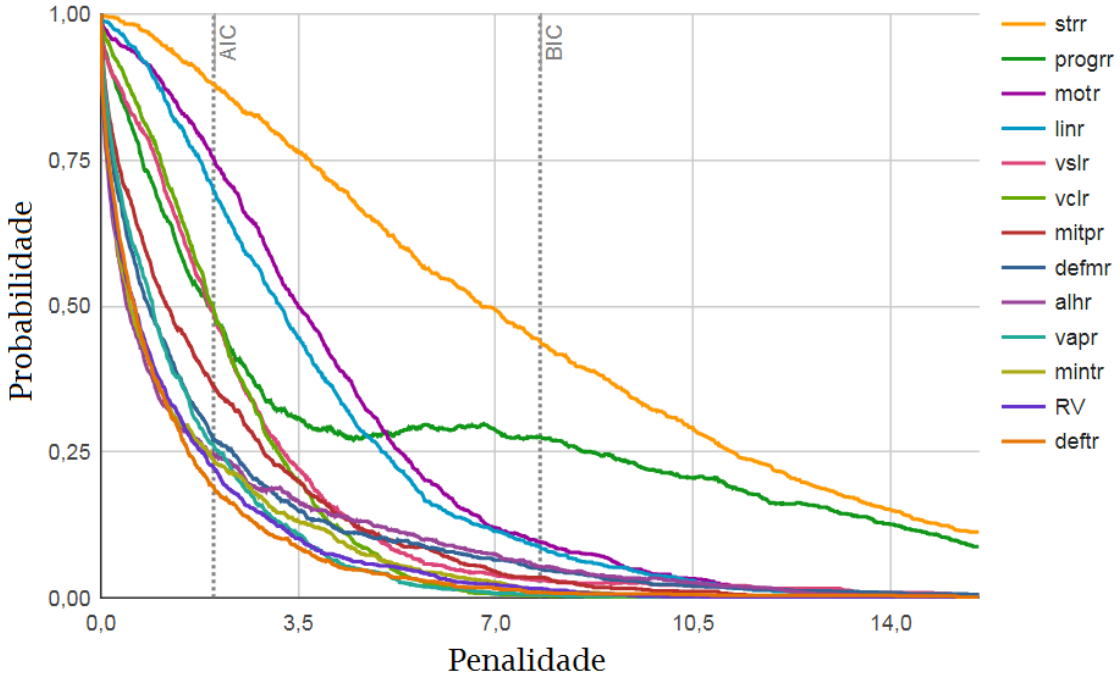

Figura 15 - Gráfico de inclusão de variáveis (VIP) para o modelo linearizado.

A Figura 16 mostra cinco gráficos do componente de perda em função da dimensão do modelo, onde cada círculo indica, proporcionalmente a frequência de vezes em que um determinado modelo foi selecionado, em uma quantidade de reamostras *bootstrap*. Em cada um dos gráficos, destaca-se uma variável para análise.

No gráfico que enfatiza a regressora `strr`, nota-se que, para todas as dimensões, ela é relevante e, quando sozinha, tem o menor valor de componente de perda com alta probabilidade. No gráfico que destaca a variável `progrr`, verifica-se também, que a mesma contém alta probabilidade quando sozinha, porém, torna-se relevante apenas em modelos acima de sete parâmetros e estes têm baixas probabilidades, não aparentando então, diminuir de forma relevante seus valores de componente de perda. Considerando o gráfico que realça `linr`, notam-se dois modelos de baixa dimensão que a contêm e que possuem probabilidades razoavelmente altas, quando comparadas às dos outros modelos de mesma dimensão. É notável também, que a partir da dimensão quatro, ela está presente em todos os modelos com os menores valores do componente de perda. Como pode ser visto no gráfico que destaca `motr`, esta variável tem comportamento similar a `linr` e está presente em todos os modelos com os menores valores para o componente de perda a partir da dimensão três. Outro modelo com dimensão três que destacou-se, contém as covariáveis `alhr` e `strr`, devido à sua alta probabilidade e baixo componente de perda, comparado aos modelos de mesma dimensão, ele se torna um candidato para análises mais detalhadas.

A Tabela 15 mostra os modelos que apareceram com probabilidade de seleção maior do que 0,15 nas reamostragens. A tabela não mostra os modelos que continham a variável redundante e um modelo com doze regressoras que foram selecionadas em 16% das reamostras, devido aos valores do componente de perda não atrativos. As variáveis `vclr` e `vslr` destacadas na Figura 15, mostraram ser pouco relevantes e só foram incluídas em modelos com baixa probabilidade e de dimensão maior que seis, logo deve-se buscar por outras alternativas com o intuito de descobrir se são relevantes.

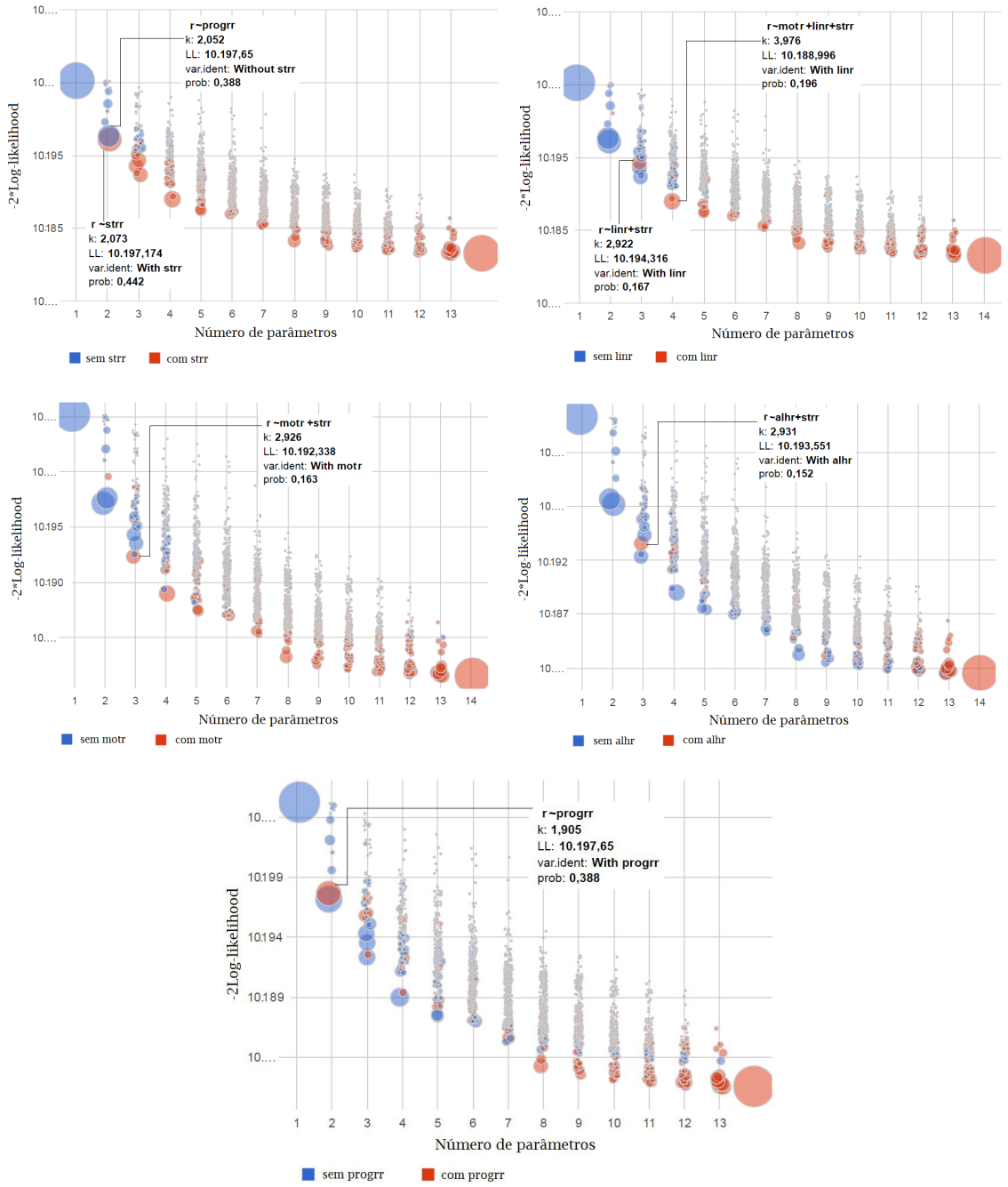

Figura 16 - Gráfico de probabilidade de seleção de modelos em função do número de parâmetros, com destaque para as regressoras, visualizadas da esquerda para a direita e de cima a baixo, `strr`, `linr`, `motr`, `alhr` e `progrr`, dados linearizados.

Tabela 15. Modelos com destaque probabilístico, conforme Figura 16, dados linearizados

Modelo	Regressora	Probabilidade
1	strr	0,44
2	progrr	0,39
3	linr e strr	0,17
4	motr e strr	0,16
5	alhr e strr	0,15
6	motr, linr e strr	0,20

Considerando a Figura 17, busca-se por modelos que estejam presentes em um intervalo de c e/ou com probabilidade relevante, para que se possa identificar novos possíveis modelos. Entretanto, o único modelo que se manifestou num intervalo considerável, mas com probabilidades baixas, é o que contém somente a covariável `strr`. Nesse sentido, não foi possível identificar modelos estáveis, o que torna os da Tabela 15 como principais ajustes para investigar possíveis interações de primeira ordem.

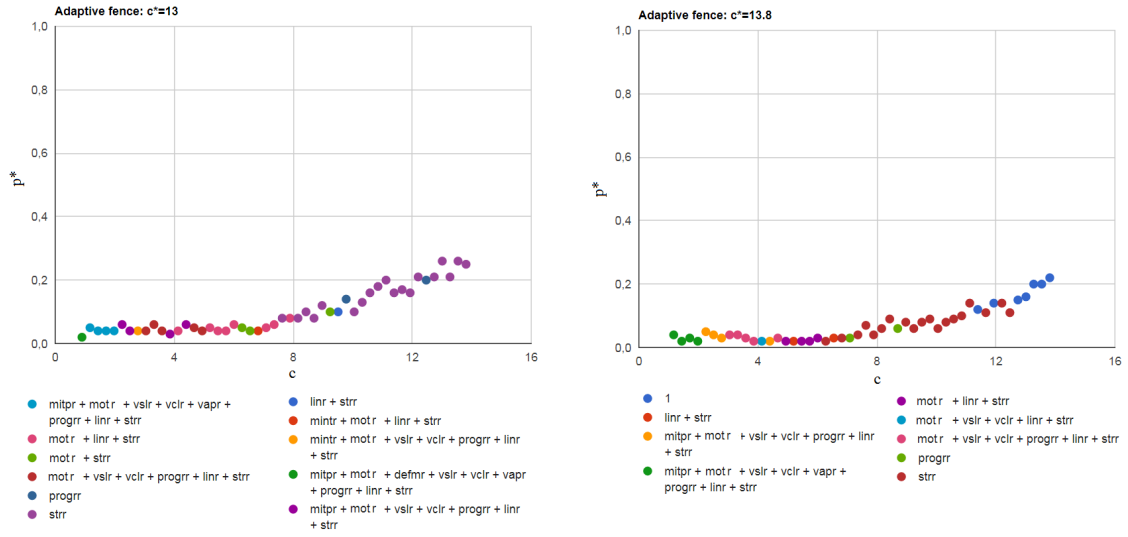

Figura 17 - Gráfico de probabilidade de seleção de modelos, dados linearizados.

Para os modelos 3, 4, 5 e 6, com o intuito de investigar a importância de termos de interação de primeira ordem entre as regressoras, foram construídos novos ajustes que contêm como variável resposta os resíduos do modelo com o efeito principal, das regressoras de interesse, e a parte preditiva contém somente as interações. Note que esses resíduos são resíduos de um modelo cuja resposta já é um termo residual, aquele do modelo que levou em conta a contribuição da covariável bloco. Dado essa construção, aplicou-se a metodologia de seleção para os termos de interação utilizando os métodos usuais e o pacote `mplot`.

Os gráficos VIP para os modelos 3 e 5 são apresentados na Figura 18. Nesse caso, conforme há aumento da penalidade, a curva de probabilidades dos termos de interação diminuem rapidamente e se confundem com a curva de RV, indicando que as interações nos modelos 3 e 5 não trazem ganho para a explicação da probabilidade de prenhez.

A Figura 19 mostra o gráfico VIP para os modelos 4 e 6. Nota-se que, para baixas penalidades, a probabilidade da interação nos dois casos é razoavelmente alta. Entretanto, com o aumento de λ as probabilidades diminuíram rapidamente, mas continuam separadas da curva RV.

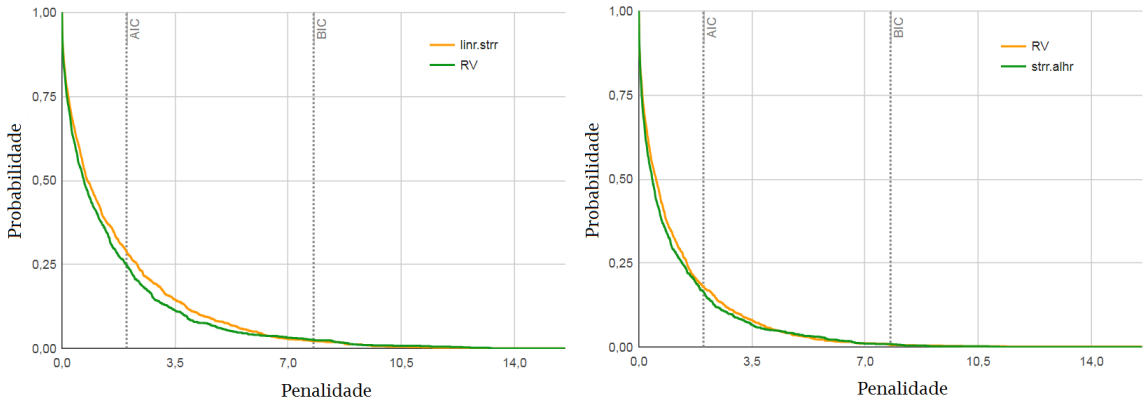

Figura 18 - Gráficos VIP para a interação dos modelos 3 (à esquerda) e 5 (à direita), dados na Tabela 15, dados linearizados.

Visto a possibilidade de incluir termos de interação, buscou-se, primeiramente para o modelo 4, usando o gráfico de estabilidade (não apresentado), a proporção de vezes em que o modelo com a interação `strr` × `motr` foi selecionada como o melhor. O valor dessa proporção foi de 0,75, sendo assim, optou-se pela inclusão da mesma.

Considerando o modelo 6, no gráfico à direita da Figura 19, os destaques ficam para as interações `strr` × `linr` e `strr` × `motr`. As probabilidades de seleção dos modelos que as incluem, dadas as reamostras, são apresentadas na Tabela 16. Como as probabilidades apresentam valores razoáveis, pode-se obter algum benefício optando-se pela inclusão dos dois termos de interação de primeira ordem.

Devido aos poucos termos de interação encontrados nos modelos 3, 4, 5 e 6, os gráficos que utilizam como base o método *fence* não são informativos, visto que suas probabilidades não se destacaram o suficiente para a inclusão de termos. Portanto, os modelos finais utilizando as ferramentas gráficas no processo de seleção são apresentados na Tabela 17.

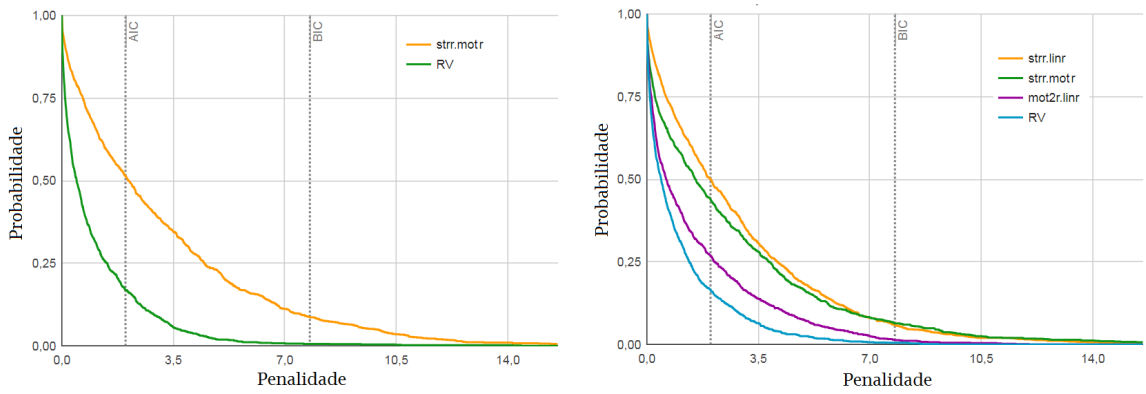

Figura 19 - Gráficos VIP para a interação dos modelos 4 (à esquerda) e 6 (à direita), dados na Tabela 15, dados linearizados.

Tabela 16. Modelos com destaque probabilístico, encontrados no gráfico de probabilidades de seleção de modelos em função do número de parâmetros (não apresentado)

Interação	Probabilidade
strr×linr	0,40
strr×motr	0,34
strr×linr + strr×motr	0,37

Aplicaram-se também os métodos tipo *stepwise* para a seleção de termos de interação nos modelos de efeitos principais encontrados na Tabela 14, cujos resultados são apresentados na Tabela 18. Para o método *forward* e critério AIC, o modelo final coincide com um dos modelos de efeitos principais destacados pela análise dos resultados via `mplot`, porém, diferentemente do indicado na Figura 19, nenhuma interação permanece no modelo. Já pelos métodos *backward* e *stepwise* e critério AIC, o modelo de efeitos principais é bem maior e ainda inclui seis termos de interação. O modelo de maior dimensão indicado nas análises das ferramentas gráficas contém três efeitos principais e duas interações.

Tabela 17. Parte preditiva dos modelos utilizando-se o `mplot`, dados linearizados

Modelo	Variáveis presentes
1	strr + motr + strr × motr
2	strr + linr + motr + strr × linr + strr × motr

Tabela 18. Regressoras presentes nos modelos finais, feita a seleção de variáveis com a metologia usual do modelo linearizado

		Regressora presente	
Critério + Método		Efeito Principal	Interação
AIC +	Forward	linr, motr e strr	-
AIC +	Backward Stepwise	linr, strr, motr, mitpr, vslr, vclr e progrr	linr×strr, strr×vslr, strr×vclr, strr×mitpr, vclr×motr e vslr×mitpr
BIC +	Forward Backward Stepwise	strr	-

Após a seleção de variáveis feita pelo `mplot`, retornou-se ao modelo de dados agregados. Considerando o modelo 1 (Tabela 17), fez-se uma tabela ANODEV que não indicou significância para o termo de interação. O diagnóstico da Figura 20 se refere ao ajuste com a variável de blocos, `str` e `mot`.

Uma análise da ANODEV também foi feita para o modelo 2 (Tabela 17), indicando que os efeitos de interação têm pouca evidência de efeito sobre a prenhez. Com isso, fez-se a análise de diagnóstico do ajuste que contém apenas os efeitos da variável de blocos, `str`, `mot` e `lin` (Figura 21). Nota-se, então, uma melhora significativa deste ajuste comparado ao anterior. Portanto, decidiu-se por uma interpretação de seus parâmetros juntamente com seus respectivos intervalos de confiança (95%), que foram obtidos a partir da função `confint` do `R`.

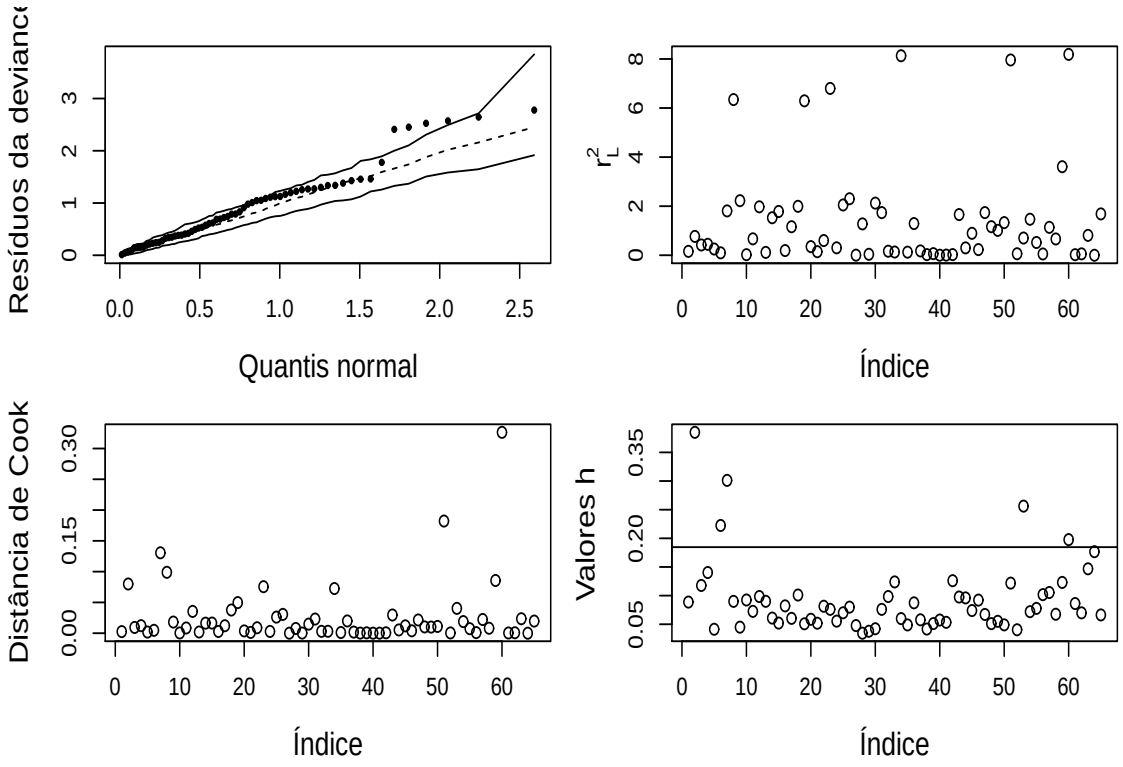

Figura 20 - Diagnóstico do modelo com efeitos da variável de blocos, `str` e `mot`, dados reduzidos (N= 65).

Tabela 19. Estimativas e IC's (95%) para os parâmetros do ajuste final, dados reduzidos (N= 65)

Regressora	Estimativa	IC (95%)
Intercepto	5,009	(1,9010; 8,1330)
Bloco1	0,5085	(0,2142; 0,8046)
Bloco2	-0,0151	(-0,2807; 0,2504)
Bloco3	-0,1100	(-0,3412; 0,1212)
str	-0,0884	(-0,1527; -0,0243)
mot	-0,0069	(-0,0128; -0,0011)
lin	0,0470	(-0,0029; 0,0971)

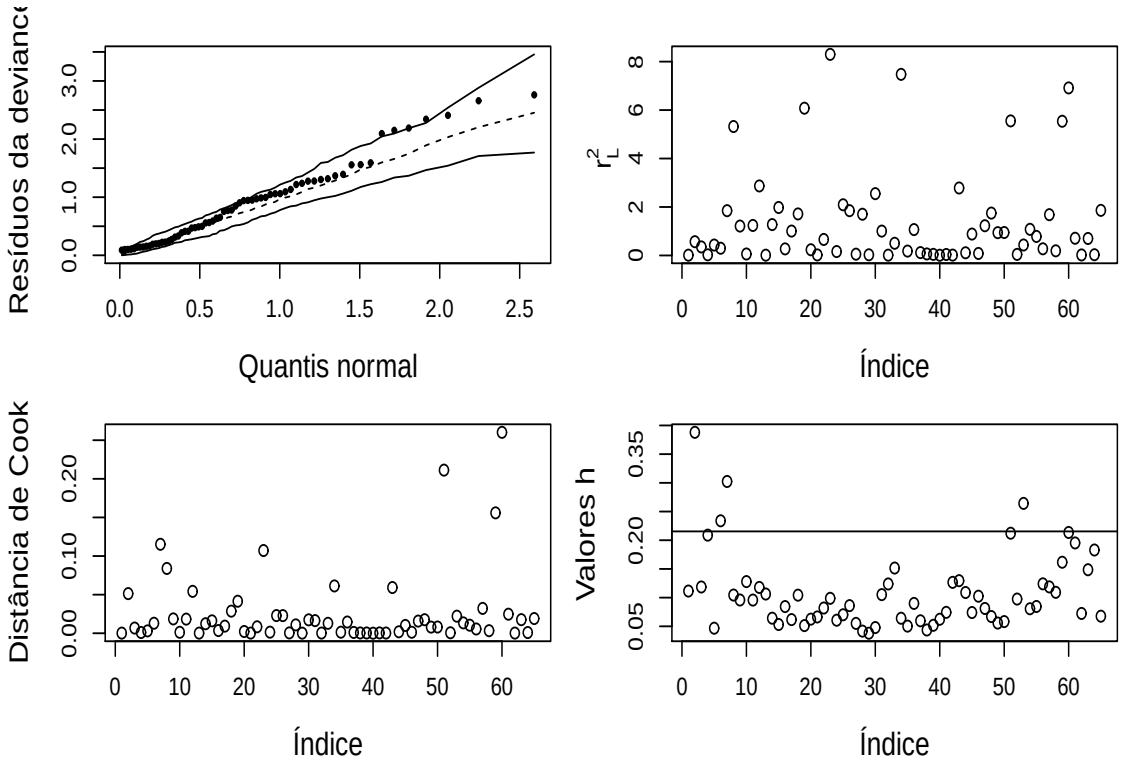

Figura 21 - Diagnóstico do modelo com efeitos da variável de blocos, `str`, `mot` e `lin`, dados reduzidos (N= 65).

Uma estimativa da probabilidade média de prenhez, usando todos os valores das covariáveis no conjunto de dados é $0,4808$. Dado a Tabela 19, estima-se a razão de chances (OR) para cada regressora. Suas estimativas ($\widehat{OR}$) e intervalo de confiança (95%) são apresentadas na Tabela 20.

Para a regressora `str`, conforme seus valores aumentam e as demais permanecem fixas, a estimativa da chance esperada de prenhez tende a diminuir. Supondo um aumento de 10 unidades, factível com os dados apresentados, tem-se, $\widehat{OR}_{str} = exp[-0,0884 \times 10] = 0,413$, ou seja, a chance de prenhez é 0,413 vezes menor do que a chance para 10 unidades a menos em `str`. A regressora `mot` indica que há uma pequena diminuição nas chances de prenhez no aumento de uma unidade. Entretanto, ao avaliar uma diferença de 50 unidades tem-se, $\widehat{OR}_{mot} =$

$exp[-0,0069 \times 50] = 0,7$. Visto que os valores de `mot` são indicados em forma de porcentagem, é plausível que haja tal diferença abordada. Logo, se o pesquisador busca por uma chance maior de prenhez, deve-se obter uma baixa porcentagem para `mot`. A chance de prenhez para a variável `lin` tende a aumentar conforme há o acréscimo em seu valor, supondo uma diferença de 20 unidades, tem-se, $\widehat{OR}_{lin} = exp[0,047 \times 20] = 2,56$. Logo, a chance de prenhez aumenta em 2,56 vezes na variação de 20 unidades proposta, factível com os dados apresentados.

Tabela 20. Estimativas da razão de chances e respectivas IC's (95%), para o modelo incluindo a variável de blocos, `mot`, `str` e `lin`, para os dados de prenhez sob IA

	Estimativa	IC (95%)
str	0,915	(0,859; 0,976)
mot	0,993	(0,987; 0,999)
lin	1,048	(0,997; 1,102)

5 CONSIDERAÇÕES FINAIS

O problema de seleção de variáveis, em modelos de regressão, é bastante complexo e não existe um método que pode ser considerado aquele que oferece a resposta mais apropriada em todas as aplicações. A bibliografia especializada apresenta uma variedade de critérios e métodos automáticos que, não raramente, quando aplicados ao mesmo conjunto de dados, oferece soluções distintas. Talvez, a maior crítica ou desvantagem do uso desses métodos automáticos seja que, por serem automáticos, não há intervenção do estatístico ou do pesquisador especialista da área da aplicação. Assim, a oportunidade de investigação de um modelo que, do ponto de vista prático, pode ser interessante, é desperdiçada quando se usa procedimentos automáticos. Efron et al. (2004), cujo argumento foi resgatado em Tarr et al. (2018), coloca um parágrafo que traduz muito bem o processo de seleção ou construção de modelos úteis na prática, que diz, numa tradução livre,

"Na prática, ou pelo menos na boa prática, acontece um ciclo de atividades entre o pesquisador, o estatístico e o computador ... O estatístico examina os resultados criticamente, discutindo os resultados com o especialista da área que pode, nesse ponto, sugerir a inclusão ou a remoção de variáveis explanatórias, e assim por diante, e outras coisas mais".

Com a disponibilidade de recursos computacionais de alto desempenho, Tarr et al. (2018) organizaram uma coleção de técnicas capazes de extrair informações sobre as contribuições de variáveis regressoras a agrupamento de modelos competitivos, sob diversos aspectos e critérios. As informações são então disponibilizadas na forma gráfica, usando uma plataforma que permite interação com o usuário. As técnicas utilizadas, são baseadas na reamostragem dos dados e são implementadas

no pacote `mplot`.

O `mplot` não oferece a resposta do "melhor" modelo e nem ordena modelos. O usuário deve, criticamente, analisar as informações disponibilizadas, adquirir conhecimento sobre o relacionamento entre as variáveis, avaliar modelos competitivos e estabelecer o racional com a prática, de preferência, em discussão com o especialista da área.

Nesse trabalho, as ferramentas foram aplicadas para formular modelos a partir de dois conjuntos de dados, um sobre o peso de recém-nascidos prematuros, e o outro sobre sucesso na inseminação artificial de vacas. Ambos os conjuntos ofereceram desafios extras para a modelagem. O primeiro, com variável resposta contínua, mostrou indícios de heterogeneidade de variâncias. O problema foi, então, amenizado utilizando uma transformação da família Box-Cox. O segundo, com variável resposta binária, inclui como covariável um fator de blocos que não deve jamais ser excluído do modelo e o maior desafio foi encontrar uma estratégia para isso, já que o `mplot` não tem tal flexibilidade. A estratégia utilizada foi usar a versão linearizada do modelo de regressão logística e trabalhar com os resíduos das variáveis ajustadas pelo fator de blocos. Nesse conjunto de dados a modelagem é bastante complexa, com correlações altíssimas entre as covariáveis, além da suspeita de que existem outras covariáveis (não observadas na pesquisa) exercendo influência nas respostas. O diagnóstico de ajuste preliminar, indicou observações problemáticas que foram removidas nesta aplicação. No entanto, a remoção de observações é um ponto bastante sensível, sendo necessário estudos mais aprofundados para sua justificativa, inclusive, avaliando a sensibilidade das ferramentas gráficas a essas observações. Esse estudo faz parte dos objetivos de pesquisas futuras.

Com as duas aplicações apresentadas nessa dissertação, os objetivos dessa pesquisa foram cumpridos, pois foi possível explorar as informações disponibilizadas e mostrar a utilidade das ferramentas gráficas no auxílio da construção de modelos. Ressalta-se que os modelos finais, destacados para cada aplicação, são apenas ilustrações de possíveis modelos que poderiam ter sido construídos. Com a

contribuição mais ativa do especialista da área, outros modelos poderiam ter sido descobertos.

Apesar do grande potencial do pacote `mplot`, ele apresenta limitações, algumas das quais já foram mencionadas. Os estudos aqui realizados permitem destacar:

1. Não é possível incluir uma variável compulsória no modelo, por exemplo blocos ou qualquer outra que o especialista da área faz questão;

2. Para regressão Binomial, os dados devem estar no formato expandido (binário) ou a resposta ser a proporção amostral nos "grupos". Neste último caso, se os grupos forem desbalanceados (n_i distintos) deve-se incluir pesos usando a opção `weights`;

3. Ainda para regressão Binomial, apenas a função de ligação logito está implementada;

4. Não está apto a trabalhar com conjuntos de dados que apresentam regressoras qualitativas com mais de duas categorias.

Anexos

Anexo 1

Abaixo encontram-se os códigos utilizados para a construção e geração das ferramentas gráficas para a Aplicação 1.

```
R> DADOS $ Y = log(DADOS $ peso)
R> fit.model.01 = lm(Y ~ idade_mae + n_consultas + n_gestantes +
n_partos + fumo + pre_natal + HAS + doencas_na_ges + corticoide +
IG + GR + GU, data=DADOS)
R> vis.art = vis(fit.model.01, B=500, redundant = TRUE)
R> af.art = af(fit.model.01, B=200, n.c=40)
R> mplot(fit.model.01, vis.art, af.art)
```

Anexo 2

Os códigos abaixo foram utilizados para construção e geração das ferramentas gráficas referentes a Aplicação 2. Note que o ajuste apresentado é o

linearizado e com o efeito de blocos.

```
R> fit.lm = lm(r ~ mintr + mitpr + motr + deftr + defmr + vslr +
vclr + vapr + alhr + progrr + linr + strr, data=Dadosr, weight=v)
R> fit.model.vis = vis(fit.lm, B=1000)
R> fit.model.af = af(fit.lm, B=200, n.c=50)
R> mplot(fit.lm, fit.model.vis, fit.model.af)
```

Anexo 3

Parcela de dados referentes a Aplicação 1.

peso	idade_mae	n_consultas	n_gestantes	n_partos	fumo	pre_natal	HAS
1380	31	5	4	3	0	1	0
1255	28	5	4	3	1	1	0
780	28	5	1	0	0	1	0
995	35	5	1	0	0	1	0
910	16	4	1	0	0	1	0
1280	23	2	3	1	0	1	0
920	24	0	1	0	0	0	0
1940	31	8	1	0	0	1	1
2450	33	3	3	1	0	1	0
2065	28	2	2	1	0	1	0
1605	24	5	1	0	0	1	0
2285	41	3	3	0	0	1	0
1825	22	3	1	0	0	1	0
1860	21	4	3	2	1	1	0
1825	18	0	1	0	0	0	0
1650	24	4	3	2	0	1	0
1540	18	3	2	0	0	1	0
1370	39	7	3	2	0	1	1
1090	16	4	1	0	0	1	0
1400	19	3	1	0	1	1	0

doenca_na_ges	corticoide	IG	GR	GU	Y
1	0	29	0	1	7,2298
1	0	33	1	1	7,1348
1	1	27	0	1	6,6592
1	1	27	1	1	6,9027
1	1	27	0	1	6,8134
1	1	31	0	1	7,1546
1	1	28	1	1	6,8243
1	0	33	1	1	7,5704
1	0	34	0	1	7,8038
1	0	33	1	1	7,6328
1	0	30	1	1	7,3808
0	1	33	1	1	7,7341
0	1	31	1	1	7,5093
1	0	34	1	1	7,5283
1	1	31	1	1	7,5093
1	1	31	1	1	7,4085
1	1	30	1	1	7,3395
1	0	33	1	1	7,2225
1	0	32	1	1	6,9939
1	0	27	1	1	7,2442

REFERÊNCIAS BIBLIOGRÁFICAS

AGRESTI, A. **Categorial data analysis**. John Wiley & Sons, 2012.

AKAIKE, H. Information theory and the maximum likelihood principle in 2nd International Symposium on Information Theory (BN Petrov and F. Cs ä ki, eds.). **Akademiai Ki à do, Budapest**, 1973.

ATKINSON, A. C. Two graphical displays for outlying and influential observations in regression. **Biometrika**, v.68, n.1, p.13–20, 1981.

BOX, G. E.; COX, D. R. An analysis of transformations. **Journal of the Royal Statistical Society. Series B (Methodological)**, p.211–252, 1964.

BOX, G. E.; DRAPER, N. R. **Empirical model-building and response surfaces.**. John Wiley & Sons, 1987.

CHARNET, R.; FREIRE, C. A. D. L.; CHARNET, E. M. R.; BONVINO, H. **Análise de modelos de regressão linear com aplicações**. Editora UNICAMP, 2015.

CHATFIELD, C.; ZIDEK, J.; LINDSEY, J. **An introduction to generalized linear models**. Chapman and Hall/CRC, 2010.

COLLETT, D. **Modelling binary data**. Chapman & Hall/CRC, 2002.

COOK, R. D. Detection of influential observation in linear regression. **Technometrics**, v.19, n.1, p.15–18, 1977.

DAVISON, A. C. **Statistical models**. Cambridge University Press, 2003. 11v.

DOBSON, A. J.; BARNETT, . A. G. **An Introduction to Generalized Linear Models**. Chapman and Hall/CRC, 2018.

DRAPER, N. R.; SMITH, H. **Applied regression analysis**. John Wiley & Sons, 2014. 326v.

EFRON, B.; HASTIE, T.; JOHNSTONE, I.; TIBSHIRANI, R.; ET AL. Least angle regression. **The Annals of statistics**, v.32, n.2, p.407–499, 2004.

FARAWAY, J. J. **Linear models with R**. Chapman and Hall/CRC, 2016.

FOX, J.; WEISBERG, S. **An R Companion to Applied Regression**. 2. ed. Sage, 2011.

GIOLO, S. R. **Introdução à análise de dados categóricos com aplicações**. Edgard Blücher Ltda., 2017.

GOTWALT, C.; XU, L.; HONG, Y.; MEEKER, W. Q. Applications of the Fractional-Random-Weight Bootstrap. **arXiv preprint arXiv:1808.08199**, 2018.

HOAGLIN, D. C.; WELSCH, R. E. The Hat Matrix in Regression and ANOVA. **The American Statistician**, v.32, n.1, p.17–22, 1978.

HOSMER, D. W.; JOVANOVIC, B.; LEMESHOW, S. Best Subsets Logistic Regression. **Biometrics**, v.45, n.4, p.1265–1270, 1989.

HOSMER JR, D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied logistic regression**. John Wiley & Sons, 2013. 398v.

JIANG, J. The fence methods. **Advances in Statistics**, v.2014, 2014.

JIANG, J.; NGUYEN, T.; RAO, J. S. A simplified adaptive fence procedure. **Statistics & Probability Letters**, v.79, n.5, p.625–629, 2009.

JIANG, J.; RAO, J. S.; GU, Z.; NGUYEN, T.; ET AL. Fence methods for mixed model selection. **The Annals of Statistics**, v.36, n.4, p.1669–1692, 2008.

JIN, Z.; YING, Z.; WEI, L.-J. A simple resampling method by perturbing the minimand. **Biometrika**, v.88, n.2, p.381–390, 2001.

KONISHI, S.; KITAGAWA, G. Generalised information criteria in model selection. **Biometrika**, v.83, n.4, p.875–890, 1996.

LITTELL, R. C.; MILLIKEN, G. A.; STROUP, W. W.; WOLFINGER, R. D.; SCHABENBERGER, O. **SAS for Mixed Models**. SAS Institute, 2006.

MCCULLAGH, P.; NELDER, J. A. **Generalized linear models**. CRC press, 1989. 37v.

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. **Introduction to linear regression analysis**. John Wiley & Sons, 2012. 821v.

MORAL, R. A.; HINDE, J.; DEMÉTRIO, C. G. B. Half-Normal Plots and Overdispersed Models in R: The hnp Package. **Journal of Statistical Software**, v.81, n.10, p.1–23, 2017.

MÜLLER, S.; WELSH, A. H. On model selection curves. **International Statistical Review**, v.78, n.2, p.240–256, 2010.

MURRAY, K.; HERITIER, S.; MÜLLER, S. Graphical tools for model selection in generalized linear models. **Statistics in medicine**, v.32, n.25, p.4438–4451, 2013.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. **Journal of the Royal Statistical Society. Series A (General)**, v.135, n.3, p.370–384, 1972.

PAULA, G. A. **Modelos de regressão: com apoio computacional**. IME-USP São Paulo, 2013.

PINHEIRO, J.; BATES, D. **Mixed-effects models in S and S-PLUS**. New York: Stringer-Verlag, 2000.

PRIGENZI, M. L. H.; TRINDADE, C. E.; RUGOLO, L. M. S. D. S.; SILVEIRA, L. V. Fatores de risco associados à mortalidade de recém-nascidos de muito baixo peso na cidade de Botucatu, São Paulo, no período 1995-2000. **Revista Brasileira de Saúde Materno Infantil**, p.93–101, 2008.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria, 2017.

SCHWARZ, G.; ET AL. Estimating the dimension of a model. **The annals of statistics**, v.6, n.2, p.461–464, 1978.

SEBER, G. A. F.; LEE, A. J. **Linear regression analysis**. John Wiley & Sons, 2012. 329v.

TARR, G.; MÜLLER, S.; WELSH, A. H. mplot: An R Package for Graphical Model Stability and Variable Selection Procedures. **Journal of Statistical Software**, v.83, n.9, p.1–28, 2018.

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

WASSERMAN, L.; ROEDER, K. High dimensional variable selection. **Annals of statistics**, v.37, p.2178, 2009.

WU, C. J.; HAMADA, M. S. **Experiments: planning, analysis, and optimization**. John Wiley & Sons, 2009. 552v.