Um estudo comparativo das técnicas de validação cruzada aplicadas a modelos mistos

João Paulo Zanola Cunha

Dissertação apresentada
ao
Instituto de Matemática e Estatística
da
Universidade de São Paulo
para
obtenção do título
de
Mestre em Ciências

Programa: Estatística

Orientador: Prof. Dr. Viviana Giampaoli

São Paulo, fevereiro de 2019

Um estudo comparativo das técnicas de validação cruzada aplicadas a modelos mistos

Esta versão da dissertação contém as correções e alterações sugeridas pela Comissão Julgadora durante a defesa da versão original do trabalho, realizada em 28/05/2019. Uma cópia da versão original está disponível no Instituto de Matemática e Estatística da Universidade de São Paulo.

Comissão Julgadora:

- Profª. Drª. Viviana Giampaoli (orientador) - IME-USP
- Profª. Drª. Nina Sumiko Tomita Hirata - IME-USP
- Prof. Dr. Jesús Enrique Garcia - IMECC-UNICAMP

Agradecimentos

Gostaria de agradecer a minha família pelo apoio e carinho durante toda a minha formação. Em especial aos meus pais, João Ferreira da Cunha e Maria Aparecida Zanola da Cunha, por terem sido a base sólida do meu desenvolvimento pessoal e profissional.

A minha orientadora, professora Viviana Giampaoli, agradeço pelas discussões sobre o desenvolvimento do trabalho, a paciência por entender minhas dificuldades na realização e pela parceria, estando sempre disposta a ajudar, sugerindo melhorias em pontos importantes que contribuíram para o sucesso desse trabalho.

Aos membros da banca, Profª Nina e Profº Jesús, por disponibilizarem o seu tempo para a leitura e compreensão do meu trabalho, assim como as contribuições e discussões trazidas durante a banca.

Gostaria de agradecer também ao apoio e incentivo dos meus chefes Felipe e Lidiane na empresa Porto Seguro, pois sem esse apoio não teria como trilhar o caminho do mestrado em conjunto com o tempo dedicado ao trabalho na empresa. Agradeço muito por terem permitido que essa jornada dupla acontecesse e pela paciência nos momentos que o excesso de trabalho me deixava desanimado e cansado. Além deles, agradeço a todos os meus colegas de trabalho pelas discussões que sempre agregaram aspectos positivos no desenvolvimento desse trabalho.

Aos meus amigos pela amizade, parceria, momentos de descontração e diversão, em especial aos meus amigos Gui, Henrique, Luan, Victor e Etienne. São pessoas que sempre pude contar, mesmo que não soubessem sobre o que esse trabalho se tratava, ouviam e davam conselhos para me ajudar. Ao Gui agradeço também por todos os anos de amizade, estando sempre ao meu lado. Ao Victor e Etienne agradeço também pelo dedicação que tiveram em ler esse texto, pelas sugestões de melhoria que propuseram e pelo convite de ser padrinho do casamento de vocês.

Agradeço também a Simone Harnik pelo apoio que sempre me deu no mestrado, mas principalmente por ter me ajudado num momento que quase pus tudo a perder por conta das minhas confusões e sem essa ajuda não teria como eu ter chegado aqui. Além dela, agradeço também a paciência e a compreensão das professoras Márcia e Airlane nesse mesmo momento.

Resumo

A avaliação da predição de um modelo por meio do cálculo do seu risco esperado é uma importante etapa no processo de escolha do um preditor eficiente para observações futuras. Porém, deve ser evitado nessa avaliação usar a mesma base em que foi criado o preditor, pois traz, no geral, estimativas abaixo do valor real do risco esperado daquele modelo. As técnicas de validação cruzada (*K-fold*, *Leave-One-Out*, *Hold-Out* e *Bootstrap*) são aconselhadas nesse caso, pois permitem a divisão de uma base em amostra de treino e validação, fazendo assim que a criação do preditor e a avaliação do seu risco sejam feitas em bases diferentes. Este trabalho apresenta uma revisão dessas técnicas e suas particularidades na estimação do risco esperado. Essas técnicas foram avaliadas em dois modelos mistos com distribuições Normal e Logístico e seus desempenhos comparados por meio de estudos de simulação. Por fim, as metodologias foram aplicadas em um conjunto de dados real.

Palavras-chave: Validação Cruzada, Modelos Mistos, Risco Esperado.

Abstract

The appraisal of model's prediction through the calculation of the expected risk is an important step on the process of the choice of an efficient predictor to future observations. However, in this evaluation it should be avoided to use the same data to calculate the predictor on which it was created, due to it brings, in general, estimates above the real expected risk value of the model. In this case, the cross-validation methods (*K-fold*, *Leave-One-Out*, *Hold-Out* and *Bootstrap*) are recommended because the partitioning of the data in training and validation samples allows the creation of the predictor and its risk evaluation on different data sets. This work presents a briefing of this methods and its particularities on the expected risk estimation. These methods were evaluated on two mixed models with Normal and Logistic distributions and their performances were compared through simulation cases. Lastly, those methods were applied on a real database.

Keywords: Cross-validation, Mixed Models, Expected risk.

Sumário

Capítulo 1

Introdução

1.1 Objetivo do trabalho

Uma medida importante da eficiência de um modelo estatístico é a sua capacidade de predição. Para essa avaliação, é necessário a coleta de informações de uma amostra da população de interesse, em que se observa uma variável objetivo, que chamaremos de Y, e também uma coleção de outras variáveis $\mathbf{X} = (X_1, X_2, ..., X_p)$, em que p é o número de variáveis coletadas, sendo que elas podem ou não ter relação com Y. A partir dessa amostra, um preditor de Y será criado através de uma função $\hat{f}(\mathbf{X})$ das variáveis $\mathbf{X}$, em que $\hat{f}(\mathbf{X})$ deve ser uma boa função que prediz Y usando $\mathbf{X}$ (James *et al.*, 2013).

Diversas técnicas foram propostas para a criação desses preditores, cada vez mais eficazes em conseguir extrair ao máximo as informações que um banco de dados possui. Em alguns casos, um preditor proposto pode ser tão preciso que, quando se compara o valor da informação observada Y e do preditor $\hat{f}(\mathbf{X})$, a diferença será quase nula. Essa comparação entre o observado e o predito é o chamado risco esperado (Borra e Ciaccio, 2010), ou seja, ela é uma medida que penaliza o uso de $\hat{f}(\mathbf{X})$ para predizer Y.

É sabido, por exemplo, que usar a mesma amostra para criar um preditor e medir o risco esperado dele produz uma estimativa otimista desse risco, ou seja, como a estrutura do preditor foi criado com essa amostra, é esperado que ele tenha uma boa perfomance em predizer a variável resposta, criando a impressão que o erro na capacidade de predição seja baixo quando na verdade ele pode ser muito maior. Em algumas técnicas de criação de preditores pode ocorrer superajuste, que indica que um preditor, embora demonstre uma grande eficiência em predizer os valores Y da amostra, não consegue ser eficiente quando o usamos para a predição de um grupo novo de informações.

Por isso, a validação do erro entre Y e seu preditor $\hat{f}(\mathbf{X})$ deve ser feita em observações que não foram usadas para criar o preditor, ou seja, submete-se novas observações ao modelo criado e avalia-se sua performance na predição da variável resposta. Entretanto, dispor de novas observações não é tão simples, pois muitos estudos possuem limitações técnicas e de custos, o que inviabiliza um grande volume de dados para se criar o preditor e depois avaliá-lo.

James *et al.* (2013) sugere que, para contornar esse problema e ter uma boa avaliação do erro de predição de um modelo, é necessário dividir a amostra coletada em duas partes. A primeira é conhecida como amostra de treino (*original sample*, *construction sample*), sendo usada para criar o preditor $\hat{f}(\mathbf{X})$ e a segunda parte é chamada de amostra de validação (*check sample*, *validation sample*), que é usada para avaliar a capacidade de predição do modelo, submetendo-a ao preditor criado com a primeira base e medindo o risco esperado em se predizer os valores de Y dessa segunda base. Alguns autores como Hastie *et al.* (2008) consideram uma divisão em três partes: amostra de treino, amostra de teste e amostra de validação. Nesse caso, a amostra de treino é usada pra criar o preditor, a amostra de validação é usada para estimar o risco esperado durante o processo de escolha da forma do preditor, escolhendo-se aquela com o menor risco esperado, e a amostra de teste é usada

para estimar o risco esperado após o preditor final ter sido escolhido para avaliar, por exemplo, se não houve superajuste. A divisão em três bases é mais apropriada quando estamos definindo qual a melhor forma do preditor. Como o interesse desse trabalho é apenas avaliar a eficiência das técnicas de validação, usaremos a divisão em duas bases descrita por James *et al.* (2013). O uso da divisão da base para avaliar a predição é parte da metodologia conhecida como validação cruzada (Stone, 1974).

A técnica de validação cruzada pode ser feita de diferentes formas, sendo que os métodos mais usados são o Hold-Out (Devroye e Wagner, 1979), K-fold (Burman, 1989), Leave-one-out e Bootstrap (Efron, 1983). A descrição desses métodos será detalhada no Capítulo 3 desse trabalho. Cada uma dessas técnicas tem seus prós e contras e é de interesse compará-las. Diversos trabalhos foram feitos nesse sentido, como por exemplo, Borra e Ciaccio (2010), Kim (2009), Rodriguez *et al.* (2013), entre outros. Esses trabalhos, geralmente avaliam modelos de classificação e modelos de regressão nos quais se tem apenas uma única fonte de variabilidade.

Essa dissertação traz como contribuição, então, a avaliação das técnicas de validação cruzada em modelos de regressão em que há uma fonte de variabilide extra, em modelos conhecidos como modelos mistos (Demidenko, 2013), que serão apresentados no Capítulo 4 desse texto.

1.2 Organização do texto

No Capítulo 2 desse trabalho é apresentado como se avalia o risco esperado de um modelo, apresentando os conceitos de função de perda e erro quadrático médio (EQM).

Diversas técnicas de validação cruzada como Hold-Out, K-fold, Leave-One-Out, Bootstrap e suas variações são apresentadas, com suas características, no Capítulo 3.

No Capítulo 4 é feita uma revisão sobre modelos de regressão e modelos mistos, tanto no caso em que a variável resposta tem a distribuição normal, quanto no caso mais geral em que se usam distribuições pertencentes à família exponencial, os modelos lineares generalizados.

Os estudos de simulação para avaliar a eficiência das técnicas de validação cruzada são apresentados no Capítulo 5 fazendo um comparativo do desempenho delas nos modelos mistos e nos modelos que possuem apenas efeitos fixos.

No Capítulo 6 aplicamos as técnicas apresentadas num banco de dados real.

As considerações finais sobre o estudo realizado são apresentadas no Capítulo 7.

Capítulo 2

Medidas de erro de predição

Um dos grandes interesses da Estatística é conseguir encontrar relações entre variáveis, identificando padrões para poder realizar predições de uma variável de interesse. Se é de interesse vender um novo produto, por exemplo, seria intessante descobrir que informações do público alvo (sexo, idade, renda média, região onde mora) podem ser relevantes para alavancar a venda desse produto ou que meio de comunicação (TV, rádio, sites de notícia, redes sociais) seria mais interessante para atingir esse público. Da mesma forma, um estudo farmacêutico pode querer avaliar a performance de uma nova droga no tratamento de uma doença, comparando-a com um tratamento que já exista, com o propósito de definir se essa nova droga é mais eficiente do que a já existente. Ou seja, temos uma informação ou objetivo de interesse e dispomos de um conjunto de outras informações que podem ser relevantes ou não para atingir esse objetivo. Por meio de modelos estatísticos e da construção de um banco de dados, podemos estabelecer essas relações por meio de uma formulação matemática.

Nessa formulação, essas informações são usadas como variáveis, sendo a informação de interesse chamada de variável dependente, identificada como Y, e as outras informações são chamadas de variáveis independentes (ou covariáveis) e serão representadas por $\mathbf{X} = (X_1, X_2, ..., X_p)$, em que p é a quantidade de variáveis independentes que dispomos, X_1 é a primeira variável, X_2 a segunda e assim por diante, até a p-ésima variável, X_p.

Assim, podemos estabelecer a relação

$$Y = f(\mathbf{X}) + \epsilon \tag{2.1}$$

em que $f(.)$ é uma função fixa das covariáveis $\mathbf{X}$, ou seja, a função $f(\mathbf{X})$ tem uma forma única que estabelece quanto a mudança nas variáveis $\mathbf{X}$ influenciam no valor da variável Y e o termo ϵ é um erro aleatório. Porém, apesar de ser fixa, não sabemos a priori qual a forma de $f(\mathbf{X})$. A coleta de uma amostra dessa população pode nos ajudar a estabelecer uma forma aproximada dessa função.

Assim coletamos n observações para compor a amostra d, tendo um conjunto de informações de interesse $\mathbf{Y} = (y_1, y_2, ..., y_n)'$ e de covariáveis $\mathbf{X}_i = (x_{i1}, x_{i2}, ..., x_{ip})$, em que $i = 1, 2, ...n$ é a quantidade de observações na amostra e p é a quantidade de covariáveis, criando a matriz $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n)'$ de dimensão n x p. Com essa amostra e alguma técnica de estimação, como, por exemplo, mínimos quadrados ou máxima verossimilhança, criamos a função $\hat{f}(\mathbf{X})$, que deve ter sua forma parecida com a forma de $f(\mathbf{X})$.

O termo ϵ em (2.1) indica que, mesmo que seja possível encontrar a função verdadeira $f(\mathbf{X})$, a relação entre Y e $f(\mathbf{X})$ não é perfeita, ainda existindo este termo, chamado de erro do modelo, que é uma variável aleatória, com média zero e variância σ^2_ϵ.

Então, dado que conseguimos estimar $f(\mathbf{X})$ por uma função $\hat{f}(\mathbf{X})$ e ϵ tem média zero, o valor

esperado de $\mathbf{Y}$ é dado por

$$\hat{\mathbf{Y}} = E(\mathbf{Y}) = E(f(\mathbf{X}) + \boldsymbol{\epsilon}) = \hat{f}(\mathbf{X}). \tag{2.2}$$

Se Y for uma variável contínua, o preditor $\hat{f}(\mathbf{X})$ será uma função que irá retornar um valor na reta real. Podemos usar nesse caso técnicas de regressão linear (Kutner *et al.*, 2004), regressão não-linear (Bates e Watts, 2007), entre outras. Para o caso de Y ser uma variável categórica, $\hat{f}(\mathbf{X})$ deverá retornar a categoria de Y onde aquele indíviduo mais se enquadra. Para esse caso, podemos usar métodos como análise discriminante (Huberty, 1975), vizinhos mais próximos (Cover e Hart, 1967) ou árvores de classificação (Breiman *et al.*, 1984).

Segundo James *et al.* (2013), a acurácia de $\hat{\mathbf{Y}}$ como uma predição dos valores esperados de $\mathbf{Y}$ depende de duas quantidades, que são chamados de erro redutível e erro irredutível.

O erro redutível é a diferença entre o valor estimado $\hat{f}(\mathbf{X})$ e o valor real de $f(\mathbf{X})$. Chamamos de redutível porque é possível diminuir esse erro se usarmos técnicas de estimação apropriadas, porém a maior dificuldade é encontrar a forma correta de $f(\mathbf{X})$, dado que não a conhecemos e usamos apenas um amostra para estimá-la.

Já o erro irredutível está associado ao valor de ϵ, representando a variabilidade que não é explicada por $\mathbf{X}$, por exemplo, no caso de existirem outras covariáveis $\mathbf{W} = (W_1, ..., W_k)$ que não foram observadas e que podem explicar Y, e portanto não estão contidas no valor de $f(\mathbf{X})$.

Outro motivo desse erro existir são variações no momento de se medir as covariáveis X_p. Por exemplo, suponha que Y seja a variável que indica hipertensão, onde, Y = 1 se o paciente tem hipertensão e 0 caso não tenha. Suponha que X_1 seja a pressão arterial do paciente e queremos encontrar a relação entre Y e X_1. Porém a pressão arterial é uma variável que oscila no tempo e, no momento em que ocorre a medida, podemos coletar uma amostra que tenha uma pressão menor do que o usual. Esse problema poderia ser resolvido se coletarmos diversas amostras de X_1 e usarmos por exemplo a média dos valores como o valor coletado de X_1, mas em muitas situações, coletar mais amostras indica um maior custo para a análise, o que pode inviabilizar o estudo. Também pode ocorrer de usarmos um equipamento de medição que esteja descalibrado, ou seja, teremos uma medição não tão precisa do valor de X_1. Esses erros são não observáveis, pois não temos como medi-los, e por isso eles acabam sendo incorporados no valor de ϵ.

Se assumirmos $\hat{Y} = \hat{f}(\mathbf{X})$ e que $\hat{f}(\mathbf{X})$ e $\mathbf{X}$ são fixos, podemos decompor o erro esperado do modelo nessas duas quantidades, no caso do erro quadrático, como

$$E(\mathbf{Y} - \hat{\mathbf{Y}})^2 = E(f(\mathbf{X}) + \boldsymbol{\epsilon} - \hat{f}(\mathbf{X}))^2 =$$

$$VAR(f(\mathbf{X}) + \boldsymbol{\epsilon} - \hat{f}(\mathbf{X})) + [E(f(\mathbf{X}) + \boldsymbol{\epsilon} - \hat{f}(\mathbf{X})]^2 =$$

$$VAR(\boldsymbol{\epsilon}) + [f(\mathbf{X}) - \hat{f}(\mathbf{X})]^2,$$

em que $E(\mathbf{Y} - \hat{\mathbf{Y}})^2$ representa o valor esperado do quadrado da diferença entre o valor predito e o verdadeiro valor do vetor de observações $\mathbf{Y}$, $VAR(\boldsymbol{\epsilon})$ é o erro irredutível e a quantidade $[f(\mathbf{X}) - \hat{f}(\mathbf{X})]^2$ é o erro redutível. A melhor predição seria aquela onde seja possível termos o menor valor do erro redutível.

2.1 Função de perda e risco esperado

Para avaliar a assertividade da predição de $\hat{f}(\mathbf{X})$ sobre o valor de Y, podemos criar uma função L que penalize $\hat{f}(\mathbf{X})$ quando este não está próxima do valor real. A função $L(Y, \hat{f}(\mathbf{X}))$, chamada de função de perda (ou custo) quantifica essa penalização, retornando um valor em $\mathbb{R}$, sendo que quanto menor for esse valor, mais aceitável será o risco de se usar $\hat{f}(\mathbf{X})$ como preditor do valor esperado de Y.

A escolha da função de perda depende do objetivo e do custo que se aceita pagar por um erro.

No caso em que Y é uma variável contínua e o nosso objetivo é encontrar uma função de regressão $f(\mathbf{X})$ que seja preditora de Y, podemos escolher, segundo Hastie *et al.* (2008), as funções:

$$L(Y, \hat{f}(\mathbf{X})) = (Y - \hat{f}(\mathbf{X}))^2, \tag{2.3}$$

$$L(Y, \hat{f}(\mathbf{X})) = |Y - \hat{f}(\mathbf{X})|. \tag{2.4}$$

A perda (2.3) é conhecida como perda quadrática e a perda (2.4) é chamada de perda absoluta. Molinaro *et al.* (2005) diz que (2.3) é mais utilizada na literatura do que (2.4) e será a função de perda que usaremos nesse trabalho.

Para o caso de Y ser uma variável categórica e $f(\mathbf{X})$ uma função de classificação, Hastie *et al.* (2008) diz que a escolha mais comum de função de perda é conhecida como perda 0-1, dada por:

$$L(Y, \hat{f}(\mathbf{X})) = I(Y \neq \hat{f}(\mathbf{X})). \tag{2.5}$$

No caso de assumirmos que Y tem uma distribuição logística, podemos usar como função de perda o método da Entropia Cruzada (Murphy, 2012). O método tem esse nome porque avalia a divergência entre duas distribuições de probabilidade. Nesse caso, iremos medir a divergência entre a distribuição de Y e a distribuição de $\hat{f}(\mathbf{X})$, através da fórmula:

$$L(Y, \hat{f}(\mathbf{X})) = -(Y log\{\hat{f}(\mathbf{X})\} + (1 - Y) log\{1 - \hat{f}(\mathbf{X})\}) \tag{2.6}$$

Escolhida a função de perda, podemos criar uma medida que leve em conta a perda em todos os pontos possíveis da função. Essa medida, chamada de risco esperado (Borra e Ciaccio, 2010), leva em conta o preditor $\hat{f}(\mathbf{X})$ e a função $L(Y, \hat{f}(\mathbf{X}))$ e pode ser definida como

$$Err = R(\hat{f}(\mathbf{X}), Y) = E_{\mathbf{X}} E_{Y|\mathbf{X}}[L(Y, \hat{f}(\mathbf{X}))|\hat{f}(\mathbf{X}), \mathbf{X}], \tag{2.7}$$

ou seja, calculamos o erro esperado sob todos os possíveis valores do vetor de covariáveis $\mathbf{X}$ e Y para uma $\hat{f}(\mathbf{X})$ fixa. Essa medida é sempre positiva, embora não seja necessariamente finita, pois

a distribuição das variáveis pode não ter valor médio.

2.2 Estimadores do risco esperado

O risco esperado dado por (2.7) depende que saibamos a distribuição conjunta das covariáveis X. Como isso é bem complicado de se obter, ainda mais quando houverem muitas covariáveis com distribuições diferentes, precisamos criar um estimador $e\hat{r}r$ que tenha boas propriedades para que seu valor seja o mais próximo possível do valor real dado por (2.7).

O estimador mais simples que foi proposto (Borra e Ciaccio, 2010) é chamado de erro aparente, dado por

$$errapt = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}(\mathbf{x}_i)) \tag{2.8}$$

sendo n o tamanho da amostra e $i = 1, ..., n$, isto é, usamos a média da função de perda das observações de uma amostra d.

Porém esse valor não é o mais apropriado, pois como o preditor foi criado com essa amostra, (2.8) tende a subestimar o valor de (2.7) (Efron e Tibshirani, 1995).

A validação cruzada pode nos ajudar a contornar esse problema, propondo uma partição da nossa amostra d, de tamanho n, em duas partes complementares sendo a primeira a amostra de treino d_t, de tamanho t, e a segunda a amostra de validação d_v, de tamanho v, com $n = t + v$. Usa-se a amostra d_t para criar o preditor $\hat{f}_t(\mathbf{X})$ e avalia-se o erro esperado submetendo esse preditor à amostra d_v, ou seja, o estimador de erro será definido como

$$err_v = \frac{1}{v}\sum_{i=1}^{v} L(y_i, \hat{f}_t(\mathbf{x}_i)). \tag{2.9}$$

Fazendo isso, temos uma estimativa mais apropriada do risco esperado, pois as observações de d_v não foram usadas para criar o preditor $\hat{f}_t(\mathbf{X})$, ou seja, o preditor é avaliado num conjunto de dados novo.

No caso de modelos de regressão, uma medida mais comumente usada para medir a assertividade do preditor é o erro quadrático médio (EQM), definido como

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\mathbf{x}_i))^2. \tag{2.10}$$

Note que, nesse caso, estamos usando para estimar o erro a função de perda quadrática definida em (2.3). O EQM é um caso do chamado erro aparente dado em (2.8), ou seja, EQM é uma estimativa otimista do risco esperado. Devemos então ter um EQM da base de validação, dado por:

$$EQM_v = \frac{1}{v}\sum_{i=1}^{v}(y_i - \hat{f}_t(\mathbf{x}_i))^2. \tag{2.11}$$

Nesse caso, o EQM_v é um caso particular de (2.9) quando a função de perda usada é a quadrática.

2.3 Comparação entre estimadores

Os estimadores que possuem a estrutura dada em (2.8) ou (2.9), que chamaremos genericamente nessa seção por $e\hat{r}r$, devem ter boas propriedades para que estimem com precisão o valor de (2.7), ou seja, é esperado que $E(e\hat{r}r) = Err$.

Para avaliar se isso ocorre, podemos usar também uma medida de risco esperado em se estimar Err usando o estimador $e\hat{r}r$. Usando a perda quadrática, podemos definir o erro quadrático médio do estimador $EQM(e\hat{r}r)$ como

$$R(e\hat{r}r, Err) = EQM(e\hat{r}r) = E[(e\hat{r}r - Err)^2].$$

Portanto, o melhor estimador será aquele que minimiza o valor de $EQM(e\hat{r}r)$. Podemos decompor essa medida em dois valores

$$E[(e\hat{r}r - Err)^2] = E[(e\hat{r}r - E[e\hat{r}r] + E[e\hat{r}r] - Err)^2] =$$

$$E[(e\hat{r}r - E[e\hat{r}r])^2 + 2(e\hat{r}r - E[e\hat{r}r])(E[e\hat{r}r] - Err) + (E[e\hat{r}r] - Err)^2] =$$

$$E[(e\hat{r}r - E[e\hat{r}r])^2] + 2E[e\hat{r}r] - Err])E[e\hat{r}r - E[e\hat{r}r]] + E[(E[e\hat{r}r] - Err)^2] =$$

$$Var[e\hat{r}r] + (E[e\hat{r}r] - Err)^2,$$

pois $E[e\hat{r}r - E[e\hat{r}r]] = 0$ e $E[(e\hat{r}r - Err)^2] = Var[e\hat{r}r]$. O resultado encontrado possui dois termos: $Var[e\hat{r}r]$ é a variância do estimador e $(E[e\hat{r}r] - Err)^2$ é o quadrado do viés do estimador. Ou seja, um bom estimador deve ser aquele que tem pouca variância e um viés baixo. Podemos ver a diferença entre esses dois conceitos na **Figura 2.1**, onde o centro do alvo seria o valor real de Err.

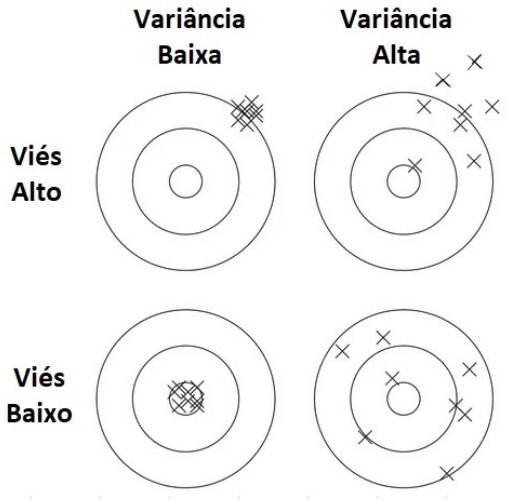

Figura 2.1: *Ilustração do conceito de viés e variância*

A variância do estimador $\hat{err}$ mede a dispersão dele ao redor do valor de Err. Segundo James *et al.* (2013), a escolha da função $\hat{f}(\mathbf{X})$ pode influenciar na variância do estimador, pois caso seja escolhida uma função que superajusta os dados, a forma dessa função pode ser completamente diferente caso tivéssemos outra amostra para criá-la. Wong (2015) indica também que usar uma técnica de partição da amostra que deixe poucas observações na amostra de teste aumenta a variância, pois, com poucas observações, $\hat{err}$ não converge apropriadamente para Err.

Por outro lado, o viés do estimador $\hat{err}$ mede sua precisão em acertar o valor de Err. A escolha da complexidade do preditor $\hat{f}(\mathbf{X})$ influencia no viés porque a função $f(\mathbf{X})$ pode ter uma complexidade diferente da escolhida. Por exemplo, se nossa função real for não linear e usarmos uma função linear para estimá-la, teremos um viés alto na precisão do nosso modelo. O mesmo pode acontecer se a função for linear e usarmos uma função mais complexa para estimá-la. Outro fator que também influencia no viés é a técnica de partição da amostra escolhida. Uma amostra de treino pequena pode não capturar toda a informação necessária para se ter um bom estimador de $f(\mathbf{X})$, criando uma versão distorcida do modelo.

A escolha da técnica de partição deve levar em conta, então, o quanto ela produz de viés e de variância. No geral, o que acontece é que uma técnica tende a ter um viés alto e uma variância baixa ou um viés baixo e uma variância alta, muitas vezes relacionado ao tamanho da amostra de treino e de validação que foram criados. Deve-se então tentar equilibrar as duas medidas, encontrando a melhor técnica que reduza os erros.

No próximo capítulo apresentaremos as formas mais conhecidas de se fazer essa partição e como a escolha da técnica modifica o estimador dado por (2.9).

Capítulo 3

Técnicas de validação de modelos

Validar um modelo é uma etapa importante para poder ter uma avaliação da capacidade de predição do mesmo. Assim, ao longo do tempo, foram propostas diferentes técnicas de validação de modelos, como as descritas brevemente a seguir.

3.1 Hold-out

O método *hold-out* (Devroye e Wagner, 1979), também conhecido como validação simples, propõe que a amostra d seja dividida em duas partes, usando uma proporção p dela como amostra de validação. Formalmente, dado um conjunto de dados d, separa-se uma proporção p dos dados, criando-se a amostra de treino d_t onde $t = n * (1 - p)$ e a amostra de validação d_v, de tamanho $v = n * p$. O estimador *hold-out* de (2.7) é dado por

$$hop^{-1} = \frac{1}{v} \sum_{i=1}^{v} L(y_i, \hat{f}_t(\mathbf{x}_i)), \tag{3.1}$$

em que $\hat{f}_t(\mathbf{X})$ é preditor criado com a amostra de treino d_t e a função de perda é avaliada em todos os pontos $(y_i, \mathbf{x}_i)$ da amostra de validação d_v.

Segundo Kohavi (1995), o estimador pelo método de *hold-out* é pessimista porque usa apenas uma parte dos dados como preditor do modelo. Quanto mais observações deixarmos para base de teste, maior será o viés do modelo. No caso de um conjunto de dados pequenos, por exemplo, deixar de usar uma parte pode causar uma grande distorção no modelo. Deixar poucas observações para a base de treino, por outro lado, aumenta a variância do estimador.

Nesse trabalho, usaremos os estimadores $ho3$ ($p = \frac{1}{3}$) e $ho10$ ($p = \frac{1}{10}$). O estimador $ho3$ é o mais usado, embora ele possa produzir um viés maior na estimativa, uma vez que o preditor $\hat{f}_t(\mathbf{X})$ usa apenas $\frac{2}{3}$ das observações como amostra de treino. Já $ho10$ diminui o viés do preditor, porém deve aumentar a variância, pois a amostra de teste é pequena, principalmente quando temos poucas observações.

Para contornar tais problemas, podemos pensar numa variação do método *hold-out* simplesmente repetindo-o R vezes (*repetead hold-out*). Assim, a cada repetição r, teremos uma base de treino d_{tr}, de tamanho $t = n * (1 - p)$ e uma base de validação d_{vr}, de tamanho $v = n * p$, diferentes. A estimativa de (2.7) nesse caso será dada por

$$rhop^{-1} = \frac{1}{R} \sum_{r=1}^{R} \frac{1}{v} \sum_{i=1}^{v} L(y_{ir}, \hat{f}_{tr}(x_{ir})) \tag{3.2}$$

em que $\hat{f}_{tr}(\mathbf{X})$ é o preditor criado com a amostra de treino da repetição r e avaliamos esse preditor nas observações $(y_{ir}, \mathbf{x}_{ir})$ da base de validação correspondente d_{vr}, $r = 1, 2, ..., R$.

Um problema apontado por Kohavi (1995) é não existir independência entre as repetições e como as repetições são aleatórias, uma parte dos dados pode estar sendo subrepresentada nessas repetições. A vantagem desse método é que ele não depende de uma única partição da base, diminuindo assim o variância do estimador (Rodriguez *et al.*, 2013).

3.2 K-fold

No método *K-fold* (Burman, 1989), a amostra d é dividida em K partes (d_1, d_2, ..., d_K) de tamanho parecido m_k, em que $\sum_{k=1}^{K} m_k = n$. O processo terá K iterações onde, em cada iteração, a amostra de validação será dada por d_k, com $k = 1, 2, ..., K$, e a amostra de treino para a criação do preditor será o conjunto das outras $K-1$ partes, ou seja, $d_{(-k)} = \{d_1, d_2, ..., d_{k-1}, d_{k+1}, ..., d_K\}$. Assim, ao final dos K passos, teríamos usado todos os dados tanto na parte de treino, quanto na parte de validação. O estimador de (2.7) pelo método *K-fold* é dado por

$$kfK = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{m_k}\sum_{i=1}^{m_k} L(y_{ik}, \hat{f}_{(-k)}(\mathbf{x}_{ik})), \tag{3.3}$$

em que o preditor $\hat{f}_{(-k)}(\mathbf{X})$ é criado com a amostra de treino $d_{(-k)}$ e esse preditor é avaliado nas observações da amostra de teste d_k para $k = 1, 2, ...K$.

Segundo Borra e Ciaccio (2010), o viés do método *K-fold* diminui quanto maior o valor do K. Porém, um K muito elevado acaba aumentando o custo computacional da técnica, além de implicar uma amostra de teste pequena, o que aumenta a variância. Na literatura, se discute qual valor de K seria o ideal, sendo como opções mais usuais os valores K= 2, 5 ou 10. Kohavi (1995) traz essa discussão e afirma que, em seus estudos, $K = 10$ tem um melhor desempenho. Borra e Ciaccio (2010) e Kim (2009) também usam em seus trabalhos o valor de K = 10. Como iremos estudar algumas variações do método K-fold que serão descritas abaixo, avaliar valores de K resultaria numa grande quantidade de estimadores para serem avaliados. Sendo $K = 10$ o valor mais frequentemente usado nos trabalhos anteriores, adotaremos, então, esse valor de K.

Um outro problema em relação a esse método apontado por Breiman (1996) é que as amostras de treino ($d_{(-1)}$, $d_{(-2)}$, ..., $d_{(-K)}$) não são independentes entre si, implicando numa variância que pode ser grande.

Burman (1989) mostrou em seu artigo que $E(kfK - Err) \approx s_0(K-1)^{-1}n^{-1}$, em que $E(kfK - Err)$ é o viés médio que ocorre quando usamos kfK para estimar Err e s_0 é uma constante que não depende de K e n, dependendo somente da função de perda $L(.)$ escolhida e da função $f(\mathbf{X})$. Ou seja, Burman mostra que, em média, kfK possui um viés. Se tivermos $K = n$ esse valor é bem pequeno e não afeta tanto a estimativa, mas quando K é pequeno, o viés médio não é necessariamente pequeno. Além disso, s_0 é da mesma ordem da quantidade de parâmetros do modelo, ou seja, se o número de parâmetros for alto, o viés de (3.3) se torna maior. Por isso, Burman (1989) propõe em seu artigo a seguinte correção:

$$bkfK = kfK + errapt - kf^{+}, \tag{3.4}$$

em que o termo kfK é dado por (3.3), $errapt$ é dado por (2.8), sendo calculada com a amostra completa d e kf^{+} é calculado em K iterações, sendo que, a cada iteração, retira-se da amostra de treino o conjunto d_k, ou seja, $d_{(-k)} = \{d_1, d_2, ..., d_{k-1}, d_{k+1}, ..., d_K\}$, mas na amostra de validação usa-se a amostra toda d em todas as iterações, ou seja, o valor de kf^{+} é dado por

$$kf^{+} = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{n}\sum_{i=1}^{n} L(y_i, \hat{f}_{(-k)}(x_i)),$$

em que $\hat{f}_{(-k)}(\mathbf{X})$ é preditor criado em cada iteração k do processo, usando a amostra de treino $d_{(-k)}$ correspondente a cada etapa.

Burman (1989) demonstra que $E(bkfK - Err) \approx s_1(K-1)^{-1}n^{-2}$, em que s_1 também é uma constante que depende apenas da função de perda $L(.)$ e da função $f(\mathbf{X})$ e possui uma ordem de grandeza semelhante a constante s_0 do viés de kfK. Como as duas constantes são próximas, o fator n^{-2} pode fazer o viés de $bkfK$ ser menor do que do estimador kfK, principalmente quando n cresce.

Podemos também pensar em fazer método *K-fold* com repetições (*repeated K-fold*), ou seja, na primeira repetição, dividimos a amostra em K partes aleatoriamente e calculamos o valor de kfK_1 usando (3.3). Na segunda repetição, dividimos a amostra em outras K partes e calculamos kfK_2 e assim sucessivamente, repetindo o procedimento R vezes. No final, fazemos a média dos valores encontrados em cada repetição, ou seja,

$$rkfK = \frac{1}{R}\sum_{r=1}^{R} kfK_r. \tag{3.5}$$

$rkfK$ tende a ter uma variância menor do que kfK. A repetição dilui o problema de alguma parcela viesar o valor da estimativa, porém temos o problema que, ao se repetir o processo, as amostras de validação deixam de ser independentes (Wong, 2015).

A correção de Burman também pode ser aplicado para cada uma das R repetições, tendo o estimador

$$rbkfK = \frac{1}{R}\sum_{r=1}^{R} bkfK_r. \tag{3.6}$$

3.3 Leave-one-out

O método *Leave-one-out* (*loo*) é um caso especial do *K-fold* para $K = n$, ou seja, a cada iteração, a amostra de validação será correspondente a uma observação $d_k = \{(y_k, \mathbf{x}_k)\}$, $k = 1, 2, ..., n$, e a amostra de treino para criar o preditor é feito com as outras $n-1$ observações, onde usaremos a notação $d_{(-k)}$, ou seja, é o conjunto de todas as observações exceto a k-ésima. A estimativa do risco esperado pode ser definida como

$$loo = \frac{1}{n}\sum_{k=1}^{n} L(y_k, \hat{f}_{(-k)}(\mathbf{x}_k)) \tag{3.7}$$

onde $\hat{f}_{(-k)}(\mathbf{X})$ é o preditor criado em cada iteração k do processo, retirando-se a observação $(y_k, \mathbf{x}_k)$ da amostra de treino.

Segundo Borra e Ciaccio (2010), *loo* é um estimador quase não viesado do erro, pois a amostra de treino é quase a base toda, principalmente quando n é grande. Porém, *loo* tem alta variabilidade,

pois as parcelas de cada etapa possuem apenas uma observação. Além disso, por treinar n vezes o modelo, o custo computacional desse método pode ser elevado se tivermos uma amostra muito grande (Kim, 2009).

3.4 Bootstrap

A técnica de *bootstrap* pode ser usada para estimar o erro de um modelo da seguinte forma: retira-se da amostra d aleatoriamente e com reposição K amostras $(B_1, B_2, \cdots, B_K)$ de tamanho m. Assim se a amostra de treino é B_k, a amostra de validação será $B_{(-k)} = \{(x_{ik}, y_{ik}) \notin B_k\}$, de tamanho m_k, que é composta por todas as observações que não estão em B_k. É importante salientar que a amostra B_k é construída com reposição, ou seja, é possível que a amostra de treino tenha observações repetidas. Isso faz com que cada amostra de validação tenha um tamanho m_k diferente e que $m + m_k \neq n$. A estimativa de (2.7), nesse caso, será dado por

$$bt = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{m_k}\sum_{i=1}^{m_k} L(y_{ik}, \hat{f}_k(\mathbf{x}_{ik})), \tag{3.8}$$

em que $\hat{f}_k(\mathbf{X})$ é o preditor criado em cada iteração do processo de *bootstrap*, com a amostra de treino B_k correspondente e a validação é feita em todos os pontos que não pertencem a B_k .

Segundo Kim (2009), a estimativa do erro via *bootstrap* tem uma boa performance em amostras pequenas porque ela tem uma menor variância, porém demanda um maior custo computacional.

O estimador (3.8) superestima o valor do (2.7), pois, dado que o tamanho da base é igual a n, a seleção do método inclui apenas $0.632n$ casos na amostra de treino. Podemos mostrar isso calculando a probabilidade de uma observação i estar numa amostra B da seguinte forma:

$$P(i \in B) = 1 - (1 - \frac{1}{n})^n \approx 1 - e^{-1} = 0.632.$$

Ou seja, uma boa parte dos dados fica na base de teste ($n - 0.632n = 0.368n$), produzindo um viés na estimativa. Para contornar esse viés, então, Efron (1983) propôs uma versão ponderada do estimador, dada por

$$bt632 = 0.632 * bt + 0.368 * errapt, \tag{3.9}$$

em que bt é o estimador dado em (3.8) e $errapt$ é calculado por (2.8) sob a amostra completa d. O valor do ponderador de 0.632 usa como referência a probabilidade de uma observação i estar numa amostra B. A ideia de Efron (1983) é equilibrar um estimador superviesado (bt) com um estimador subviesado ($errapt$) trazendo a estimativa para mais próxima do valor esperado.

No artigo de Efron (1983), $bt632$ tem uma alta performance, mas no caso de superajuste do modelo, como pode acontecer no método dos vizinhos mais próximos, $errapt = 0$ e (3.9) ficará sub-viesado.

Efron e Tibshirani (1995) propõem, então, uma nova modificação, conhecida como estimador de *bootstrap 0.632+*, que dará um maior peso para bt nos casos de superajuste.

Para chegar nesse valor do ponderador, eles primeiro usam uma medida chamada de erro não informativo, γ, que corresponde ao valor do erro esperado quando Y e X são independentes. Podemos estimar esse valor através de gma, que é obtida considerando a perda $L(.)$ em todos os pares (x_i, y_j), $i, j = 1, 2, ..., n$, e a função $\hat{f}(\mathbf{X})$ é criada com a amostra d completa, ou seja,

$$gma = \frac{1}{n^2} \sum_{i,j} L(y_i, \hat{f}(x_j))$$

$$(i, j = 1, 2, \cdots, n),$$

em que o denominador n^2 indica o total de combinações possíveis entre os valores de x e y.

Após, Efron e Tibshirani (1995) criam uma taxa relativa de superajuste, dada por:

$$\hat{R} = \frac{bt - errapt}{gma - errapt},$$

que deve ser um número entre 0 e 1, onde quando mais perto de 1, maior o superajuste dos dados.

Então, o estimador *bootstrap 0.632+* é dado por

$$bt632+ = \hat{w} * bt + (1 - \hat{w}) * errapt \tag{3.10}$$

em que

$$\hat{w} = \frac{0.632}{1 - 0.368 * \hat{R}},$$

bt é o estimador bootstrap dado em (3.8) e $errapt$ é calculado por (2.8) sob a amostra completa d. Ou seja, se $\hat{R} = 1$, estaremos num caso de superajuste, $\hat{w} = 1$ e o estimador $bt632+$ será igual ao bt. Se $\hat{R}$ for igual a zero, $\hat{w} = 0.632$ e o estimador $bt632+$ será igual ao $bt632$.

Capítulo 4

Modelos Mistos

Os modelos lineares foram umas das primeiras técnicas para estabelecer relações entre variáveis e fazer predições da variável de interesse. Embora hoje tenham surgido diversas novas técnicas computacionais para criar preditores eficientes, os modelos lineares ainda são amplamente usados na literatura e são muito úteis para a resolução de grande parte dos problemas de predição.

Neste capítulo apresentaremos algumas noções dos modelos lineares generalizados (GLM) e modelos lineares generalizados mistos (GLMM) e como utilizá-los para predição da variável de interesse para os leitores menos familiarizados com a notação do assunto.

4.1 Modelos Lineares

Os modelos lineares (Kutner *et al.*, 2004) são chamados assim pois seus parâmetros são obtidos de forma linear. Dada uma amostra de n observações, seja $\mathbf{Y} = (y_1, y_2, \cdots, y_n)'$ o vetor das n observações da variável de interesse e $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ uma matrix n x p das p variáveis explicativas, em que cada $\mathbf{x}_i = (x_{i1}, x_{i2}, ..., x_{ip})'$, com $i = 1, 2, ..., n$. Os modelos lineares podem ser escritos na forma matricial como

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.1}$$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)'$ é o vetor de parâmetros desconhecidos que precisam ser estimados por algum método de estimação e $\boldsymbol{\epsilon}$ é um vetor de erros aleatórios que segue uma distribuição normal com média zero e matriz de covariância $\mathbf{R} = \mathbf{1}'_n\sigma_\epsilon$, em que $\mathbf{1}'_n$ é um vetor de uns transposto.

Os parâmetros dos modelos lineares são considerados fixos, ou seja, eles têm um valor único, porém desconhecido. Assim, usando técnicas de estimação como mínimos quadrados ou máxima verossimilhança, obtemos como estimativa para o vetor $\boldsymbol{\beta}$ o valor

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

Assim, podemos predizer o valor da variável resposta da observação y_i por meio da estimativa do seu valor esperado dada por

$$\hat{Y}_i = \mathbf{X}_i\hat{\boldsymbol{\beta}}, \tag{4.2}$$

em que $i = 1, 2, ..., n$ e $\mathbf{X}_i = (x_{i1}, x_{i2}, ..., x_{ip})$ é o vetor das p covariáveis da observação i.

4.2 Modelos Lineares Mistos

Existem situações nas quais as observações da nossa análise não são totalmente independentes entre si, mas organizadas em grupos de dependência ou indíviduos. Nesses casos, devemos considerar que exista um fator que influencia nossa variável resposta e que dependa do indivíduo analisado. Quando esse fator existe, nós dizemos que existe um efeito aleatório na função $f(\mathbf{X})$ que é usada. Quando temos um conjunto de variáveis que, entre elas existem variáveis de efeitos fixos e variáveis de efeitos aleatórios, temos um tipo de modelo chamado de misto.

Dada uma amostra de n observações tal que para cada indivíduo j temos n_j observações, $\sum n_j = n$ e $j = 1, 2, ..., J$, podemos escrever os modelos mistos normais (Demidenko, 2013) pela expressão

$$\mathbf{Y}_j = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{b}_j + \boldsymbol{\epsilon}_j, \tag{4.3}$$

em que $j = 1, 2, ..., J$ indica os indivíduos, $\mathbf{Y}_j = (Y_{1j}, Y_{2j}, ..., Y_{n_jj})$ é o vetor das observações do individuo j, $\mathbf{X}_j$ é a matriz n_j x p de variáveis explicativas com efeito fixo, $\boldsymbol{\beta}$ é o vetor dos p parâmetros com efeitos fixos, $\mathbf{Z}$ é a matriz n_j x q de variáveis com efeitos aleatórios e $\mathbf{b}_j = (b_{j1}, b_{j2}, ..., b_{jq})'$ é o vetor das q variáveis de efeitos aleatórios com distribuição $N(\mathbf{0}, \mathbf{D}_j)$. Para o vetor $\boldsymbol{\epsilon_j}$ é assumida a distribuição $N(\mathbf{0}, \mathbf{R}_j)$. As matrizes $\mathbf{D}_j$ e $\mathbf{R}_j$ são as matrizes de covariância de cada variável aleatória.

É assumido que os efeitos aleatórios são independentes entre si e que eles são independentes de ϵ_j.

Podemos escrever o conjunto das J equações dadas por (4.3) como

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}, \tag{4.4}$$

em que $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, ..., \mathbf{Y}_J)'$ é o vetor das n observações, $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_J)$ é uma matriz n x p, $\mathbf{Z} = diag(\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_J)$ é uma matriz diagonal das J matrizes Z_j dos efeitos aleatórios, $\mathbf{b} = (\mathbf{b}_1, \mathbf{b}_2, ..., \mathbf{b}_J)'$ é o vetor dos Jk efeitos aleatórios do modelo que possui distribuição $N(\mathbf{0}, \mathbf{D})$ e $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1, \boldsymbol{\epsilon}_2, ..., \boldsymbol{\epsilon}_J)$ é o vetor dos erros do modelo com $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{R})$. Assim, de (4.4) pode se obter a distribuição condicional

$$\mathbf{Y}|\mathbf{b} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}, \mathbf{R})$$
$$\mathbf{b} \sim N(\mathbf{0}, \mathbf{D}).$$

Para encontrar a média e a covariância da distribuição marginal de $\mathbf{Y}$, fazemos

$$E(\mathbf{Y}) = E(E(\mathbf{Y}|\mathbf{b})) = E(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}E(\mathbf{b}) = \mathbf{X}\boldsymbol{\beta}$$
$$Cov(\mathbf{Y}) = E(Cov(\mathbf{Y}|\mathbf{b})) + Cov(E(\mathbf{Y}|\mathbf{b})) = \mathbf{R} + Cov(\mathbf{Z}\mathbf{b}) = \mathbf{R} + \mathbf{Z}\mathbf{D}\mathbf{Z}'.$$

Facilitando a notação, podemos escrever $Cov(\mathbf{Y}) = \mathbf{R} + \mathbf{ZDZ}' = \mathbf{V}$. Assim a distribuição marginal de $\mathbf{Y}$ será dada por

$$\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V}).$$

Para estimar o vetor $\boldsymbol{\beta}$ de parâmetros pode-se usar, por exemplo, o método da máxima verossimilhança (ML) obtendo a estimativa

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}\hat{\mathbf{V}}^{-1}\mathbf{Y}, \tag{4.5}$$

em que $\hat{\mathbf{V}}$ deve ser estimado iterativamente pelas equações

$$tr(\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i') = \mathbf{Y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Y}$$
$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}.$$

Para mais detalhes do processo de estimação de $\boldsymbol{\beta}$, ver Demidenko (2013) e Searle *et al.* (1992).

Para os efeitos aleatórios $\mathbf{b}$, devemos predizer o seu valor por meio do seu valor esperado condicionado ao valores realizados de Y, isto é,

$$E[\mathbf{b}|\mathbf{Y}] = \mathbf{DZ}'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

Fazendo as substituições de $\boldsymbol{\beta}$ por $\hat{\boldsymbol{\beta}}$, $\mathbf{V}$ por $\hat{\mathbf{V}}$ e $\mathbf{D}$ por $\hat{\mathbf{D}}$ temos

$$\tilde{\mathbf{b}} = \hat{\mathbf{D}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}). \tag{4.6}$$

Esse valor é chamado de *best linear prediction* (BLUP) (McCulloch e Searle, 2005).

Assim, a melhor predição do valor esperado de Y_{ij} da observação i do indivíduo j será dada por

$$\hat{Y}_{ij} = \hat{\boldsymbol{\beta}}\mathbf{X}_{ij}' + \tilde{\mathbf{b}}_j\mathbf{Z}_{ij}', \tag{4.7}$$

em que $i = 1, 2, ..., n_j$, $j = 1, 2, ..., J$, $\hat{\boldsymbol{\beta}}$ é o vetor das estimativas dos efeitos fixos do modelo, $\mathbf{X}_{ij}$ é o vetor das covariáveis de efeitos fixos da observação i do indivíduo j, $\tilde{\mathbf{b}}_j$ é o vetor das predições dos efeitos aleatórios do individuo j e $\mathbf{Z}_{ij}$ é o vetor das covariáveis de efeitos aleatórios da observação i do indivíduo j.

4.3 Modelos Lineares Generalizados

Os modelos citados nas **Seções 4.1** e **4.2** tinham a suposição que nossa variável resposta seguia a distribuição normal. Na prática, essa suposição se adequa a grande parte dos problemas, mas em alguns casos essa suposição é violada. Para contornar esse problema, durante muito tempo, se utilizava o método de transformação da variável resposta para que esta nova variável possuísse a distribuição normal e se pudesse aplicar a metodologia usual. Porém, essa transformação trazia como consequência a dificuldade de interpretação dos parâmetros e o fato dela não ser a única possível.

Para solucionar esse problema, Nelder e Wedderburn (1972) propuseram os modelos lineares generalizados (GLM, *Generalized Linear Models*), que permitem que a variável resposta pertença a um grupo de distribuições chamado de família exponencial.

Para pertencer à família exponencial, o função de distribuição da variável Y deve ser tal que sua escrita pode ser feita da forma

$$f(\mathbf{Y}) = exp[\phi(\mathbf{Y}\boldsymbol{\theta} - b(\boldsymbol{\theta}))] + c(\mathbf{Y}, \phi), \tag{4.8}$$

$E(\mathbf{Y}) = \boldsymbol{\mu} = b'(\boldsymbol{\theta})$, $Var(\mathbf{Y}) = \phi^{-1}\mathbf{V}$, $\mathbf{V} = \frac{d\boldsymbol{\mu}}{d\boldsymbol{\theta}}$ é a função variância, ϕ^{-1} é o parâmetro de dispersão e $b(\boldsymbol{\theta})$ e $c(\mathbf{Y}, \phi)$ são funções que dependem apenas dos parâmetros envolvidos.

As distribuições Normal, Gama, Poisson, Binomial e Normal Inversa são exemplos de distribuições que atendem a escrita de (4.8). Além de (4.8), os modelos lineares generalizados estão caracterizados pela parte sistemática dada por

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} \tag{4.9}$$

em que $\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$ é o preditor linear, $\boldsymbol{\beta}$ é o vetor de parâmetros a serem estimados e $\mathbf{X}$ é a matriz n x p das variáveis explicativas. $g(.)$ é uma função monótona e diferenciável, chamada de função de ligação. Assim, os valores de $\boldsymbol{\eta}$ podem variar livremente em $\mathbb{R}$ e a inversa $\boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta})$ garante que o modelo atenda a distribuição da variável resposta. Para cada distribuição da família exponencial, algumas funções $g(.)$ podem ser utilizadas, sendo um caso particular a ligação canônica, que ocorre quando $\boldsymbol{\theta} = \boldsymbol{\eta}$.

Para a estimação do vetor $\boldsymbol{\beta}$, se requer um processo iterativo baseado na função escore e descrito sucintamente a seguir. Dada a função de verossimilhança $L(\boldsymbol{\beta}, \phi)$, a função escore é dada por

$$\mathbf{U}_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \phi) = \frac{dL(\boldsymbol{\beta}, \phi)}{d\boldsymbol{\beta}} = \phi\mathbf{X}'\mathbf{W}^{\frac{1}{2}}\mathbf{V}^{-\frac{1}{2}}(\mathbf{Y} - \boldsymbol{\mu}),$$

em que $\mathbf{W} = diag(w_1, ..., w_n)$ é a matriz de pesos, onde cada w_i é dado por $w_i = (d\mu_i/d\eta_i)^2/V_i$.

A obtenção da estimativa de máxima verossimilhança para $\boldsymbol{\beta}$ é dada pelo processo iterativo de Newton-Raphson, expandindo a função escore $\mathbf{U}_{\boldsymbol{\beta}}$ em torno de um $\boldsymbol{\beta}^{(0)}$

$$\mathbf{U}_{\boldsymbol{\beta}} \cong \mathbf{U}_{\boldsymbol{\beta}}^{(0)} + \mathbf{U}_{\boldsymbol{\beta}}^{(0)'}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}),$$

que resulta num processo iterativo de mínimos quadrados reponderados, dado por

$$\boldsymbol{\beta}^{(m+1)} = (\mathbf{X}'\mathbf{W}^{(m)}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(m)}\mathbf{z}^{(m)}, \tag{4.10}$$

em que $\mathbf{z} = \boldsymbol{\eta} + \mathbf{W}^{-\frac{1}{2}}\mathbf{V}^{-\frac{1}{2}}(\mathbf{Y} - \boldsymbol{\mu})$. A convergência de (4.10) acontece num número finito de passos m, resultando na estimativa $\hat{\boldsymbol{\beta}}$.

Com o valor da estimativa $\hat{\boldsymbol{\beta}}$, podemos predizer o valor esperado de Y_i por (4.2). Para mais detalhes destes modelos, consultar Paula (2013).

4.4 Modelos Mistos Generalizados

Podemos pensar também em acrescentar efeitos aleatórios na parte sistemática do modelo linear generalizado. Esse modelo é conhecido como Modelo Linear Generalizado Misto (McCulloch e Searle, 2005), também conhecido por GLMM (*Generalized Linear Mixed Model*), que é expressado por

$$\mathbf{Y}_j|\mathbf{b}_j \sim f_{\mathbf{Y}_j|\mathbf{b}_j}(\mathbf{Y}_j|\mathbf{b}_j),$$
$$f_{\mathbf{Y}_j|\mathbf{b}_j}(\mathbf{Y}_j|\mathbf{b}_j) = exp[\phi(\mathbf{Y}_j\boldsymbol{\theta}_j - h(\boldsymbol{\theta}_j))] + c(\mathbf{Y}_j, \phi),$$

em que $Y_j = (Y_{1j}, Y_{2j}, ..., Y_{n_jj})'$ é o vetor das n_j observações do indivíduo j e assim como (4.8), $E(\mathbf{Y}_j|\mathbf{b}_j) = \boldsymbol{\mu}_j = h'(\boldsymbol{\theta}_j)$, e essa parte sistemática é dada por

$$g(\boldsymbol{\mu}_j) = \mathbf{X}_j\boldsymbol{\beta} + \mathbf{Z}_j\mathbf{b}_j,$$

em que $\mathbf{X}_j$ é a matriz n_j x p das variáveis de efeito fixo do indivíduo j, $\boldsymbol{\beta}$ é o vetor dos p parâmetros de efeitos fixos, $\mathbf{Z}_j$ é a matrix n_j x q das variáveis de efeitos aleatórios do indivíduo j e $\mathbf{b}_j$ é o vetor dos q parâmetros de efeito aleatório do indivíduo j.

Para a estimação do vetor de parâmetros $\boldsymbol{\beta}$ e a predição dos efeitos aletórios $\mathbf{b}_j$ pode-se usar métodos baseados na máxima verossimilhança. Porém, como as derivadas não têm forma fechada, deve-se utilizar aproximações como a de Laplace ou a Quadratura de Gauss-Hermite (Demidenko, 2013).

Após se obter as estimativas dos parâmetros e os valores preditos dos efeitos aleatórios, a predição do valor esperado de Y_{ij} será dada por 4.7.

Capítulo 5

Estudos de Simulação

Neste capítulo será avaliado o comportamento das técnicas de validação de modelos apresentadas no **Capítulo 2**. Serão usados conjuntos de dados simulados de distribuições Normal e Logística em amostras com tamanhos diferentes. Serão analisados a diferença de comportamento entre elas, a influência do tamanho da amostra no resultado e, principalmente, o impacto que ocorre quando o modelo proposto possui efeitos aleatórios, comparando os resultados com o modelo de efeitos fixos associado. Os dados simulados que usaremos nessa seção e todos os cenários de simulação foram criados atráves do *software R 3.5.0*.

A partir do que foi apresentado no **Capítulo 2**, foi realizado a comparação de desempenho de 13 estimatidores de (2.7), que estão descritas na **Tabela 5.1**, em 4 modelos: Normal com efeitos fixos, Normal Misto, Logístico com efeitos fixos e Logístico Misto.

Técnica	Descrição	Parâmetros	Fórmula
errapt	Erro aparente	-	(2.8)
kf10	*K-fold*	$K = 10$	(3.3)
bkf10	*Burman K-fold*	$K = 10$	(3.4)
rkf10	*Repeated K-fold*	$K = 10$ e $R = 5$	(3.5)
rbkf10	*Repeated Burman K-fold*	$K = 10$ e $R = 5$	(3.6)
loo	*Leave One Out*	-	(3.7)
ho3	*Hold-out*	$p = 1/3$	(3.1)
ho10	*Hold-out*	$p = 1/10$	(3.1)
rho3	*Repeated Hold-out*	$p = 1/3$ e $R = 50$	(3.2)
rho10	*Repeated Hold-out*	$p = 1/10$ e $R = 50$	(3.2)
bt	*Bootstrap*	$K = 50$ e $m = \frac{2n}{3}$	(3.8)
bt632	*Bootstrap* Ponderado	*bt* e *errapt*	(3.9)
bt632+	*Bootstrap* Ponderado	*bt*, *errapt* e $\hat{w}$	(3.10)

Tabela 5.1: *Descrição das estimativas usadas na simulação*

Para realizar esta avaliação nos modelos normais foi gerada a base $\mathbf{B}_1$, com as covariáveis $\mathbf{X}$, de acordo com a descrição dada na **Seção 5.1**. Outra base $\mathbf{B}_2$ foi gerada para os modelos logísticos, com a construção das variáveis $\mathbf{X}$ seguindo os mesmos critérios da base $\mathbf{B}_1$.

Após a escolha da função preditora de cada modelo, as variáveis respostas foram geradas de acordo com o que será apresentado na **Seção 5.2**, usando as bases $\mathbf{B}_1$ ou $\mathbf{B}_2$, obtendo para cada uma dela uma nova base com a inclusão das variáveis respostas $\mathbf{BC}_1$ e $\mathbf{BC}_2$.

Seguindo as propostas nos esquemas de simulação de Borra e Ciaccio (2010) e Kim (2009), foi analisado o comportamento das técnicas de validação em amostras de 120, 160, 200, 400, 600, 800, 1000 observações. Para isso, para um determinado tamanho de amostra foram retiradas sem reposição 100 amostras aleatórias da base $\mathbf{BC}_1$ para os modelos normais e da $\mathbf{BC}_2$ para os modelos

logísticos.

Para cada amostra d retirada foram aplicadas as seguintes etapas, mostradas também na **Figura 5.1**:

1. Usando a amostra completa criou-se o preditor $\hat{f}(\mathbf{X})$ usando a técnica de estimação apropriada para cada modelo.

2. Dado $\hat{f}(\mathbf{X})$ obtido no passo anterior e usando-se uma amostra de validação de 5 mil observações retirada da base $\mathbf{BC}_1$ para os modelos normais e $\mathbf{BC}_2$ para os modelos logísticos se obteve os valores preditos para Y para essas 5 mil observações. Assim, tendo os valores verdadeiros e os preditos se obteve a ERR por meio do estimador dado em (2.9), usando a função de perda apropriada a cada modelo, ou seja, a perda quadrática, dada por (2.3), nas simulações dos modelos normais e a perda Entropia Cruzada, dada por (2.6), nas simulações dos modelos logísticos.

3. Cada uma das técnicas (**Tabela 5.1**) foi aplicada para para a obtenção da correspondente estimativa err, dividindo-se a amostra d de acordo com cada técnica e usando-se a função de perda correspondente a cada modelo.

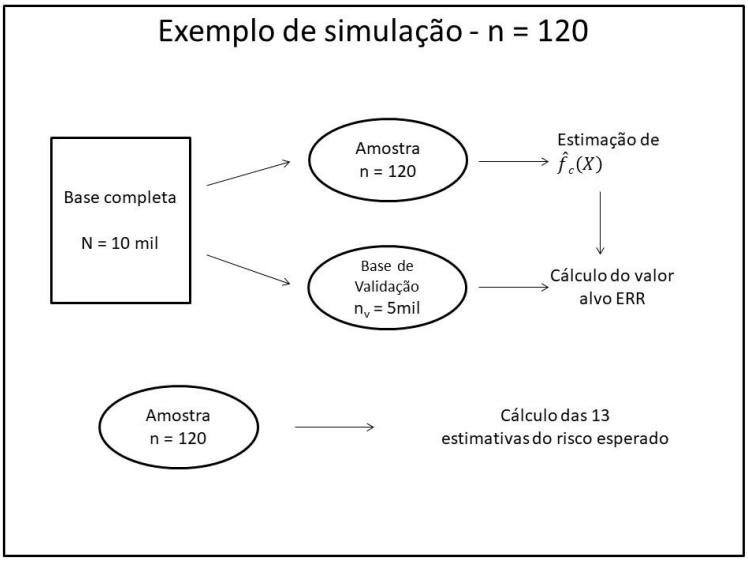

Figura 5.1: *Esquema de uma amostra de simulação*

As etapas 1 a 3 se repetem 100 vezes. A média dos 100 valores de ERR obtidos foi considerada com uma boa aproximação do valor de (2.7).

Foi considerado também como o valor obtido de cada um dos 13 estimadores a média das 100 amostras, obtendo-se $\bar{err}$.

O viés de cada estimativa foi obtido considerando a diferença entre o valor resultante em cada amostra e o valor de ERR correspondente daquela amostra. O viés médio será calculado usando

$$\bar{vies} = \sum_{i=1}^{100} \frac{(err_i - ERR_i)}{100}, \tag{5.1}$$

em que err_i é o valor da estimativa de cada técnica na amostra i, $i = 1, ..., 100$.

Por fim, foi calculado a variância amostral de cada técnica avaliada usando

$$var = \sum_{i=1}^{100} \frac{(err_i - e\bar{r}r)^2}{99}, \tag{5.2}$$

em que $e\bar{r}r$ é a média das estimativas de err das 100 simulações, para cada técnica avaliada.

Resumindo todas as etapas descritas aqui, apresentamos a **Figura 5.2**.

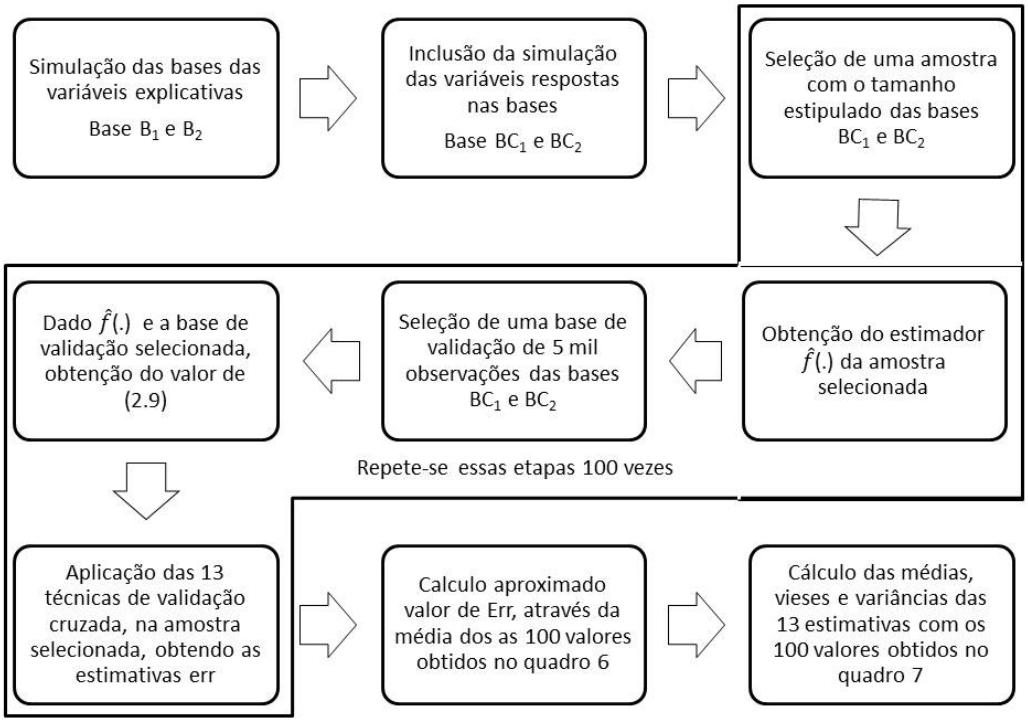

Figura 5.2: *Esquemas das etapas de simulação*

5.1 Descrição das bases de dados

Para esse estudo foram criadas a base $\mathbf{B}_1$, usada nos modelos normais, e a base $\mathbf{B}_2$, usada nos modelos logísticos. Para cada base, foram geradas $N = 10000$ observações. Para que seja possível usar a técnica dos modelos mistos, foram considerados 40 indíviduos, com 250 observações cada, ou seja, $N_j = 250$ e $\sum_{j=1}^{40} N_j = N = 10000$. As variáveis criadas terão as mesmas estruturas nos dois bancos de dados criados.

As covariáveis criadas foram geradas através de funções de números aleatórios do R de acordo com as distribuições abaixo:

$X_1 \sim N(0,1)$;

$$X_{2w} = \begin{cases} 1, \text{se } \frac{w-1}{3} < u <= \frac{w}{3} \text{ com } u \sim U(0,1) \\ 0, \text{caso contrário.} \end{cases} \text{, para } w = 1,2,3;$$

$X_3 \sim Gamma(1, 4/3)$;

$X_4 \sim Beta(5, 2)$;

$Z_1 \sim Beta(2, 7)$.

As covariáveis $\mathbf{X}$ são consideradas de efeito fixo e a variável Z de efeito aleatório.

Os parâmetros do intercepto e de cada covariável de efeito fixo foram determinados pelos seguintes valores: $\beta_0 = 0.75$, $\beta_1 = -2.75$, $\beta_{21} = -1.5$, $\beta_{22} = 0$, $\beta_{23} = 1.5$, $\beta_3 = 1.25$ e $\beta_4 = 0.75$.

Para os casos dos modelos mistos, o vetor de interceptos aleatórios $\mathbf{b}_0 = (b_{01}, b_{02}, ..., b_{040})$ e vetor de parâmetros da covariável de efeito aleatório $\mathbf{b}_1 = (b_{11}, b_{12}, ..., b_{140})$ foram gerados aleatoriamente através de distribuições normais com média zero e desvios padrão iguais a $\sigma_{b0} = 0.33$ e $\sigma_{b1} = 0.7$, respectivamente, ou seja, $b_{0j} \sim N(0, (0.33)^2)$ e $b_{1j} \sim N(0, (0.7)^2)$, para $j = 1, 2, ..., 40$.

5.2 Descrição das variáveis respostas

Iremos analisar 4 tipos de modelos usando os bancos de dados descritos na seção anterior. Nessa seção apresentaremos a estrutura definida para esses modelos e a forma de construção da variável resposta.

5.2.1 Modelo 1: Normal com efeitos fixos

A função preditora para este modelo tem as seguintes características:

$$\eta_i = \beta_0 + \beta_1 * X_{1i} + \beta_{21i} * X_{21i} + \beta_{22} * X_{22i}\beta_{23} * X_{23i} + \beta_3 * X_{3i} + \beta_4 * X_{4i},$$

$$f_1(\mathbf{X}_i) = \eta_i.$$

em que $i = 1, 2, ..., 10000$ e η_i é o valor médio da distribuição normal associado a observação i e $\mathbf{B}_1$ foi o banco de dados usado. Note que nesse caso não estamos considerando a estrutura de efeitos aleatórios, ou seja, as observações são consideradas independentes.

Para simular os valores de Y_{1i}, foram gerados os valores de ϵ_{1i}, $i = 1, 2, ..., 10000$, através de uma distribuição normal com média zero e desvio padrão $\sigma_\epsilon = 0.33$ e assim,

$Y_{1i} = f_1(X_i) + \epsilon_{1i}$.

Portanto $\mathbf{Y}_1 = \{Y_{1i}, i = 1, 2, ..., 10000\}$ terá uma distribuição normal com média $\boldsymbol{\eta} = (\eta_i, i = 1, 2, ..., 10000)$ e variância $\mathbf{1}_n\sigma_\epsilon^2$.

O preditor $\hat{f}_1(\mathbf{X})$ deverá retornar o valor médio da distribuição normal associado a Y e o erro esperado será calculado com a função de perda (2.3), onde a penalização da predição será maior quanto mais distante $\hat{f}_1(\mathbf{X})$ estiver de $\mathbf{Y}_1$.

5.2.2 Modelo 2: Normal Misto

A função preditora desse modelo tem as seguintes características:

$$\eta_{ij} = \beta_0 + \beta_1 * X_{1ij} + \beta_{21} * X_{21ij} + \beta_{22} * X_{22ij} + \beta_{23} * X_{23ij} + \beta_3 * X_{3ij} + \beta_4 * X_{4ij} + b_{0j} + b_{1j} * Z_{1ij},$$

$$f_2(X_{ij}) = \eta_{ij},$$

em que $i = 1, 2, ..., 250$, $j = 1, 2, ..., 40$, η_{ij} é o valor médio da distribuição normal associado a observação i do indivíduo j e o banco de dados usados foi o $\mathbf{B}_1$. Nesse caso, como usamos os efeitos aleatórios, estamos considerando que as observações foram retiradas de 40 indivíduos e que as 250 observações de cada indivíduo possuem dependência entre si.

Para simular os valores de Y_{2ij}, foram gerados os valores de ϵ_{2ij}, $i = 1, 2, ..., 250$, $j = 1, 2, ..., 40$, através de uma distribuição normal com média zero e desvio padrão $\sigma_\epsilon = 0.33$ e assim,

$$Y_{2ij} = f_2(X_{ij}) + \epsilon_{2ij}.$$

Portanto $\mathbf{Y}_2 = \{Y_{2ij}, i = 1, 2, ..., 250, j = 1, 2, ..., 40\}$ terá uma distribuição normal com média $\boldsymbol{\eta} = \{\eta_{ij}, i = 1, 2, ..., 250, j = 1, 2, ..., 40\}$ e variância $\mathbf{1}_n\sigma_\epsilon^2$.

O preditor $\hat{f}_2(\mathbf{X})$ deverá retornar o valor médio da distribuição normal associado a Ye o erro esperado será calculado com a função de perda (2.3), onde a penalização da predição será maior quanto mais distante $\hat{f}_2(\mathbf{X})$ estiver de $\mathbf{Y}_2$.

5.2.3 Modelo 3: Logístico com efeitos fixos

A função preditora desse modelo tem as seguintes características:

$$\eta_i = \beta_0 + \beta_1 * X_{1i} + \beta_{21} * X_{21i} + \beta_{22} * X_{22i} + \beta_{23} * X_{23i} + \beta_3 * X_{3i} + \beta_4 * X_{4i},$$

$$f_3(X_i) = exp(\eta_i)/(1 + exp(\eta_i)),$$

em que $i = 1, 2, ..., 10000$, $f_3(X_i)$ é a probabilidade de sucesso do indivíduo i e o banco de dados usado foi o $\mathbf{B}_2$. Note que nesse caso não estamos considerando a estrutura de efeitos aleatórios, ou seja, as observações são consideradas independentes.

Para simular a variável resposta Y_{3i} foi usada a função de números aleátorios da binomial do *software R*, com cada $f_3(X_i)$ como média de cada Y_{3i}:

$Y_{3i} = rbinom(10000, 1, f_3(X_i))$,
$(i = 1, 2, ..., 10000)$,

em que i é o índice de cada observação e $f_3(X_i)$ é o valor da probabilidade de sucesso de cada observação. Assim $\mathbf{Y}_3 = \{Y_{3i}, i = 1, ..., 10000\}$ será o vetor de variáveis respostas do nosso modelo, com valores indicando 1 (sucesso) ou 0 (fracasso).

O preditor $\hat{f}_3(\mathbf{X})$ deverá retornar a probabilidade esperada de um indíviduo ser considerado sucesso. O erro esperado desse preditor usará a função de perda (2.6), onde a penalização ocorre quando a probabilidade esperada de sucesso é baixa, mas o indíviduo possui $Y = 1$ ou quando a probabilidade esperada de sucesso é alta, mas o indíviduo possui $Y = 0$.

5.2.4 Modelo 4: Logístico Misto

A função preditora desse modelo tem as seguintes características:

$$\eta_{ij} = \beta_0 + \beta_1 * X_{1ij} + \beta_{21} * X_{21ij} + \beta_{22} * X_{22ij} + \beta_{23} * X_{23ij} + \beta_3 * X_{3ij} + \beta_4 * X_{4ij} + b_{0j} + b_{1j} * Z_{1ij},$$

$$f_4(X_{ij}) = exp(\eta_{ij})/(1 + exp(\eta_{ij})),$$

em que $i = 1, 2, ..., 250$, $j = 1, 2, ..., 40$ e a base de dados usada foi a $\mathbf{B}_2$. Nesse caso, como usamos os efeitos aleatórios, estamos considerando que as observações foram retiradas de 40 indivíduos e que as 250 observações de cada indivíduo possuem dependência entre si.

Para simular o vetor de variáveis respostas $\mathbf{Y}_4$ foi usado a função de números aleátorios da binomial, com cada $f_4(X_{ij})$ como média de cada Y_{4ij}:

$Y_{4ij} = rbinom(10000, 1, f_4(X_{ij}))$,
$(i = 1, 2, ..., 250), (j = 1, 2, ..., 40)$.

em que i é o índice de cada observação, j corresponde a cada indivíduo e $f_4(X_{ij})$ é a função simulada correspondente a cada observação. Assim $\mathbf{Y}_4 = \{Y_{4ij}, i = 1, ..., 250, j = 1, ...40\}$ será o vetor de variáveis respostas do nosso modelo, com valores indicando 1 (sucesso) ou 0 (fracasso).

O preditor $\hat{f}_4(\mathbf{X})$ deverá retornar a probabilidade esperada de uma observação de um indíviduo ser considerado sucesso. O erro esperado desse preditor usará a função de perda (2.6), onde a penalização ocorre quando a probabilidade esperada de sucesso é baixa, mas a observação possui $Y = 1$ ou quando a probabilidade esperada de sucesso é alta, mas a observação possui $Y = 0$.

5.3 Discussão dos resultados das simulações

5.3.1 Modelo 1: Normal com efeitos fixos

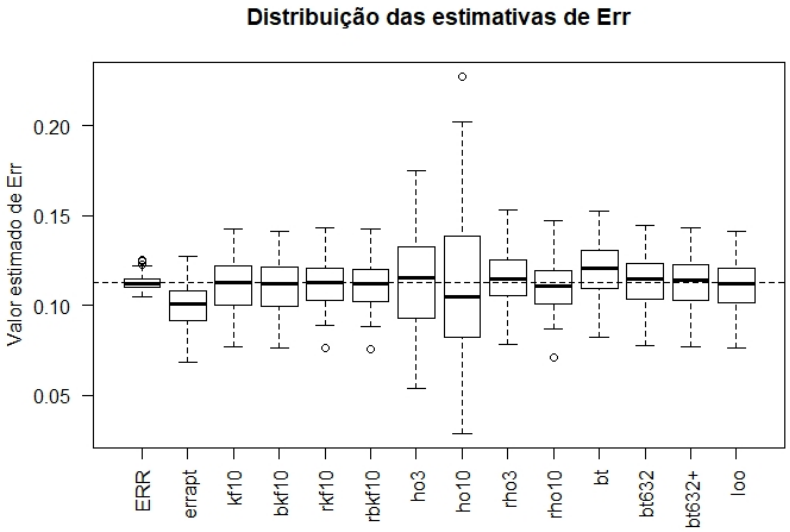

Figura 5.3: *Distribuição dos estimadores de Err para n=120 no Modelo 1*

A **Figura 5.3** mostra o comportamento da distribuição das estimativas do valor de (2.7) para as 100 amostras de tamanho $n = 120$ e indica também a variabilidade de cada estimativa. A linha tracejada na figura indica o valor da média de ERR.

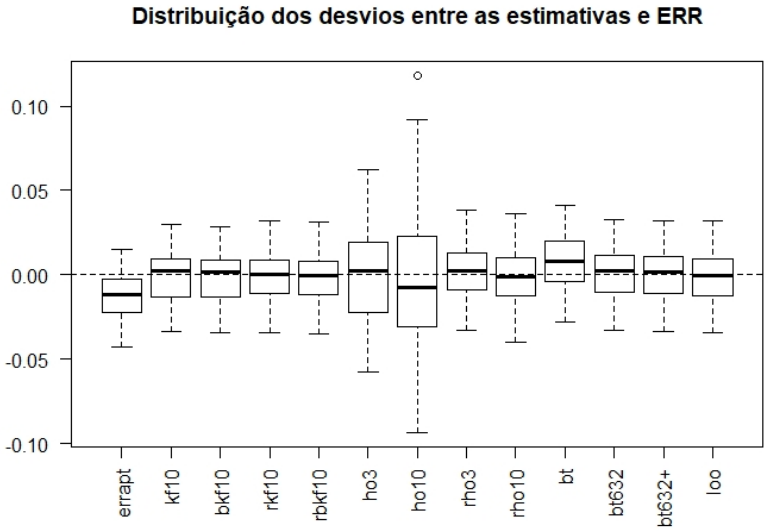

Figura 5.4: *Distribuição das diferenças entre o valor de ERR e o valor obtido pelos estimadores para $n = 120$ no Modelo 1*

Podemos notar distribuições muito próximas entre si na maior parte das estimativas, com medianas próximas da linha da média de ERR, com exceção de *errapt* e dispersões parecidas. O que pode se notar de diferente uma forte dispersão nos valores para $ho10$ e $ho3$ e um deslocamento para cima da distribuição de bt. A ponderação dos métodos $bt632$ e $bt632+$, nesse caso, as favorecem, tornando-as mais competitivas com relação as demais.

Como usamos a mesma amostra para calcular o valor obtido pelos 13 estimadores, podemos calcular a diferença entre o valor de ERR e cada estimativa para cada amostra. A **Figura 5.4** indica a distribuição dessas diferenças para as 100 amostras de tamanho $n = 120$. Podemos interpretar esse gráfico como a distribuição do viés, sendo que quanto mais centrada em zero estiver a distribuição, mais assertiva é a técnica para estimar Err. Analisando o gráfico, vemos que as medianas dos desvios para o valor de ERR da maioria das estimativas estão centradas em 0, porém podemos notar uma leve assimetria para valores positivos nas estimativas $kf10$, $bkf10$, $rkf10$ e $rbkf10$ indicando que essas técnicas podem superestimar o risco esperado. Os estimadores *errapt* e $ho10$ possuem uma distribuição subviesada e a distribuição dos desvios para ERR da estimativa bt apresenta vieses positivos.

As médias $\bar{err}$ das 100 amostras de cada estimativa estão representadas nas **Figuras 5.5** a **5.8**. A separação em 4 gráficos foi realizada para facilitar a visualização das trajetórias, agrupando num mesmo gráfico técnicas similares. A **Figura** 5.9 mostra um resumo das melhores técnicas, após a análise geral.

Além disso, a **Tabela 5.2** apresenta o valor do viés médio de cada estimador, obtido através de (5.1) e a **Tabela**. **5.3** apresenta as variâncias dos estimadores, que foram calculadas por (5.2).

Podemos notar que *errapt* é o estimador mais subviesado do risco esperado, principalmente nas amostras pequenas (**Tabela 5.2** e **Figura 5.5**).

Na **Figura 5.6** podemos observar que $kf10$ e $rkf10$ estimam, no geral, quase sempre o mesmo valor, assim como $bkf10$ e $rbkf10$. Na **Tabela 5.2** vemos a semelhança dos vieses médios entre esses estimadores. Portanto, podemos concluir a repetição do método *K-fold* não provoca grandes

mudanças no valor, o que torna o custo de realizá-la desnecessário. Os estimadores com correção de Burman ($bkf10$ e $rkf10$) apresentam valores menores e mais próximos de ERR do que suas versões sem correção (**Figura 5.6**).

Com relação aos estimadores pelo método *Hold-Out*, como esperado, $ho10$ e $ho3$ são os estimadores que apresentam maior variabilidade entre todos os estimadores (**Tabela 5.3**) e todos os tamanhos de amostra realizados sendo que suas versões com repetições $rho3$ e $rho10$ possuem variabilidade próximas das demais técnicas.

O estimador bt, em geral, é o estimador mais superviesado (**Figura 5.8** e **Tabela 5.2**). As ponderações $bt632$ e $bt632+$ estão mais próximas do valor de ERR, sendo que nesse caso, $bt632+$ é a mais próxima da trajetória de ERR (**Figura 5.8**) e com os menores vieses (**Tabela 5.2**).

Por fim, podemos ver que os estimadores mais próximos de ERR são $bkf10$, loo nas amostras menores (**Figura 5.9**). A partir do tamanho de amostra 400, as estimativas ficam bem próximas entre si, sendo que as estimativas $bkf10$, $rho10$ e $bt632+$ são as mais próximas do esperado.

Técnica	120	160	200	400	600	800	1000
errapt	-11.571	-5.800	-4.671	-3.181	-1.236	-1.493	-1.494
kf10	-0.026	2.819	2.443	0.153	1.006	0.220	-0.179
bkf10	-0.651	2.355	2.062	-0.024	0.888	0.129	-0.248
rkf10	-0.046	3.066	2.375	0.149	1.068	0.204	-0.157
rbkf10	-0.670	2.589	1.998	-0.028	0.946	0.114	-0.227
ho3	0.996	4.584	2.476	1.284	1.707	0.514	0.986
ho10	-2.051	2.608	5.039	-0.363	3.149	0.514	1.256
rho3	2.903	4.706	4.106	0.928	1.575	0.737	0.062
rho10	-0.927	3.579	2.633	-0.088	1.478	0.324	-0.377
bt	7.945	9.126	7.255	2.243	2.624	1.259	0.754
bt632	1.486	4.351	3.570	0.910	1.867	0.901	0.576
bt632+	0.810	3.657	2.881	0.250	1.205	0.247	-0.073
loo	-0.572	2.497	1.957				

Tabela 5.2: *Viés médio ($\times 10^{-3}$) dos 13 estimadores testados no Modelo 1*

Técnica	120	160	200	400	600	800	1000
errapt	14.457	15.394	9.746	6.040	3.493	2.349	2.478
kf10	18.161	18.152	11.182	6.267	3.634	2.384	2.547
bkf10	17.916	17.984	11.093	6.253	3.625	2.381	2.543
rkf10	17.874	18.428	11.023	6.391	3.706	2.471	2.537
rbkf10	17.669	18.252	10.951	6.371	3.695	2.464	2.534
ho3	73.798	50.142	32.676	19.813	15.082	9.161	7.356
ho10	160.385	185.738	112.531	60.143	43.656	25.116	25.865
rho3	20.499	19.302	11.820	6.671	3.964	2.721	2.622
rho10	21.769	21.031	14.352	6.542	3.918	3.146	2.669
bt	20.730	19.978	12.471	6.604	3.904	2.632	2.674
bt632	18.364	18.360	11.499	6.441	3.775	2.544	2.620
bt632+	18.154	18.143	11.361	6.364	3.73	2.514	2.589
loo	17.836	18.112	10.978				

Tabela 5.3: *Variância ($\times 10^{-5}$) dos 13 estimadores testados no Modelo 1*

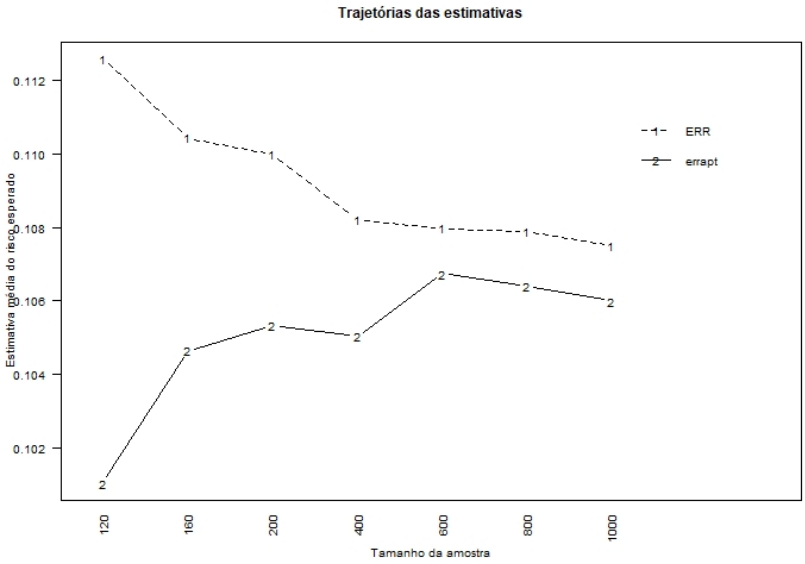

Figura 5.5: *Trajetória das médias dos valores obtidos nas 100 amostras para o estimador errapt no Modelo 1*

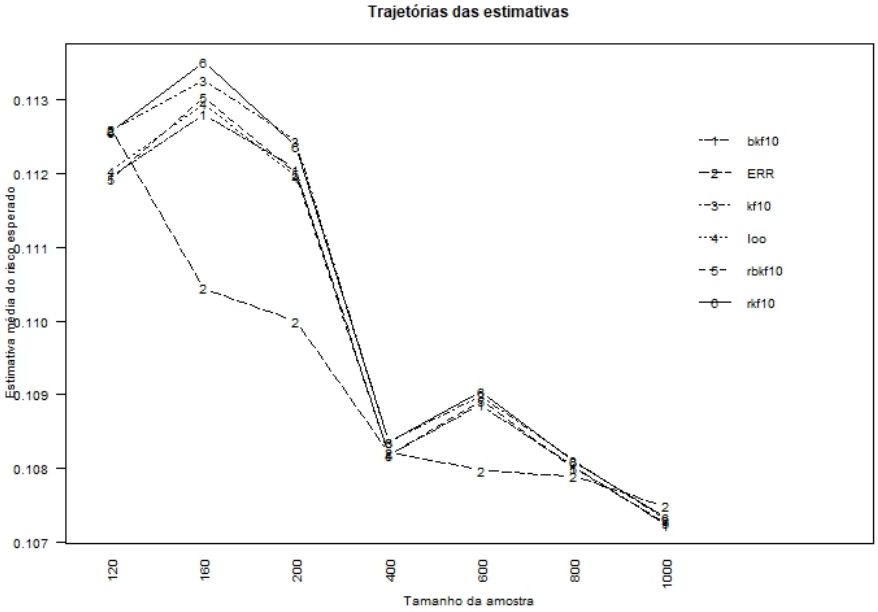

Figura 5.6: *Trajetória das médias dos valores obtidos nas 100 amostras para os estimadores K-fold ($kf10$, $bkf10$, $rkf10$ e $rbkf10$ e Leave-One-Out (loo) no Modelo 1*

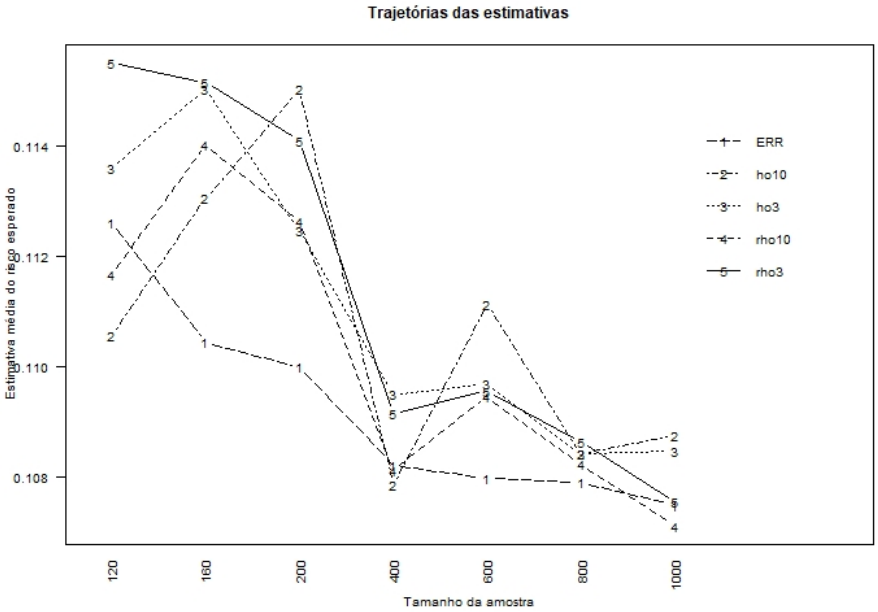

Figura 5.7: *Trajetória das médias dos valores obtidos nas 100 amostras para os estimadores Hold-out ($ho3$, $ho10$, $rho3$ e $rho10$ no Modelo 1*

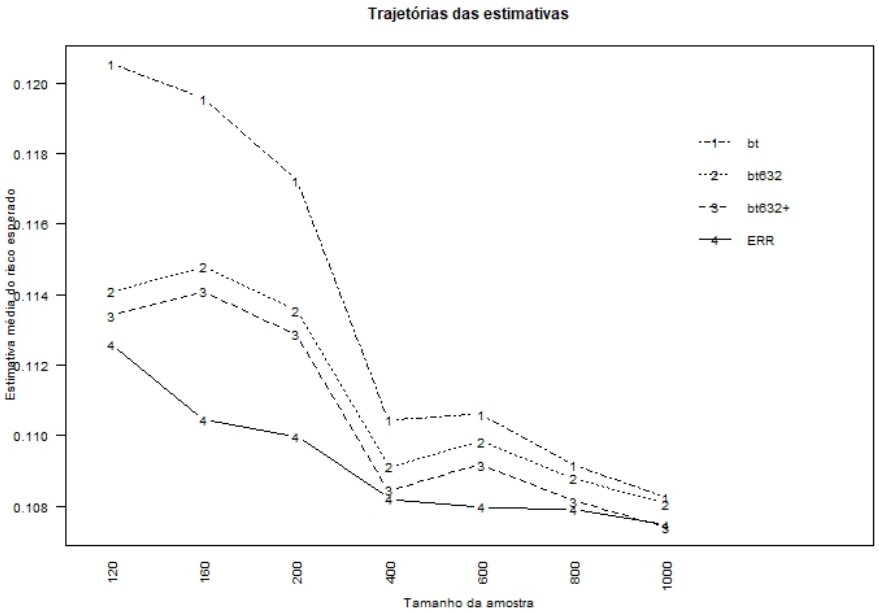

Figura 5.8: *Trajetória das médias dos valores obtidos nas 100 amostras para os estimadores Bootstrap (bt, bt*632*, bt*632+*) no Modelo 1*

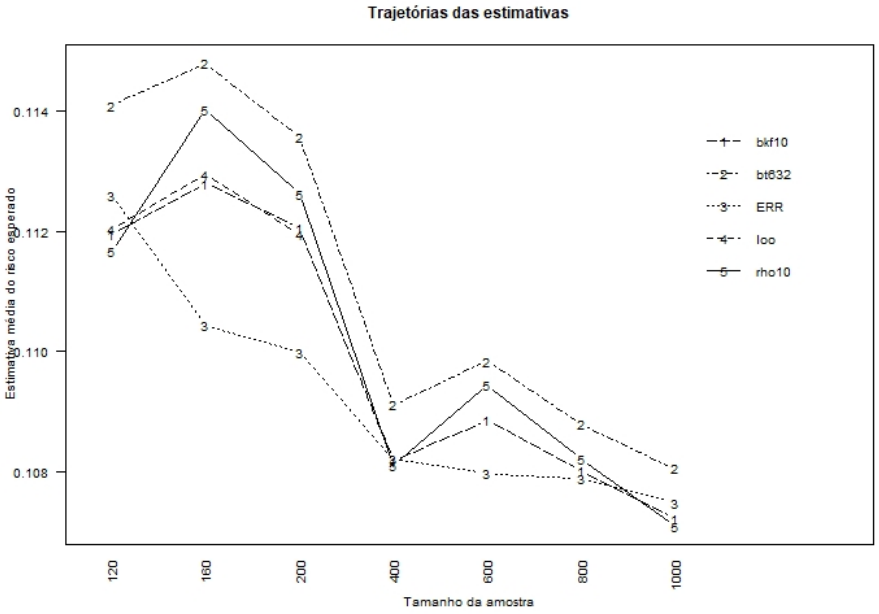

Figura 5.9: *Trajetória das médias dos valores obtidos nas 100 amostras para os estimadores mais assertivos no Modelo 1*

5.3.2 Modelo 2: Normal Misto

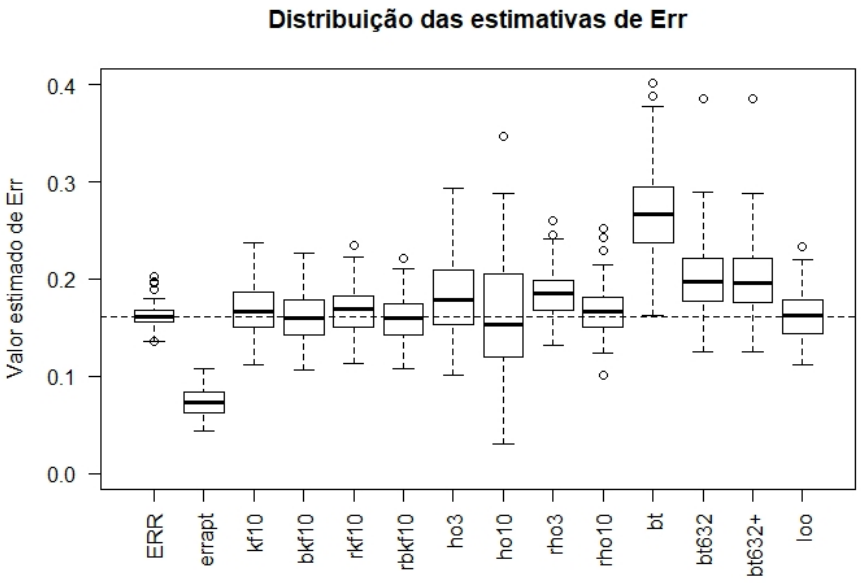

Figura 5.10: *Distribuição dos estimadores de Err para n=120 no Modelo 2*

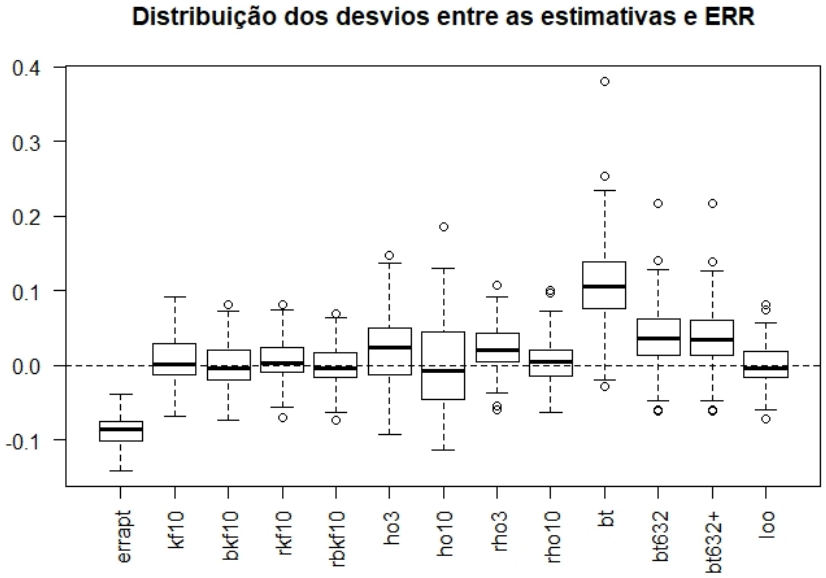

Figura 5.11: *Distribuição das diferenças entre o valor de ERR e do valor obtido pelos 13 estimadores para n = 120 no Modelo 2*

Na **Figura** 5.10 temos a distribuição dos estimadores do valor de Err para as 100 amostras de tamanho $n = 120$ no modelo 2. Comparando com o modelo com efeitos fixos, podemos ver que o estimador *errapt* fica mais subviesado e *bt* fica bem acima do esperado. Os estimadores $bt632$ e $bt632+$ também possuem distribuições com valores bem acima da distribuição de ERR, tendo uma performance pior que no caso de efeitos fixos. Os estimadores $ho3$ e $ho10$ possuem as maiores variâncias entre as estimativas. Suas versões com repetições $rho3$ e $rho10$ possuem as menores variâncias, embora $rho3$ esteja superviesado. Os demais estimadores apresentam distribuições bem próximas entre si.

Ao analisar a distribuição dos desvios dos valores obtido pelos estimadores para o valor de ERR por meio da **Figura 5.11**, vemos novamente a evidência do viés forte para valores negativos da técnica *errapt*. As técnicas $ho3$, $rho3$, bt, $bt632$ e $bt632+$ tem um tendência de viés positivo, sendo bt o que tem o maior viés. As demais técnicas possuem a mediana próxima ao valor 0.

A mesma análise do comportamento da média $e\hat{r}r$ das 100 simulações para os tamanhos de amostras analisadas é feita aqui, com a separação da estimativas nas **Figuras 5.12** a **5.15** de acordo com a proximidade das técnicas. O comportamento do viés médio (5.1) pode ser analisado na **Tabela 5.4** e o comportamento das variâncias (5.2) é mostrada na **Tabela 5.5**.

Podemos notar que os estimadores possuem vieses mais agravados (**Tabela 5.4**) e maior varibilidade (**Tabela 5.5**) nos modelos normais mistos do que quando aplicados ao modelo normal com efeito fixos (**Tabela 5.2** e **Tabela 5.3**).

O estimador *errapt* permanece sendo o estimador mais subviesado (**Figura 5.12** e **Tabela 5.4**), tendo esse comportamento agravado quando comparado ao modelo de efeitos fixos.

O estimador por *Bootstrap* é o mais superviesado entre os estimadores (**Tabela 5.4**), sendo que, diferente do caso de efeitos fixos, as ponderações $bt632$ e $bt632+$ também se tornam superviesados nas amostras até 200 observações. Os estimadores $ho3$ e $rho3$ também apresentaram viéses positivos elevados em todos os tamanhos de amostras realizados, enquanto $ho10$ e $rho10$ superviesam apenas nas amostras até 200 observações (**Figura 5.14**)

Na **Figura 5.13** vemos que, assim como no caso de efeitos fixos, os estimadores $kf10$ e $rkf10$ apresentam estimativas equivalentes, como também acontece com $bkf10$ e $rbkf10$. Os estimadores com correção de Burman apresentam valores mais próximos do valor de ERR.

Com relação a variabilidade (**Tabela 5.5**), vemos que $ho10$ apresenta a maior variabilidade entre os estimadores em todos os tamanhos de amostras realizados, sendo que $ho3$ também mantém uma alta variabilidade em todos os tamanhos de amostras. bt, $bt632$ e $bt632+$ apresentam alta variabilidade nas amostras menores, mas ela diminui ao patamar dos demais estimadores a medida que o tamanho de amostra aumenta.

Concluindo, vemos que $bkf10$ e loo são as estimadores mais assertivos nas amostras de tamanho até 200 observações (**Figura 5.16**). Nas amostras maiores, $rho10$, $bkf10$, $bt632+$ e $bt632$ são bons estimadores para o risco esperado.

Técnica	120	160	200	400	600	800	1000
errapt	-87.454	-66.817	-56.654	-32.399	-22.786	-18.558	-15.149
kf10	7.285	6.794	6.733	1.314	1.804	0.583	0.630
bkf10	-0.048	1.264	1.974	-1.06	0.112	-0.710	-0.407
rkf10	6.647	8.228	7.494	1.314	1.810	0.412	0.642
rbkf10	-0.739	2.636	2.694	-1.057	0.112	-0.864	-0.393
ho3	25.261	20.819	19.385	6.683	6.358	3.825	3.974
ho10	1.911	13.333	9.926	0.753	2.970	1.101	1.130
rho3	23.584	22.927	19.682	7.754	6.381	3.954	3.428
rho10	5.696	7.500	7.292	1.043	2.085	0.707	0.443
bt	110.985	63.655	50.370	22.487	16.782	11.578	9.989
bt632	39.591	16.907	12.132	3.174	3.040	1.26	1.485
bt632+	38.471	15.850	11.120	2.325	2.239	0.499	0.745
loo	2.022	4.310	3.414				

Tabela 5.4: *Viés médio ($\times 10^{-3}$) dos 13 estimadores testados no Modelo 2*

Técnica	120	160	200	400	600	800	1000
errapt	21.666	23.020	14.844	6.331	4.048	2.745	2.621
kf10	70.075	38.947	28.498	9.893	5.078	3.797	3.118
bkf10	64.001	37.466	26.658	9.459	4.943	3.693	3.070
rkf10	57.981	39.032	27.732	9.410	5.062	3.638	3.081
rbkf10	52.665	37.563	26.041	8.995	4.954	3.554	3.046
ho3	228.376	143.397	90.545	30.492	22.880	11.723	9.539
ho10	368.637	340.255	205.845	72.142	53.643	29.923	30.223
rho3	64.437	40.203	31.634	10.94	5.416	3.943	3.297
rho10	67.072	42.655	31.963	9.668	5.113	4.297	3.148
bt	318.349	67.977	41.550	12.166	6.230	4.415	3.489
bt632	147.314	44.613	28.502	9.356	5.246	3.694	3.137
bt632+	147.001	44.145	28.172	9.238	5.18	3.647	3.098
loo	58.812	36.441	27.035				

Tabela 5.5: *Variância ($\times 10^{-5}$) dos 13 estimadores testados no Modelo 2*

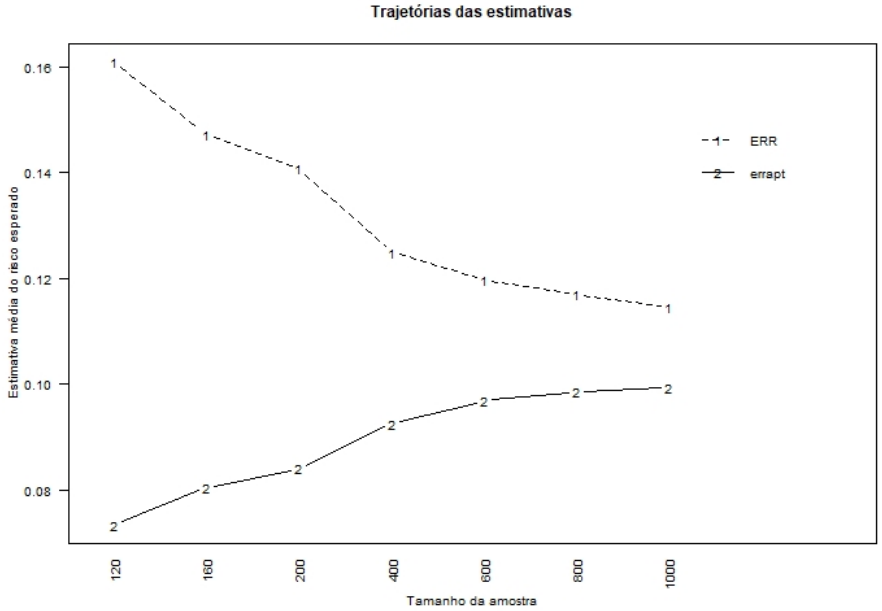

Figura 5.12: *Trajetória das médias dos valores obtidos nas 100 amostras para o estimador errapt no Modelo 2*

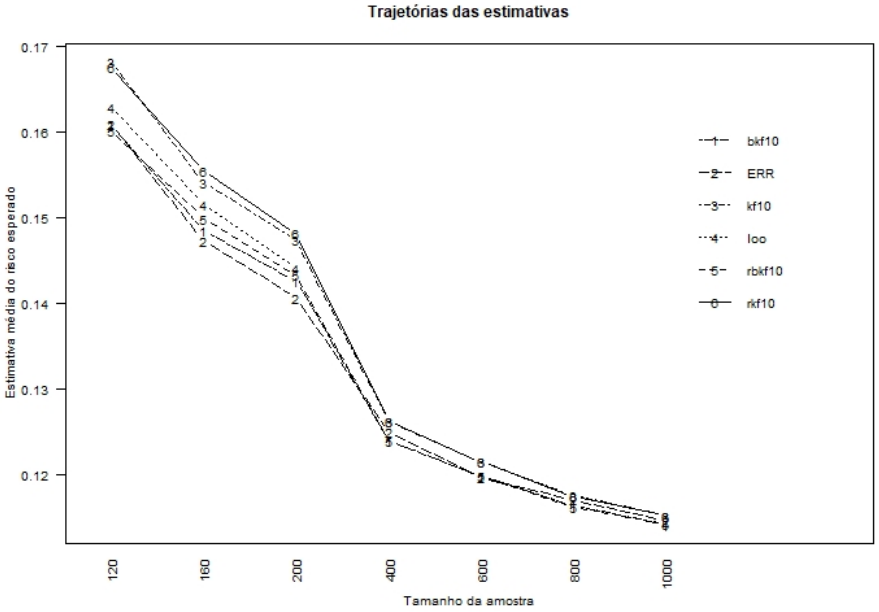

Figura 5.13: *Trajetória das médias dos valores obtidos nas 100 amostras para os estimadores K-fold ($kf10$, $bkf10$, $rkf10$ e $rbkf10$) e Leave-One-Out (loo) no Modelo 2*

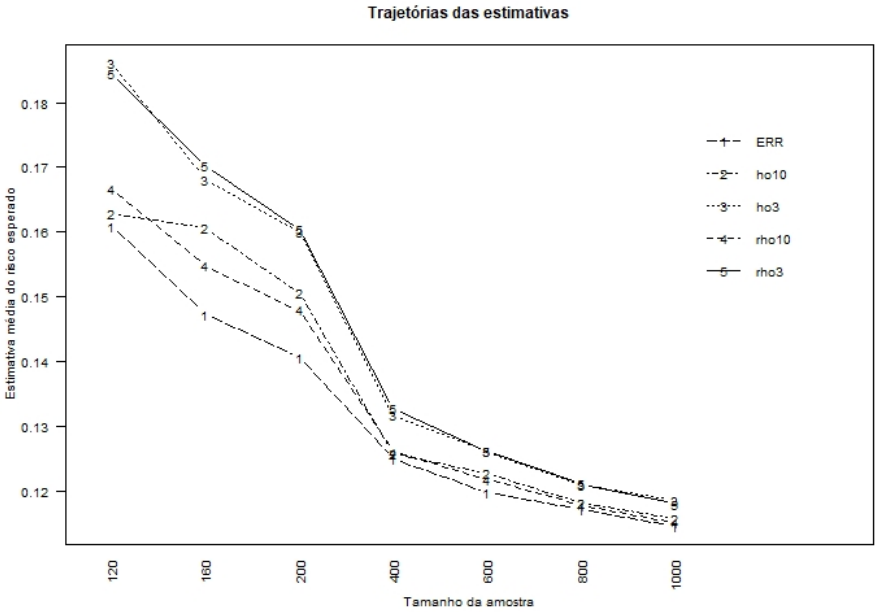

Figura 5.14: *Trajetória das médias dos valores obtidos nas 100 amostras para os estimadores Hold-out ($ho3$, $ho10$, $rho3$ e $rho10$) no Modelo 2*

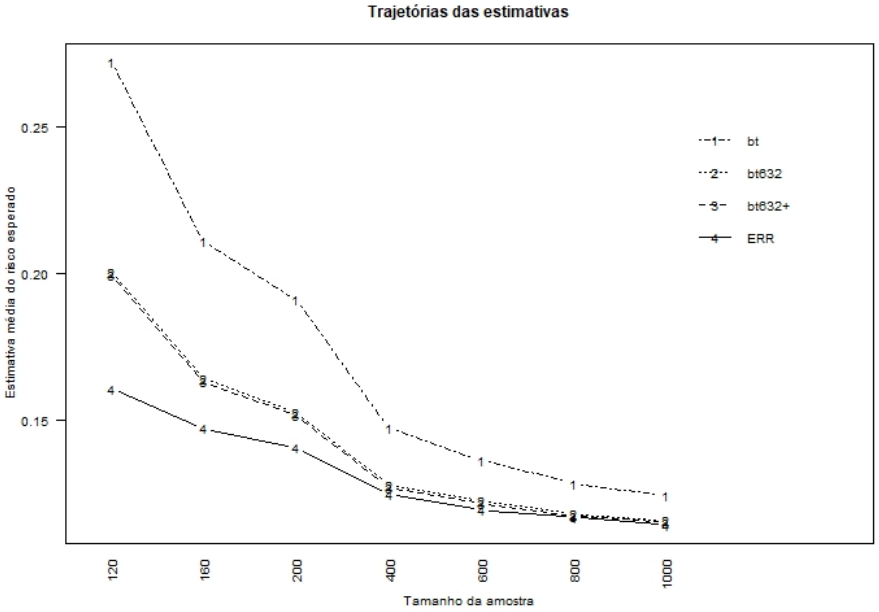

Figura 5.15: *Trajetória das médias dos valores obtidos nas 100 amostras para os estimadores Bootstrap (bt, bt*632*, bt*632*+) no Modelo 2*

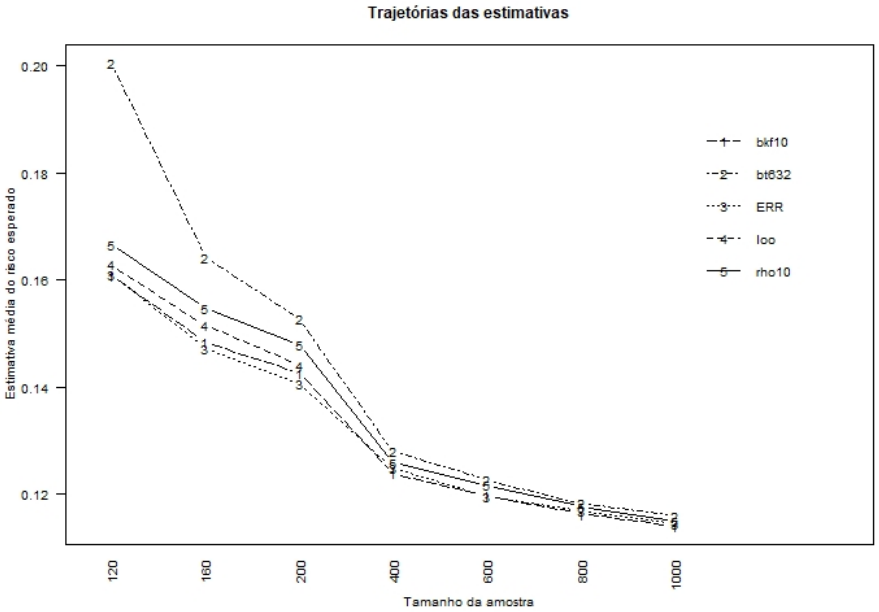

Figura 5.16: *Trajetória das médias dos valores obtidos nas 100 amostras para os estimadores mais assertivos no Modelo 2*

5.3.3 Modelo 3: Logístico com efeitos fixos

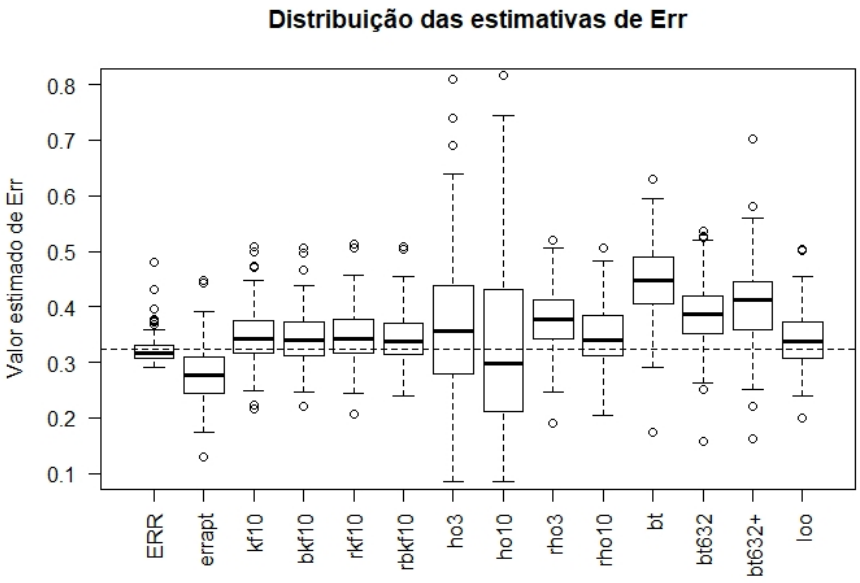

Figura 5.17: *Distribuição dos estimadores de Err para n = 120 no Modelo 3*

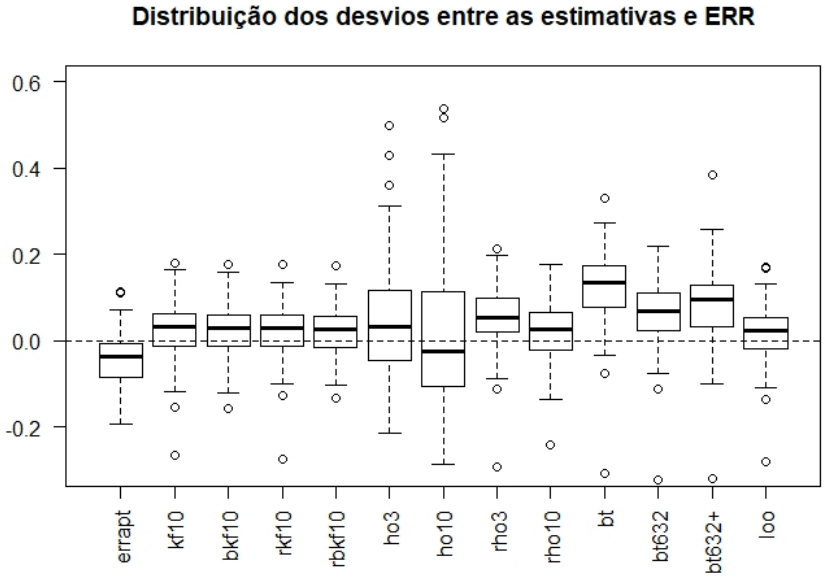

Figura 5.18: *Distribuição dos desvios entre o valor de ERR e dos valores obtidos pelos 13 estimadores para n = 120 no Modelo 3*

A **Figura 5.17** mostra o comportamento da distribuição das estimativas do valor de (2.7) para as 100 amostras de tamanho $n = 120$, indicando também a variabilidade das técnicas. A linha tracejada na figura indica o valor da média de ERR.

Com relação ao viés de cada distribuição, *errapt* apresenta uma distribuição de valores abaixo das demais, enquanto a maior parte das distribuição das estimativas estão deslocadas para cima com relação a média de ERR, sendo que as estimativas obtidas por *bootstrat* (bt, $bt632$ e $bt632+$) evidenciam distribuições mais elevadas, com uma variabilidade um pouco maior que as demais.

Ainda com relação a variabilidade, as distribuições de $ho10$ e $ho3$ demonstram ter as maiores variâncias.

A **Figura 5.18** indica a distribuição dos vieses entre os valores obtidos pelos estimadores e o valor de ERR para as 100 amostras de tamanho $n = 120$. A estimativa *errapt* subestimou o valor do risco esperado em quase todas as simulações realizadas. Os estimadores *bootstrap* bt, $bt632+$ superestimaram a medida, enquanto a versão $bt632$ está mais próxima. O estimador $rho3$ também superestimou os valores na maior parte das vezes. A estimativa $ho10$ embora apresente a mediana próxima de zero, possui as maiores dispersões, ou seja, existem amostras em que o estimador superestima severamente e em outras subestima severamente. Já as distribuições dos vieses das técnicas K-fold ($kf10$, $bkf10$, $rkf10$ e $rbkf10$), assim $rho10$ e loo possuem distribuições muito similares entre si e as medianas levemente acima da linha do zero.

As médias $\bar{err}$ das 100 amostras de cada estimador estão representadas nas **Figuras 5.19** a **5.22**. A separação em 4 gráficos foi realizada para facilitar a visualização das trajetórias de cada estimativa. A **Figura** 5.23 mostra um resumo das melhores técnicas, após a análise geral.

Além disso, a **Tabela 5.6** apresenta o valor do viés médio de cada estimador, obtido através de (5.1) e a **Tabela**. **5.7** apresenta as variâncias das estimadores, que foram calculadas por (5.2).

O estimador *errapt* é o que apresenta o menor viés entre todas técnicas em todos os tamanhos de amostra estudados (Na **Figura 5.19** e **Tabela 5.6**), sendo esse viés bem forte nas amostras de tamanho menor.

O estimador bt é mais superviesado em todos os tamanhos de amostra (**Tabela 5.6**). Outros estimadores com vieses positivos elevados são $bt632$, $bt632+$, $ho3$ e $rho3$ na maioria das amostras realizadas.

Entre os estimadores pelo método *K-fold*, percebemos algumas particularidades: $kf10$ e $rkf10$ estimam, em média, quase sempre o mesmo valor (**Figura 5.20**) e seus vieses médios são bem próximos (**Tabela 5.6**), o que nos leva a concluir que a repetição não traz grandes ganhos em estimar o risco esperado. A mesma conclusão podemos chegar quando olhamos as linhas $bkf10$ e $rbkf10$. Outra informação importante é que os estimadores com correção de Burman são menores e mais próximos de ERR na maioria dos tamanhos de amostra realizados.

Com relação a variabilidade dos estimadores, $ho10$ possui os maiores valores de variância em todos os tamanhos de amostras, exceto nas simulações de tamanho 120, onde $ho3$ teve a maior variância (**Tabela 5.7**). Essa alta variabilidade de $ho10$ provoca uma certa instabilidade na estimativa média (**Figura 5.21**), fazendo que o estimador ora superviese ERR, ora subviese. Os estimadores por *bootstrap* também possuem uma alta variabilidade em amostras pequenas. Os demais estimadores possuem valores de variância próximos.

Podemos concluir que, para amostras pequenas no caso de modelos logísticos, os estimadores mais indicados são $rho10$, $bkf10$ e loo (**Figura 5.23**), sendo $bkf10$ a que possui os menores vieses (**Tabela 5.6**). Para as amostras acima de 400 observações, $rho10$ e $bkf10$ apresentaram uma melhor performance.

Técnica	120	160	200	400	600	800	1000
errapt	-46.221	-38.201	-32.425	-15.389	-6.650	-3.954	-5.011
kf10	23.894	10.687	5.368	1.300	4.477	4.199	1.701
bkf10	22.734	7.921	3.253	0.401	3.883	3.765	1.345
rkf10	24.200	9.898	4.716	1.375	4.716	4.228	1.585
rbkf10	22.937	7.180	2.651	0.471	4.108	3.792	1.234
ho3	78.471	18.465	30.761	0.186	13.281	8.060	0.537
ho10	20.860	24.956	36.092	17.250	7.170	8.981	-1.023
rho3	53.981	29.409	15.815	5.293	7.366	5.999	2.432
rho10	25.869	8.288	3.299	1.115	5.783	3.428	1.224
bt	123.442	71.543	43.209	14.822	13.185	10.459	6.216
bt632	63.688	33.481	17.490	5.591	7.729	6.982	3.879
bt632+	80.935	38.951	20.299	4.615	6.257	5.368	2.207
loo	17.275	5.968	2.045				

Tabela 5.6: *Viés médio ($\times 10^{-3}$) dos 13 estimadores testados no Modelo 3*

Técnica	120	160	200	400	600	800	1000
errapt	2.991	1.864	1.860	0.872	0.474	0.344	0.309
kf10	3.005	2.098	1.722	0.892	0.471	0.346	0.314
bkf10	2.827	2.076	1.725	0.891	0.471	0.345	0.313
rkf10	2.729	1.832	1.758	0.877	0.480	0.345	0.310
rbkf10	2.534	1.835	1.764	0.876	0.48 0	0.345	0.310
ho3	69.984	9.536	10.256	2.452	1.842	1.083	0.851
ho10	32.141	33.755	21.651	9.560	5.187	3.711	3.253
rho3	3.17	1.993	1.958	0.900	0.500	0.354	0.338
rho10	3.709	2.157	2.134	1.053	0.589	0.394	0.349
bt	4.812	2.172	2.256	0.846	0.531	0.363	0.308
bt632	3.743	1.901	2.052	0.859	0.513	0.358	0.310
bt632+	6.359	2.175	2.043	0.848	0.508	0.354	0.307
loo	2.815	1.893	1.764				

Tabela 5.7: *Variância ($\times 10^{-3}$) dos 13 estimadores testados no Modelo 3*

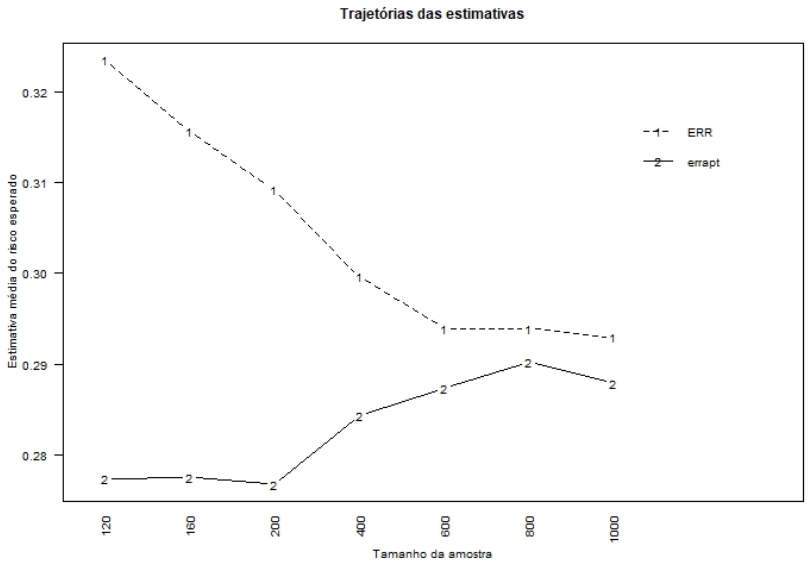

Figura 5.19: *Trajetória das médias dos valores obtidos nas 100 amostras para o estimador errapt no Modelo 3*

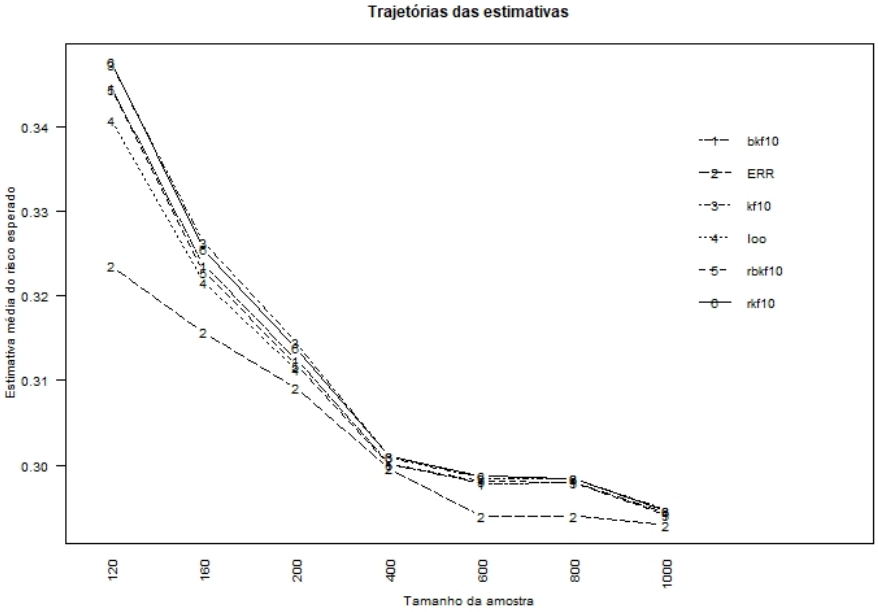

Figura 5.20: *Trajetória das médias dos valores obtidos nas 100 amostras para os estimadores K-fold ($kf10$, $bkf10$, $rkf10$ e $brkf10$) e Leave-One-Out (loo) no Modelo 3*

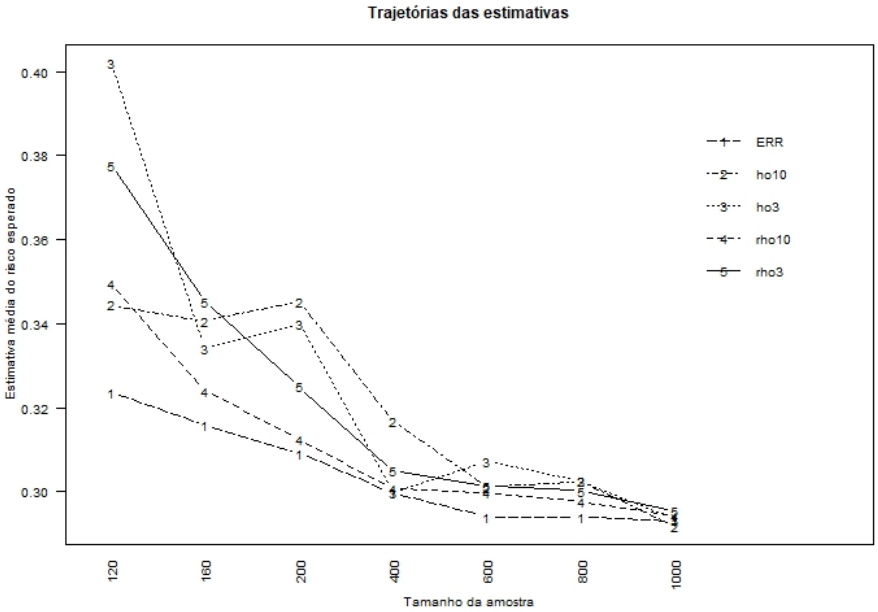

Figura 5.21: *Trajetória das médias dos valores obtidos nas 100 amostras para os estimadores Hold-out ($ho3$, $ho10$, $rho3$ e $rho10$) no Modelo 3*

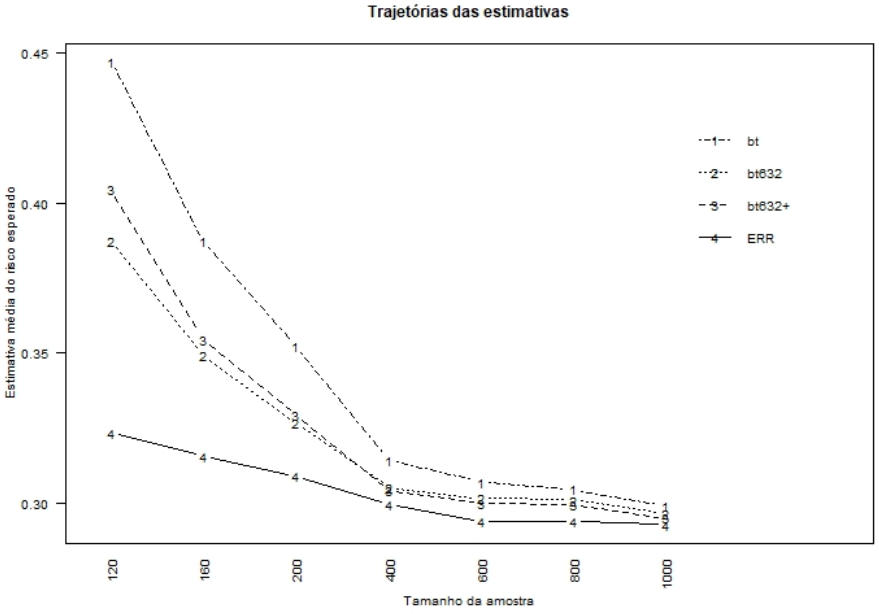

Figura 5.22: *Trajetória das médias dos valores obtidos nas 100 amostras para os estimadores Bootstrap (bt, bt*632*, bt*632*+) no Modelo 3*

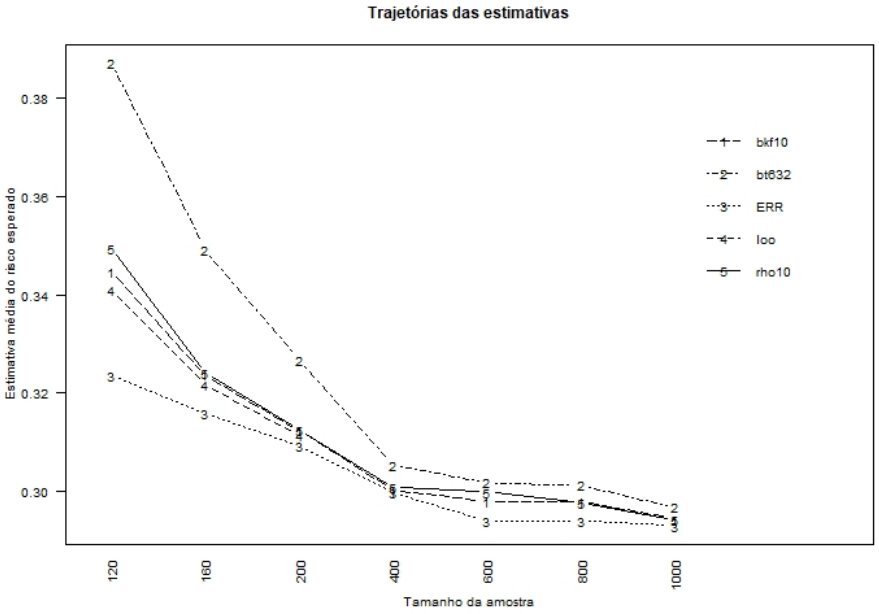

Figura 5.23: *Trajetória das médias dos valores obtidos nas 100 amostras para os estimadores mais assertivos no Modelo 3*

5.3.4 Modelo 4: Logístico Misto

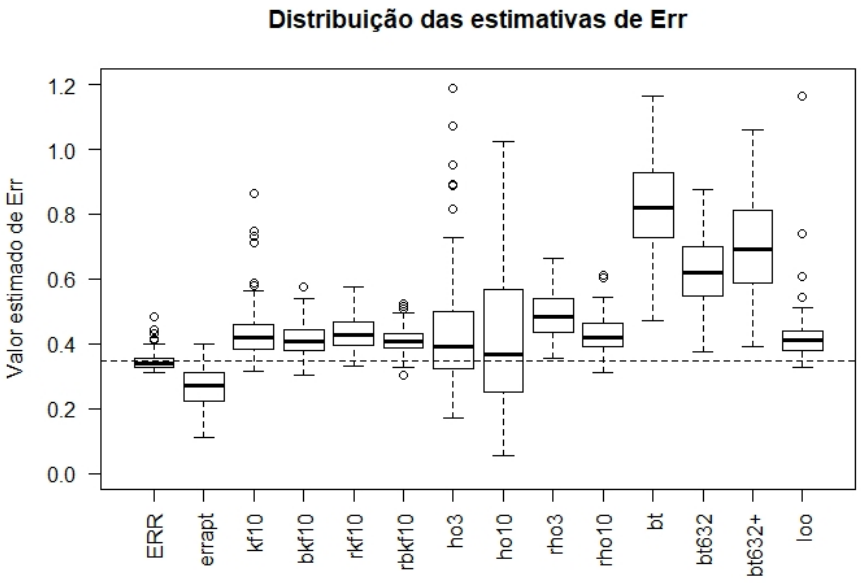

Figura 5.24: *Distribuição dos estimadores de Err para n = 120 no modelo 4*

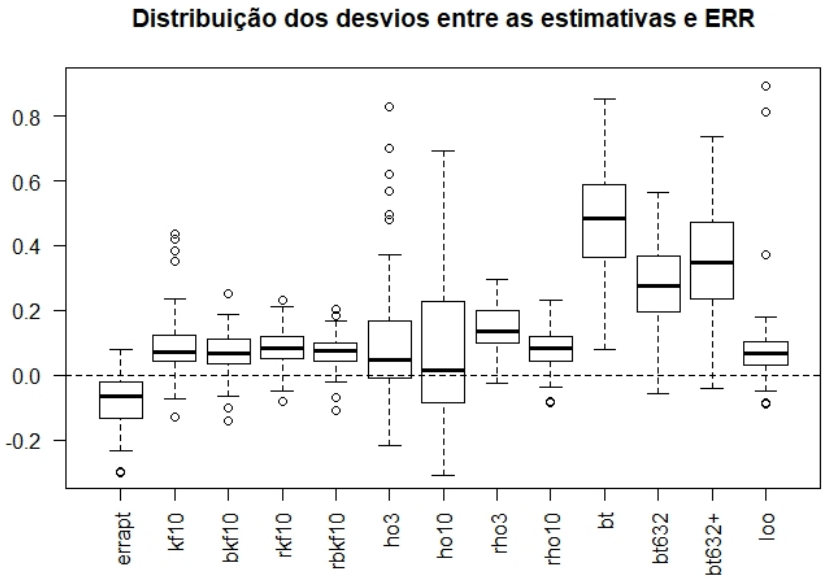

Figura 5.25: *Distribuição das diferenças entre o valor de ERR e o valor obtido pelos estimadores nas 100 amostras com n = 120 no Modelo 4*

Da mesma forma que fizemos nos modelos anteriores, a **Figura 5.24** representa a distribuição das estimativas de (2.7) para as 100 amostras de tamanho $n = 120$ no modelo 4. A linha tracejada na figura indica o valor da média de ERR.

Podemos notar que, em modelos logísticos mistos, todas as distribuições das estimativas estão acima da mediana, exceto *errapt* que está abaixo. Além disso, esse deslocamento para cima é maior do que o ocorrido no modelo com efeito fixo em todos todos os casos. As estimativas *bootstrap* superestimam severamente os valores, chegando a ser quase 3 vezes o valor de ERR em algumas

simulações. A distribuição de $bt632$ é a mais comportada dentre as três.

Com relação a variabilidade, a estimativa $ho10$ manteve-se como a distribuição com maior variância. $ho3$ tem uma variância bem menor do que a ocorrida no modelo com efeitos fixos, enquanto bt e $bt632+$ tiveram suas variâncias aumentadas. Os estimadores $kf10$, $bkf10$, $rkf10$, $rbkf10$, loo e $rho10$ apresentaram distribuições parecidas, porém no caso de loo, pontos extremos foram observados, indicando uma maior variabilidade.

Pela **Figura 5.25** vemos a distribuição dos vieses dos estimadores para as 100 amostras de tamanho $n = 120$ no modelo 4. Podemos notar que *errapt* possui quase todos os vieses negativos com relação a ERR, indicando a subestimação da estimativa. Já bt superestimou em todas as 100 simulações, enquanto $bt632$ e $bt632$ superestimaram em quase todas elas. Um outro ponto importante a se destacar é que $ho10$ tem a mediana dos desvios para ERR mais próxima de zero, porém sua dispersão é bem alta, ou seja, em alguns casos ela superestima em outros ela subestima.

As médias $\bar{err}$ das 100 amostras em cada caso estão representadas nas **Figuras 5.26** a **5.29**. A separação em 4 gráficos foi realizada para facilitar a visualização das trajetórias de cada estimativa. A **Figura** 5.30 mostra um resumo das melhores técnicas, após a análise geral.

O comportamento dos desvios médios pode ser analisado na **Tabela 5.8**, em que os valores foram obtidos através de (5.1), e as variâncias de cada estimativa, calculados por (5.2), são apresentados na **Tabela 5.9**.

O estimador *errapt* foi o único estimador que apresenta vieses negativos em todas as amostras analisadas (**Tabela 5.8**), sendo esse viés severo em todos os casos e, comparando-se com o modelo de efeitos fixos (**Figura** textbf5.19), o estimador *errapt* apresenta um comportamento (**Figura 5.26**) mais subviesado.

Os estimadores por *bootstrap* apresentam os maiores vieses positivos em todos os tamanhos de amostra realizados. Podemos apontar que no caso do modelo de efeitos fixos, $bt632$ e $bt632+$ se aproximam dos vieses dos demais estimadores (**Tabela 5.6**) nas amostras maiores, mas aqui ele se apresentaram superviesados mesmo nas amostras com mais observações (**Tabela 5.6** e **Figura 5.29**).

Todos estimadores possuem vieses bem elevados nas amostras de tamanho 120 e 160 (**Tabela 5.6**), com destaques para os estimadores bt, $bt632$, $bt632+$, $ho3$, $ho10$ e $kf10$ (nas amostras de tamanho 120).

Entre os estimadores pelo método *K-fold* podemos notar que, diferente do ocorrido no modelo com efeitos fixos, os estimadores $kf10$ e $rkf10$ deixaram de ser paralelas nas amostras pequenas, assim como também deixaram de ser paralelas as estimativas $bkf10$ e $rbkf10$. (**Figura 5.27**). Isso indica que, nesse caso, a repetição influencia na estimativa, embora não exista um indicativo que repetir o método melhore a estimação. Na comparação entre as técnicas com e sem correção de Burman, vemos que as que possuem a correção erram menos o valor de ERR nas amostras pequenas.

Na **Figura** 5.28, podemos destacar a instabilidade de $ho10$, assim como o ocorrido no modelo de efeitos fixos, o que o torna pouco recomendado para estimar o risco esperado.

Com relação a variabilidade dos estimadores (**Tabela 5.9**), vemos que $ho10$ continua sendo o estimador com os maiores valores de variância em todas os tamanhos de amostra, seguido por $ho3$. Os estimadores $kf10$, bt, $bt632$ e $bt632+$ possuem alta variabilidade nas amostras de tamanho 120, sendo que os estimadores por *bootstrap* mantém essa alta varibilidade nas amostras de tamanho 160

e 200.

Podemos notar que, nas menores amostras (120 e 160), nenhum estimador tem boa performance, pois todos possuem alto viés e alguns também possuem alta variabilidade. O fato de haver poucas observações faz com que haja poucas observações por indíviduo nas amostras. Isso faz com que a predição dos efeitos aleatórios seja menos precisa. Além disso, ao quebrar a amostra em base de treino e validação, pode acontecer de algum indivíduo não estar presente na amostra de treino. Como consequência, a predição usará apenas a parte fixa do modelo como preditora criando um viés maior. Quando as amostras aumentam de tamanho, ganhamos mais observações por indíviduo, tornando mais precisa as predições dos efeitos aleatórios.

Com essa visão em mente, podemos concluir que, no caso dos modelos com efeitos mistos, para amostras pequenas até 200 observações, os estimadores com melhores desempenhos são $rho10$, $bkf10$ e *loo* (**Figura** 5.30), embora apresentem viés alto (**Tabela 5.8**). Já em amostras acima de 200 observações, os estimadores mais assertivos são $rho10$, $bkf10$ (**Figura** 5.30), com $bkf10$ apresentando os menores vieses (**Tabela 5.8**).

Técnica	120	160	200	400	600	800	1000
errapt	-82.955	-94.498	-73.242	-38.424	-25.079	-23.891	-15.332
kf10	89.846	34.427	13.123	2.901	2.638	1.529	2.511
bkf10	69.462	24.725	11.983	0.816	1.465	0.062	1.798
rkf10	83.193	27.528	11.343	2.677	2.944	1.225	2.816
rbkf10	72.886	24.082	12.325	0.573	1.841	-0.311	2.004
ho3	159.514	35.961	28.841	7.550	0.279	2.060	2.873
ho10	131.474	42.259	6.552	7.996	20.000	-0.730	3.898
rho3	142.180	60.277	32.617	8.107	5.882	3.713	3.848
rho10	82.006	28.341	7.198	2.169	4.212	0.682	2.096
bt	451.391	247.847	165.044	72.892	44.804	30.955	25.067
bt632	258.627	126.074	80.340	34.270	21.240	12.835	12.210
bt632+	371.785	160.311	94.074	35.404	20.447	11.613	10.651
loo	83.486	20.795	7.711				

Tabela 5.8: *Viés médio ($\times 10^{-3}$) dos 13 estimadores testados no Modelo 4*

Técnica	120	160	200	400	600	800	1000
errapt	3.879	3.399	2.486	0.697	0.486	0.465	0.396
kf10	7.853	4.157	1.663	0.663	0.383	0.385	0.329
bkf10	2.639	2.417	1.597	0.659	0.400	0.387	0.338
rkf10	2.542	2.205	1.312	0.680	0.377	0.376	0.323
rbkf10	2.103	1.809	1.467	0.670	0.386	0.378	0.334
ho3	158.541	23.266	10.449	2.780	1.531	1.197	1.142
ho10	224.822	39.478	20.061	8.905	4.697	4.104	3.092
rho3	4.188	2.693	1.762	0.726	0.426	0.397	0.340
rho10	3.224	2.233	1.928	0.762	0.463	0.416	0.365
bt	19.236	12.884	5.613	0.950	0.442	0.418	0.327
bt632	10.224	6.846	3.447	0.753	0.416	0.412	0.343
bt632+	41.740	12.099	4.456	0.784	0.415	0.410	0.339
loo	16.718	2.682	1.377				

Tabela 5.9: *Variância ($\times 10^{-3}$) dos 13 estimadores testados no Modelo 4*

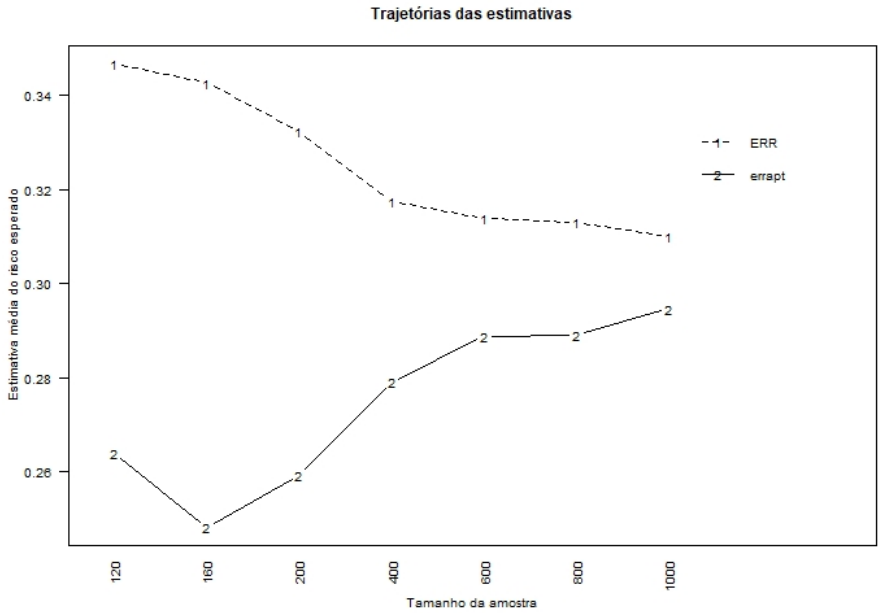

Figura 5.26: *Trajetória das médias dos valores obtidos nas 100 amostras para o estimador errapt no Modelo 4*

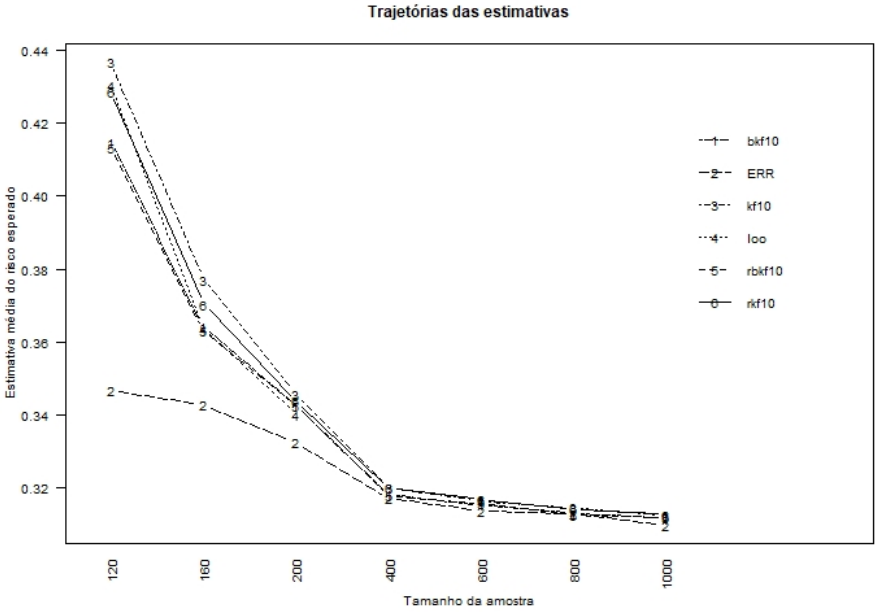

Figura 5.27: *Trajetória das médias dos valores obtidos nas 100 amostras para os estimadores K-fold ($kf10$, $bkf10$, $rkf10$ e $rbkf10$) e Leave-One-Out (loo) no Modelo 4*

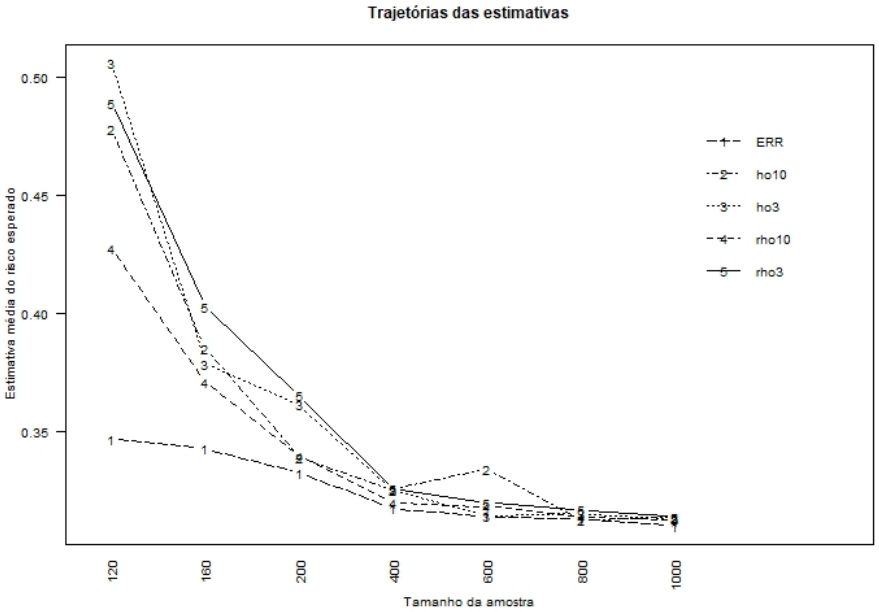

Figura 5.28: *Trajetória das médias dos valores obtidos nas 100 amostras para os estimadores Hold-out ($ho3$, $ho10$, $rho3$ e $rho10$ no Modelo 4*

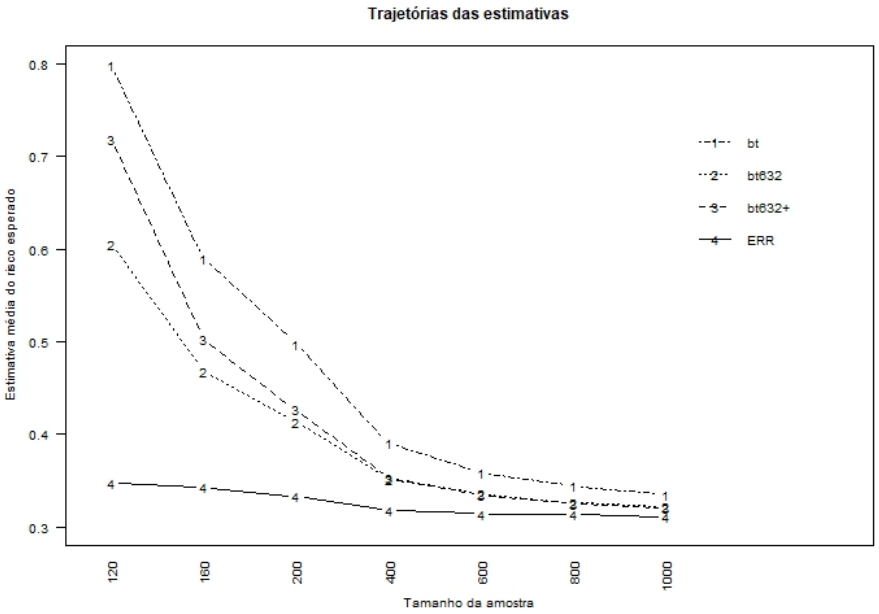

Figura 5.29: *Trajetória das médias dos valores obtidos nas 100 amostras para os estimadores Bootstrap (bt, bt*632*, bt*632*+) no Modelo 4*

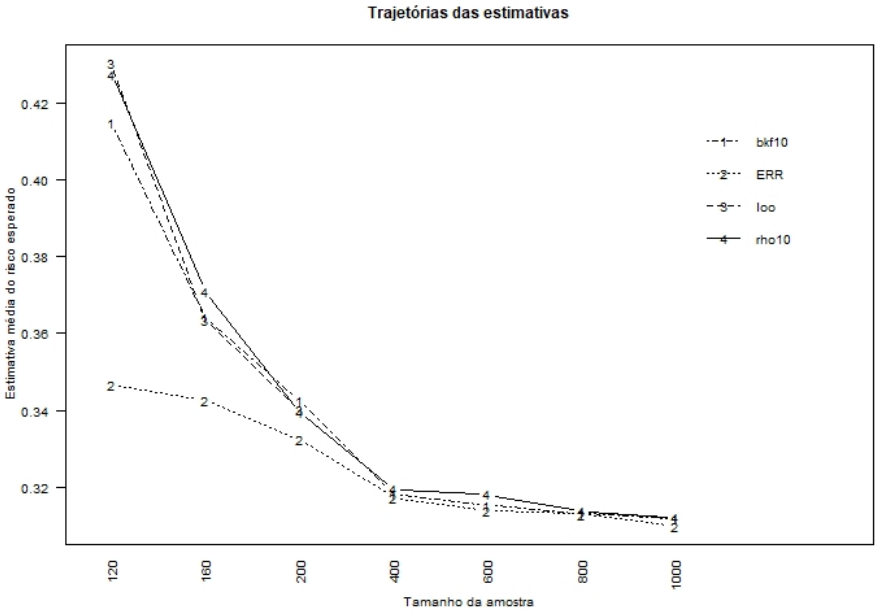

Figura 5.30: *Trajetória das médias dos valores obtidos nas 100 amostras para os estimadores mais assertivos no Modelo 4*

Capítulo 6

Aplicação

Neste capítulo iremos aplicar as técnicas de validação apresentadas no Capítulo 3 numa base de um estudo da área biológica. As etapas desse capítulo são: Descrição do estudo e da amostra, definição do modelo (preditor) $\hat{f}(X)$ a ser analisado e o seu valor estimado $\hat{f}(X)$, aplicação das técnicas de validação cruzada para calcular o erro esperado em se predizer a variavel resposta usando $\hat{f}(X)$ e, por fim, a discussão dos resultados.

6.1 Descrição da amostra

O estudo que iremos usar nese trabalho foi apresentado em Nelder e Wedderburn (1972) e está representada na amostra **salamandras**. Trata-se de um experimento com duas populações de salamandras que, no meio ambiente, estão separadas geograficamente uma da outra. As espécies analisadas são a *Rough Butt* (RB) e *Whiteside* (WS). O objetivo do estudo é verificar se essa separação geográfica contribui para que o acasalamento entre indivíduos de espécies iguais tenha maior probabilidade de ocorrência do que o acasalamento entre espécies diferentes. Foram usadas 10 salamandras de cada combinação Sexo (Macho ou Fêmea) e Espécie (RB ou WS), num total de 40 salamandras. Foram realizados 3 experimentos, sendo um na época do verão e dois no inverno, todos no mesmo ano. Porém, nesse estudo, foi ignorado o fato que os mesmos animais foram usados em experimentos diferentes, ou seja, foi considerado que, em cada experimento, foram usadas 20 salamandras de cada gênero e que os experimentos são independentes entre si. Cada fêmea foi pareada com 6 machos, sendo 3 machos da mesma espécie e 3 machos da outra espécie, num total de 120 pareamentos em cada experimento realizado, ou seja, não houve todos os cruzamentos possíveis. A variável resposta do modelo indica 1, quando ocorre acasalamento entre duas salamandras e 0, caso contrário. As covariáveis usadas são Sexo, Espécie e Experimento (que leva em consideração também a estação do ano).

6.2 Modelo proposto

Como o objetivo da análise é avaliar a chance de acasalamento entre espécies e cada salamandra pode ter um efeito diferente no acasalamento, usaremos como modelo proposto um modelo logístico misto com dois efeitos aleatórios.

Seja Y_{ijk} a variável resposta que indica se houve acasalamento entre as salamandras fêmea (i) e macho (j) no experimento k. Seja b^f_{ik} o efeito aleatório da i-ésima salamandra fêmea no acasalamento, $i = 1, 2, ..., 20$, $k = 1, 2, 3$ e b^m_{jk} o efeito aleatório do j-ésimo macho no acasalamento, $j = 1, 2, ..., 20$, $k = 1, 2, 3$. Então $Y_{ijk}|b^f_{ik}, b^m_{jk} \sim Bernouli(f_{ijk}(\mathbf{X}))$, com $b^f_{ik} \sim N(0, \sigma^2_f)$ e $b^m_{jk} \sim N(0, \sigma^2_m)$ e

$$\nu_{ijk} = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{21i} + \beta_4 x_{22j} + \beta_5 x_{21i} x_{22j} + b^f_{ik} + b^m_{jk},$$

$i = 1, 2, ..., 20$;
$j = 1, 2, ...20$;

$k = 1, 2, 3$

$$f_{ijk}(X) = exp(\nu_{ijk})/(1 + exp(\nu_{ijk}))$$

em que x_{11} indica o experimento realizado no verão (1, se verão e 0 , caso contrário); x_{12} indica o primeiro experimento do inverno (1, se primeiro experimento do inverno e 0, caso contrário); x_{21i} a espécie da salamandra fêmea (1, se WS e 0, caso contrário); x_{22j} indica a espécie da salamandra macho (1, se WS e 0, caso contrário) e $x_{21i}x_{22j}$ indica a interação entre as espécies. Os parâmetros $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ são os parâmetros de efeitos fixos. Os efeitos aleatórios b^f_{ik} e b^m_{jk} são os interceptos aleatórios correspondentes a cada animal usado nos experimentos.

Analisando descritivamente os dados, dos 360 pareamentos realizados no estudo, em 52.5% dos casos houve acasalamento.

Com relação aos experimentos, houve 58.3% de sucesso nos acasalamentos do experimento 1 (no Verão), 49.2% de sucesso no experimento 2 (Primeiro experimento do Inverno) e 50% de sucesso no experimento 3 (Segundo experimento do Inverno). Ou seja, a estação do ano indica que no Verão ocorrem mais acasalamentos.

Na **Tabela 6.1** temos a proporção de acasalamento entre cada combinação possível de Sexo e Espécie, em que a sigla WSF indica uma salamandra Whiteside fêmea, WSM uma Whiteside macho, RBF uma Rough Butt Fêmea e RBM uma Rough Butt macho.

	WSF	**RBF**
WSM	66.7%	55.6%
RBM	21.1%	66.7%

Tabela 6.1: *Proporção de acasalamentos ocorridos em cada tipo de pareamento*

Podemos notar na **Tabela** 6.1 que os pares das mesmas espécies tiveram a mesma proporção de acasalamentos ocorridos. No geral, pares de espécies diferentes tiveram menos acasalamentos, com destaque para o caso RB Macho e WS Fêmea, com uma proporção de 21.1% de acasalamentos ocorridos com sucesso.

Usando o função *glmer* do pacote *lme*4, com método de estimação a Aproximação de Laplace, obtivemos as estimativas dos parâmetros $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \sigma^2_m, \sigma^2_f)$ do modelo misto proposto de acordo com a **Tabela 6.2**

Parâmetro	Estimativa	Erro Padrão	P-valor
$\hat{\beta}_0$	0.71018	0.47530	0.135
$\hat{\beta}_1$	0.44609	0.54552	0.414
$\hat{\beta}_2$	-0.07114	0.54191	0.896
$\hat{\beta}_3$	-0.57558	0.42316	0.174
$\hat{\beta}_4$	-2.42655	0.47133	$<$0.001
$\hat{\beta}_5$	2.98229	0.51921	$<$0.001
$\hat{\sigma}^2_m$	0.9382		
$\hat{\sigma}^2_f$	1.0555		

Tabela 6.2: *Estimativas dos parâmetros do modelo misto*

Substituindo as estimativas dos parâmetros e as predições dos interceptos aleatórios, podemos calcular $\hat{\nu}_{ijk}$ e $\hat{f}_{ijk}(\mathbf{X})$, obtendo assim como preditor do valor esperado de Y_{ijk}

$$\hat{Y}_{ijk} = \hat{f}_{ijk}(\mathbf{X}).$$

Pelo contexto do problema se entende que o modelo misto é o mais adequado para analisar essa base de dados, como tem sido feito em diversos artigos. Espera-se que isto esteja refletido em termos de predição, por isso analisamos também o modelo apenas com efeitos fixos, isto é:

$$\nu_{2ijk} = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \beta_3 x_{21i} + \beta_4 x_{22j} + \beta_5 x_{21i} x_{22j}$$

$i = 1, 2, ..., 20;$
$j = 1, 2, ...20;$
$k = 1, 2, 3$

$$f_{2ijk}(X) = exp(\nu_{2ijk})/(1 + exp(\nu_{2ijk})).$$

Neste caso, foi usado a função *glm* do pacote *stats*, com método de estimação da máxima verossimilhança, e foi obtido as estimativas dos parâmetros $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ de acordo com a **Tabela 6.3**.

Parâmetro	Estimativa	Erro Padrão	P-valor
$\hat{\beta}_0$	0.58146	0.27434	0.034
$\hat{\beta}_1$	0.39351	0.28161	0.162
$\hat{\beta}_2$	-0.03857	0.27774	0.890
$\hat{\beta}_3$	-0.47411	0.30959	0.126
$\hat{\beta}_4$	-2.02902	0.34353	$<$0.001
$\hat{\beta}_5$	2.50312	0.46266	$<$0.001

Tabela 6.3: *Estimativas dos parâmetros do modelo com apenas efeitos fixos*

Podemos notar, na comparação dos parâmetros do modelo misto (**Tabela 6.2**) com os parâmetros do modelo apenas com efeitos fixos (**Tabela 6.3**) que o sentido das variáveis de efeitos fixo permanece o mesmo, mas houve mudança nos valores dos parâmetros.

As estimativas do risco esperado segundo cada técnica de validação cruzada obtidas nos dois modelos estão apresentadas na **Tabela 6.4**. Nessa tabela podemos observar que todas as estimativas do risco esperado para o modelos misto são inferiores as do modelo apenas com efeito fixos, com exceção de *bt*. Isso evidencia a importância de um seleção de modelos adequada.

Podemos notar que as estimativas do modelo misto indicam um erro esperado entre 0.54 a 0.72 (exceto *errapt*) aproximadamente, e as estimativas do modelo fixo do erro esperado indicam aproximadamente um valor entre 0.61 a 0.74.

O estimador *errapt* estima um risco esperado bem abaixo das outras estimativas no modelo misto, enquanto no modelo fixo o valor é a menor estimativa, porém seu valor obtido é próximo das demais estimativas. Isso indica que seu uso não é apropriado, pois o seu valor pode indicar um risco esperado menor que o real.

Técnica	Estimativa Modelo Misto	Estimativa Modelo Fixo
errapt	0.3973574	0.6148331
kf10	0.5845433	0.6345492
bkf10	0.5669108	0.6334956
rkf10	0.5787065	0.6326995
rbkf10	0.5607359	0.6317495
ho3	0.5913844	0.6940065
ho10	0.5449735	0.7417377
rho3	0.6041084	0.6847131
rho10	0.5651045	0.6766745
bt	0.7221565	0.6361828
bt632	0.6069634	0.6321432
bt632+	0.6664749	0.6290064
loo	0.5644003	0.6319995

Tabela 6.4: *Estimativas do risco esperado das 13 técnicas de validação para o modelo misto e para o modelo apenas com efeitos fixos criados com a amostra* ***salamandras***

A estimativa obtida por $ho10$ é o segundo valor mais baixo dentre as valores no modelo misto enquanto é o maior entre os valores do modelo fixo, o que evidencia sua instabilidade, ora subestimando, ora superestimando. As estimativas obtidas por $ho3$ tem valor alto em ambos os modelos, o que pode indicar um viés positivo da técnica nas duas situações. O estimador $rho10$ estima um valor um pouco acima dos demais no modelo de efeitos fixos, o que indicaria um viés, embora na simulação apresentada no **Capítulo 5** ela tenha tido uma performance equivalente às melhores estimativas.

Entre os estimadores por *bootstrap* podemos destacar que, no modelo misto, o estimador bt obteve um valor bem afastado dos demais, comprovando seu forte viés positivo. As ponderações $bt632+$ e $bt632$ apresentam valores menores que bt, mas ainda um pouco acima dos demais.

A aproximação entre os estimadores $kf10$ e $rkf10$, assim como de $bkf10$ e $rbkf10$, evidencia que a repetição do método *K-fold* traz poucos ganhos em melhora da estimativa, tornando-se um custo computacional que poderia ser evitado. Os estimadores com correção de Burman ($bkf10$ e $rbkf10$) trouxeram valores menores do que os estimadores correspondentes ($kf10$ e $rkf10$, respectivamente) indicando que a correção é uma boa escolha porque reduz o viés do estimador.

Por fim, se analisarmos os valores obtidos pelas estimativas que melhor performaram na simulação, a estimativa do erro esperado do modelo misto pode ser indicada pelas estimativas $bkf10$, $rho10$ e loo. As três indicam um valor próximo de 0.56. Esse valor é o menor entre os valores se considerarmos que $errapt$ é subviesado e $ho10$ é instável. É importante destacar que loo perde para os demais em eficiência computacional, pois demora mais para ser calculada.

Portanto, com essa base de dados real pode-se avaliar e também conferir o desempenho das técnicas de validação cruzada e sua relação com a escolha do modelo mais apropriado à situação.

Capítulo 7

Conclusão

Neste trabalho foi apresentada uma revisão das principais técnicas de validação cruzada existentes e a eficiência delas em estimar o risco esperado de um modelo. Na apresentação das técnicas no **Capítulo 3** foram resumidas as principais conclusões de eficiência dessas técnicas na estimação do risco esperado em outros artigos, quando foram aplicadas em modelos de regressão de efeitos fixos e modelos de classificação. O estudo dessas técnicas no contexto dos modelos mistos é a principal contribuição desse trabalho, confirmando muitas das conclusões apresentadas na literatura, mas trazendo algumas especificidades.

Os estimadores relacionados aos métodos *K-fold* e *Leave-One-Out* apresentaram no geral boas estimativas do risco esperado. O fator de repetir o método *K-fold* não mostrou trazer melhoras na estimação, obtendo-se, em média, quase sempre um valor muito próximo da versão sem repetição, com variabidades semelhantes nas duas situações. Já as versões do método com a correção de Burman trouxeram, em geral, valores melhores de estimativas do risco esperado do que as versões sem correção, sendo o estimador *Burman 10fold* ($bkf10$) o que apresentou os melhores resultados em termos de redução de viés da estimativa do risco esperado. A mudança no valor de K não foi avaliada nesse trabalho e fica como sugestão para ser testada em trabalhos futuros. O estimador *Leave-One-Out* também apresentou vieses baixos, porém demandou um esforço computacional muito elevado em comparação com as demais técnicas e portanto não é recomendado.

Entre os estimadores pelo método *K-fold* foi notado que a aplicação do método sem repetição apresentou alta variabilidade para a estimativa, além de viés alto no caso de $p = \frac{1}{3}$ e instabilidade no valor da estimativa no caso de $p = \frac{1}{10}$. Portanto, a repetição do método é aconselhavél pois reduz a variabilidade, sendo que o estimador *Repeated Hold-Out* com $p = \frac{1}{10}$ se mostrou bem eficiente em estimar o risco esperado em modelos mistos.

Os estimadores pelo método *Bootstrap* trouxeram, no geral, vieses positivos elevados comparados com os estimadores pelos outros métodos. No contexto de modelos mistos estes vieses foram agravados quando comparados com a aplicação em modelos de efeitos fixos. Além disso, no caso do modelo Logístico Misto, esses estimadores apresentaram também variabilidade maior do que a maioria dos estimadores. Portanto, é desaconselhado o uso dos estimadores pelo método *Bootstrap* no caso de modelos mistos.

Podemos concluir que, nesse contexto dos modelos mistos abordados, os estimadores *Burman 10fold* e *Repeated Hold-Out* com $p = \frac{1}{10}$ são os mais indicados para a estimação do risco esperado, trazendo resultados próximos ao valor real, com baixo viés e semelhança entre suas variabilidades.

Para finalizar, acreditamos que este trabalho possa servir como base para o estudo e aplicação destas técnicas em outros modelos mistos, os quais são uma ferramenta importante nas metodologias estatísticas mais recentes.

Referências Bibliográficas

Bates e Watts(2007) Douglas M. Bates e Donald G. Watts. *Nonlinear Regression Analysis and Its Applications*. Wiley. Citado na pág. 4

Borra e Ciaccio(2010) Simone Borra e Agostino Di Ciaccio. Measuring the prediction error. a comparison of cross-validation, bootstrap and covariance penalty methods. *Computational Statistics and Data Analysis*, 54:2976–2989. Citado na pág. 1, 2, 5, 6, 10, 11, 19

Breiman(1996) Leo Breiman. Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24 (6):2350–2383. Citado na pág. 10

Breiman *et al.***(1984)** Leo Breiman, Jerome Friedman, Charles J. Stone e R.A. Olses. *Classification and Regression Trees*. TaylorFrancis. Citado na pág. 4

Burman(1989) Prabir Burman. A comparative study of ordinary croos-validation, v-fold cross-validation and the learning-testing methods. *Biometrika*, 76:503–514. Citado na pág. 2, 10, 11

Cover e Hart(1967) Thomas Cover e Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27. Citado na pág. 4

Demidenko(2013) Eugene Demidenko. *Mixed Models: Theory and Applications with R*. Wiley. Citado na pág. 2, 15, 16, 18

Devroye e Wagner(1979) Luc Devroye e Thomas Wagner. Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory*, 25:601–604. Citado na pág. 2, 9

Efron(1983) Bladley Efron. Estimation the error rate of a prediction rule: improvement on cross-validation. *Journal of American Statistical Association*, 78:316–331. Citado na pág. 2, 12

Efron e Tibshirani(1995) Bladley Efron e Robert Tibshirani. *Cross-Validation and the Bootstrap: Estimating the error rate of a prediction rule*. Tese de Doutorado, Stanford University, California, Estados Unidos. Citado na pág. 6, 12, 13

Hastie *et al.***(2008)** Trevor Hastie, Robert Tibshirani e Jerome Friedman. *The Elements of Statistical Learning - Data Mining, Inference, and Predicion*. Springer. Citado na pág. 1, 5

Huberty(1975) Carl J. Huberty. Discriminant analysis. *Review of Education Research*, 45(4):543 – 598. Citado na pág. 4

James *et al.***(2013)** Gareth James, Daniela Wintten, Trevor Hastie e Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer. Citado na pág. 1, 2, 4, 8

Kim(2009) Ji-Hyun Kim. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computacional Statistics and Data Analysis*, 53:3735–3745. Citado na pág. 2, 10, 12, 19

Kohavi(1995) Ron Kohavi. Measuring the prediction error. a comparison of c. *Computacional Statistics and Data Analysis*, 1:2976–2989. Citado na pág. 9, 10

Kutner *et al.***(2004)** Michael H. Kutner, Christopher J. Nachtsheim, John Neter e William Li. *Applied Linear Statistical Models*. McGraw-Hill/Irwin. Citado na pág. 4, 14

McCulloch e Searle(2005) Charles E. McCulloch e Shayle R. Searle. *Generalized, Linear and Mixed Models*. Wiley. Citado na pág. 16, 18

Molinaro *et al.***(2005)** Annette M. Molinaro, Richard Simon e Ruth M. Pfeiffer. Prediction error estimation: a comparision of resampling methods. *Bioinformatics*, 21 (15):3301–3307. Citado na pág. 5

Murphy(2012) Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MITT Press. Citado na pág. 5

Nelder e Wedderburn(1972) John A. Nelder e Robert W. M. Wedderburn. Generalized linear models. *Journal of Royal Statistical Society*, 135:370 – 384. Citado na pág. 17, 46

Paula(2013) Gilberto A. Paula. *Modelos de Regressão com apoio computacional*. IME-USP São Paulo. Citado na pág. 18

Rodriguez *et al.***(2013)** Juan Diego Rodriguez, Aritz Pérez e Jose A. Lozano. A general framework for the statistical analysis of the source of variance for classification error estimators. *Pattern Recognition*, 46:855–864. Citado na pág. 2, 10

Searle *et al.***(1992)** Shayle R. Searle, George Casella e Charles E. McCulloch. *Variance Components*. Wiley. Citado na pág. 16

Stone(1974) Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of Royal Statistical Society*, 36(2):111–147. Citado na pág. 2

Wong(2015) Tzu-Tsung Wong. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recpgnition*, 48:2839–2846. Citado na pág. 8, 11