UNIVERSIDADE FEDERAL DE JUIZ DE FORA

INSTITUTO DE CIÊNCIAS EXATAS

ESTÁTISTICA

José Roberto Peres Junior

Aplicação de Modelo de Regressão Logística Multinomial em Predição

Juiz de Fora

2021

José Roberto Peres Junior

Aplicação de Modelo de Regressão Logística Multinomial em Predição

Trabalho de Conclusão de Curso apresentado ao curso de Estatística da Universidade Federal de Juiz de Fora como requisito para obtenção do título de Bacharel em Estatística

Orientador: Prof. Lupércio França Bessegato

Juiz de Fora

2021

José Roberto Peres Junior

Aplicação de Modelo de Regressão Logística Multinomial em Predição

Trabalho de Conclusão de Curso apresentado ao curso de Estatística da Universidade Federal de Juiz de Fora como requisito para obtenção do título de Bacharel em Estatística

Aprovado em

BANCA EXAMINADORA

_____

Dr. Lupércio França Bessegato - Orientador

Universidade Federal de Juiz de Fora

_____

Dr. Marcel de Toledo Viera

Universidade Federal de Juiz de Fora

_____

Dr. Augusto Carvalho Souza

Universidade Federal de Juiz de Fora

RESUMO

A análise de regressão é uma ferramenta estatística muito utilizada em modelos preditivos, entretanto, vários dos problemas apresentados possuem como resposta de interesse variáveis categóricas multinomiais. Nesses casos, a regressão logística multinomial, com função de ligação logit, é uma ferramenta que pode ser usada na construção de modelos preditivos. A seleção das variáveis explicativas do modelo de regressão logística multinomial é uma etapa importante na aplicação da técnica. Há vários procedimentos para que essa tarefa seja efetuada, sendo elas automatizadas ou não. Então, mesmo com várias alternativas disponíveis, optamos por utilizar neste trabalho o método *stepwise* e o método não automatizado proposto por Hosmer e Lemershow (1989) aplicados no conjunto de dados penguins, disponível no R.

Palavras-chave: Regressão logística multinomial; Variáveis categóricas multinominais; Logit; Conjunto de dados penguins; Seleção de variáveis.

.

ABSTRACT

Regression analysis is a statistical tool widely used in predictive models, however several of the problems presented have multinomial categorical variables as an answer of interest. In these cases, multinomial logistic regression, with a logit link function, is a tool that can be used in the construction of predictive models. The selection of the explanatory variables of the multinomial logistic regression model is an important step in the application of the technique. There are several procedures for this task to be performed, whether automated or not. Among other possibilities, we chose to use in this work the *stepwise* method and the non-automated method proposed by HOSMER E LEMESHOW (1989) applied to the penguins dataset, available in R.

Keywords: Multinomial logistic regression; Multinominal categorical variables; Logit; Penguin data set; Selection of variables.

AGRADECIMENTOS

Aos meus pais, Maristela e José Roberto, por todos os esforços que fizeram pra me dar suporte emocional, financeiro e serem minha inspiração de vida.

Ao meu amor, Ligia, que me “atura” e me completa, e que esteve ao meu lado nas alegrias e principalmente nos obstáculos ajudando-me à superá-los.

Aos meus amigos do início da longa caminhada que foi a faculdade, que carinhosamente construímos o ICEberg, mais que um clube, uma família.

Aos meus companheiros da Atlética, onde pude compartilhar momentos de vitórias, derrotas e de muito espírito esportivo universitário.

Aos professores do departamento de Estatística, em especial aos professores que contribuíram para minha formação acadêmica e humana.

À todos que de alguma forma me ajudaram a finalizar esse longo ciclo da minha vida.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

O estudo de estatística envolve diversas áreas do conhecimento, e tem o intuito de tentar responder diversos questionamentos como por exemplo a análise de regressão, ou seja, uma técnica de predição que consiste em descrever a relação entre uma variável de estudo e um conjunto de variáveis explicativas através de um modelo.

Os primeiros passos da regressão foram através do cientista Francis Galton em Galton (1889), quando estudou a relação entre altura dos filhos com a altura dos pais, estudo esse apontado como uma continuação ao estudo da Eugenia como cita De Souza (2016).

Na atualidade, a regressão está presente em vários estudos e em áreas diversas, como, por exemplo, na saúde, em estudos de fatores de risco como em Ribeiro (2009), e na economia, em análises de crédito como em Minussi (2002). Os exemplos são diversos e certamente muitos problemas que ainda não surgiram terão a regressão como uma possível ferramenta de análise. Alguns dos diferentes tipos de regressão são a linear, polinomial, quantílica, ordinal, de Poisson, de Cox e a logística que será o objetivo desse estudo.

A regressão logística é aplicada em dados que possuem como variável respostas variáveis categóricas. No decorrer do trabalho será mostrado que quando utilizamos a função de ligação logit, poderemos conseguir relacionar as variáveis explicativas com a variável resposta.

Outro ponto abordado é a diferença de métodos automatizados e não automatizados para a escolha das variáveis explicativas que serão utilizadas para o modelo preditivo. Dentre as diversas metodologias disponíveis, serão comparados o método *stepwise* e o não automatizado proposto em Hosmer e Lemeshow (1989), os quais, através dos dados estudados terão a sua acurácia medida e serão avaliadas as diferenças entre os modelos finais apresentados.

Para a exemplificação nesse trabalho, foram utilizados os dados “penguins” do pacote “palmerpenguins” do software R criado por Horst, Hill e Gorman (2020) que possuem três espécies como os elementos da variável resposta.

2 ANÁLISE DE REGRESSÃO

2.1 REGRESSÃO LOGÍSTICA

O modelo de Regressão Logística se diferencia dos demais porque apresenta a sua variável resposta como qualitativa nominal, podendo ser binária/dicotômica ou multinominal, além de ser bastante flexível, não impondo diversos pressupostos, tais como normalidade de resíduos e homogeneidade de variância como argumentam Barboza, Kimura e Altman (2017). Sendo assim, as variáveis explicativas podem ser tanto qualitativas como quantitativas.

O modelo de regressão logístico, de acordo com Hosmer e Lemeshow (1989) é definido como:

$$Y_i = \pi(\boldsymbol{x}_i) + \varepsilon_i \tag{1}$$

em que ε_i é o erro aleatório, $\boldsymbol{x}_i$ é o vetor das variáveis explicativas de cada indivíduo *i* e $\pi(x_i)$ é a probabilidade de sucesso definida abaixo. Consideramos $Y_i \sim Bernoulli\ (\pi(x_i))$, ou seja, a variável resposta Y_i assume o valor 1 para a ocorrência do evento de interesse e o valor 0 para o evento complementar. A probabilidade de sucesso é dada como:

$$P(\boldsymbol{Y}_i = 1\ |\boldsymbol{x}_i) = \pi(\boldsymbol{x}_i) = \pi_i$$

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_n x_{ip}}}{1+ e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}}} \tag{2}$$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \cdots, \beta_p)$ é o vetor de parâmetros desconhecidos a serem estimados. Por outro lado. a probabilidade de fracasso é:

$$P(\boldsymbol{Y}_i = 0\ |\boldsymbol{x}_i) = 1 - \pi(\boldsymbol{x}_i) = 1 - \pi_i$$

$$1 - \pi_i = \frac{1}{1+ e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\cdots+\beta_p x_{ip}}} \tag{3}$$

Dessa maneira, o valor esperado da variável resposta, dado os valores das variáveis explicativas, é o valor a ser modelado. Então temos que:

$$0 \leq E(Y_i|x_i) = \pi_i \leq 1$$

Devido a característica binária, precisamos de uma função de ligação para melhor interpretação do vetor β. Existem diversas funções de ligação como a probit, log natural, log-log, mas usaremos a transformação logit conforme Hosmer e Lemeshow (1989). Essta transformação é definida por:

$$g(x_i) = ln\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \quad (4)$$

A transformação logit é linear em β_0 e $\beta_j, j = 1, \cdots, p$, pode ser contínuo e variar de $(-\infty ; +\infty)$, dependendo dos valores das variáveis explicativas. O logit é o logaritmo da razão de chance, a qual é a razão entre a probabilidade de ocorrência do evento de interesse (sucesso), ou seja, π_i, em nosso modelo, e a probabilidade de seu complementar (insucesso), ou seja, 1 - π_i, em nosso modelo.

2.2 REGRESSÃO LOGÍSTICA MULTINOMIAL

A regressão logística multinominal é uma expansão da regressão logística binária. A grande diferença é que não possui mais sucesso e fracasso, e sim 3 ou mais respostas categóricas. Essa diferença é contornada com a escolha de uma das categorias como base de acordo com a explicação de Zelterman (2015) então, a categoria a ser escolhida pode ser qualquer uma. A definição dependerá da escolha do pesquisador, mas, independentemente da escolha, a resposta pode ser a mesma mudando os vetores β.

Escolhida uma categoria de referência, serão gerados pares de transformação logit, relacionando cada categoria de 1 até c, com a categoria base, da seguinte forma:

$$g_1(x) = ln\left[\frac{P(Y=1)}{P(Y=0)}\right] = \beta_{10} + \beta_{11} x_{i1} + \cdots + \beta_{1p} x_{ip}$$

$$g_2(x) = ln\left[\frac{P(Y=2)}{P(Y=0)}\right] = \beta_{20} + \beta_{21} x_{i1} + \cdots + \beta_{2p} x_{ip}$$

$$\vdots$$

$$g_c(x) = ln\left[\frac{P(\,Y = c)}{P\,(\,Y = 0)}\right] = \beta_{c0} + \beta_{c1}x_{i1} + \cdots + \beta_{cp}x_{ip}$$

As probabilidades de cada categoria são:

$$P(Y = 0\,|x_i) = \frac{1}{1 - e^{g_1(x)} + e^{g_2(x)} + \cdots + e^{g_c(x)}}$$

$$P(Y = 1\,|x_i) = \frac{e^{g_1(x)}}{1 - e^{g_1(x)} + e^{g_2(x)} + \cdots + e^{g_c(x)}}$$

$$P(Y = 2\,|x_i) = \frac{e^{g_2(x)}}{1 - e^{g_1(x)} + e^{g_2(x)} + \cdots + e^{g_c(x)}}$$

$$\vdots$$

$$P(Y = c\,|x_i) = \frac{e^{g_c(x)}}{1 - e^{g_1(x)} + e^{g_2(x)} + \cdots + e^{g_c(x)}}$$

Como o modelo multinominal é expansão do modelo binomial, iremos apresentar nos próximos pontos apenas as deduções para 1 par de categorias, pois para os outros pares as expressões poderão ser as mesmas.

2.3 ESTIMAÇÃO

A função de densidade do modelo de regressão logística é $Y_i \sim Bernoulli\ (\pi(x_i))$, com Y_i de 1 até n indivíduos independentes, relacionado com o vetor de variáveis explicativas x_i de cada unidade amostral. O método de máximo verossimilhança é utilizado para estimarmos o vetor de parâmetros $\beta = (\beta_0, \beta_1, \cdots, \beta_j)$ associado em cada variável explicativa e o intercepto.

A função de verossimilhança é dada por:

$$L(\beta) = \prod_{i=1}^{n} L_i(\beta) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i} \quad (5)$$

Aplicando o logaritmo natural na equação (5) temos que

$$l(\beta) = l(\beta|Y,X) = \sum_{i=1}^{n}\{y_i \ln \pi(x_i) + (1-y_i)\ln(1-\pi(x_i))\} \quad (6)$$

Derivando em relação aos parâmetros, para encontrar o vetor β que maximize a função, encontramos o seguinte sistema de equação:

$$\frac{\partial l(\beta)}{\partial \beta_0} = \sum_{i=1}^{n}[y_i - \pi(x_i)] = 0 \quad (7)$$

$$\frac{\partial l(\beta)}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}[y_i - \pi(x_i)] = 0 \quad (8)$$

em que $\pi(x_i)$ é apresentado na expressão (2).

As equações (7) e (8) são não lineares em β_0 e $\beta_j, j = 1, \cdots, p$, sendo necessário métodos interativos para a resolução do sistema. Neste caso proposto aqui foi utilizado o método de Newton Raphson, que podem ser encontrados com mais detalhes em Nunes (2011), Louzada e Diniz (2012) e Dantas e Desouza (2008) definido como:

$$\begin{aligned}\hat{\beta}^{(t+1)} &= \hat{\beta}^{(t)} + \left[I\left(\hat{\beta}\right)^{(t)}\right]^{-1} U(\hat{\beta}^{(t)}) \\ &= \hat{\beta}^{(t)} + \left[X^t \hat{V}^{(t)} X\right]^{-1} X^t\left(Y - \hat{\pi}^{(t)}\right), \quad t = 0,1,\cdots\end{aligned}$$

sendo que $\hat{\beta}^{(t)}$ e $\hat{\beta}^{(t+1)}$ os estimadores de β nos tempos t e t+1, U(β) é o vetor escore e *I*(β) é a matriz de informação de Fisher, que são os erros padrão para as estimativas, em que

$$U(\beta) = X^T Y - X^T \pi = X^T (Y - \pi)$$

$$I(\beta) = E\left[\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^T}\right] = X^T \hat{V} X$$

E X é uma matriz de dimensão n x (p+1) que contém os valores das variáveis explicativas, $\hat{V}$ uma matriz diagonal de dimensão n x n que contém os elementos $\hat{\pi}_i(1-\hat{\pi}_i)$, $Y = (y_1, y_2, \cdots, y_n)^T$ e $\pi = (\pi_1, \pi_2, \cdots, \pi_n)$, ou seja

$$X = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}, Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \; e \; \hat{V} = \begin{bmatrix} \hat{\pi}_1(1-\hat{\pi}_1) & 0 & \cdots & 0 \\ 0 & \hat{\pi}_2(1-\hat{\pi}_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \hat{\pi}_n(1-\hat{\pi}_n) \end{bmatrix}$$

As interações são repetidas até que uma regra de convergência adequada seja satisfeita, como, por exemplo $\left\|\hat{\beta}^{(t+1)} - \hat{\beta}^t\right\| < \varepsilon$. Valores iniciais são necessários para implementar esse algoritmo, portanto considera-se $\beta^{(0)} = 0_{(p+1)}$, vetor de zeros com dimensão (p + 1) x 1 Souza (2006).

2.4 TESTE DE SIGNIFICÂNCIA DO MODELO

Após ajustar o modelo de regressão, encontrando o vetor β, é iniciado o processo de avaliação do mesmo. Para isso, realizaremos um teste de hipótese que consiste em verificar quais as variáveis explicativas predizem bem a variável resposta, e quais não. A estatística de teste usada é a Razão de Verossimilhança, definida como:

$$\Lambda = -2 \ln \left[\frac{L(\widetilde{\beta})}{L(\widehat{\beta})}\right] = 2[l(\hat{\beta}) - l(\tilde{\beta})]$$

onde $\hat{\beta}$ é o vetor dos estimadores dos parâmetros do modelo de regressão logística com todas as variáveis explicativas (modelo saturado) e o $\tilde{\beta}$ é o vetor dos estimadores dos parâmetros do modelo de regressão logística com q variáveis retiradas (modelo ajustado). Essa estatística converge para uma distribuição qui-quadrado com os graus de liberdade dado pela diferença entre o número de parâmetros do modelo saturado e o número de parâmetros do modelo ajustado.

O teste de Razão de Verossimilhança pode ser utilizado para encontrar o modelo no qual as variáveis explicativas melhor descrevem a informação da variável resposta, com o menor número possível de variáveis. A estatística *Deviance* é uma

medida de qualidade de ajuste para modelos estatísticos. Ela generaliza a ideia de usar a soma de quadrados dos resíduos para os casos em que o ajuste é obtido por meio de procedimentos de máxima verossimilhança. Para o modelo de regressão logística, utiliza-se a expressão:

$$Deviance = -2\ln\left[\frac{L(\hat{\beta}_0, \cdots, \hat{\beta}_p)}{L(y_1, \cdots, y_n)}\right]$$

em que $L(y_1, \cdots, y_n)$ corresponde à verossimilhança do modelo saturado e $L(\hat{\beta}_0, \cdots, \hat{\beta}_p)$ corresponde à verossimilhança do modelo ajustado, em que o modelo saturado atribui à componente sistemática toda a variação dos dados e não fornece, portanto, qualquer simplificação . Por outro lado, a *Deviance* pode ser também escrita como:

$$\begin{aligned} D &= -2\sum_{i=1}^{n}[y_i \ln(\hat{\pi}_i) + (1-y_i)\ln(1-\hat{\pi}_i)] - \sum_{i=1}^{n}[y_i \ln(y_i) + (1-y_i)\ln(1-y_i)] \\ &= -2\sum_{i=1}^{n} y_i[\ln(\hat{\pi}_i) - \ln(y_i)] + (1-y_i)[\ln(1-\hat{\pi}_i) - \ln(1-y_i)] \\ &= -2\sum_{i=1}^{n}\left[y_i ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i) ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)\right] \end{aligned}$$

Assintoticamente, a *Deviance* tem uma distribuição qui-quadrada com n – p graus de liberdade. Quanto menor for o valor da *Deviance* melhor é o ajuste.

Outro parâmetro utilizado para interpretar o quanto o modelo está ajustado em regressão é o pseudo R^2. Como na regressão logística a variável resposta não é contínua, o R^2 não pode ser calculado. Porém podemos calcular um R^2 ajustado ou pseudo R^2 usando a modificação proposta por Cox e Sneel (1989), definida como:

$$R^2 = 1 - \left(\frac{L(\beta)_0}{L(\beta)_M}\right)^{\frac{2}{n}}$$

em que $L(\beta)_0$ é a função de verossimilhança sem variáveis explicativas, $L(\beta)_M$ é a função de verossimilhança do modelo estimado e n o tamanho da amostra.

De acordo com Lattin, Carroll E Green (2011), o Pseudo R2 é uma medida interpretável como a fração da incerteza explicada por um modelo em relação ao modelo sem nenhuma variável explicativa, que é o de incerteza máxima.

Diferente do R^2, o valor máximo não chega a 1. Devido a falta de um valor de referência na literatura, o valor encontrado serve para medir o desempenho de modelos concorrentes, em que se preferirá o modelo com maior valor.

Outra estatística também utilizada para medir a eficiência do modelo é o AIC-Critério de Informação de Akaike, proposto em Akaike (1974). A estatística AIC busca penalizar modelos que apresentam um elevado numero de variáveis priorizando modelos de complexidade reduzida e com predição satisfatória. A AIC não é uma estatística proveniente de um teste de hipótese, ela utiliza a razão da logverossimilhança. A medida AIC é definida como:

$$\mathrm{AIC} = -2\left\{\ln\left[\frac{L(\widetilde{\beta})}{L(\widehat{\beta})}\right] - p\right\}$$

em que p é o número de parâmetros do modelo.

Os modelos de regressão que apresentam um melhor ajuste são os modelos que possuírem o menor AIC, como definiu Alvarenga (2015). Portanto, encontrar o modelo com menor AIC é um critério de parada para métodos de seleção de variáveis automáticos.

2.5 INTERPRETEÇÃO DOS COEFICIENTES

Na regressão logística, os coeficientes não possuem interpretações diretas, pois estão presentes dentro da função de transformação logit. Ou seja, os coeficientes representam a inclinação, ou a taxa de mudança da função, para cada incremento de unidade do valor dessa variável com as demais explicativas fixas, sendo assim representada:

$$\beta_j = g(x_i + 1) - g(x_i)$$

em que a função g(.) é a função descrita (4), logo β_j pode ser interpretado como o logaritmo da razão de chances e $\exp(\beta_j)$ o valor da razão de chances, também denominada de "*odds ratio*".

A razão de chances é uma medida de associação que relaciona uma probabilidade de "sucesso" com uma probabilidade de "fracasso" como define Szumilas (2010). Como não pretendemos estudar os efeitos causais das variáveis explicativas na resposta, não interpretaremos isoladamente os coeficientes β deste modelo de regressão logística, o qual foi construído com o objetivo de predizer a resposta.

Usando a expressão para o modelo de regressão logística definido em (2) e (3) a razão de chances é definida como:

$$\varphi = \frac{\left(\frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)}\right)\Big/\left(\frac{1}{1+\exp(\beta_0+\beta_1)}\right)}{\left(\frac{\exp(\beta_0)}{1+\exp(\beta_0)}\right)\Big/\left(\frac{1}{1+\exp(\beta_0)}\right)} = \frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$$

e o logaritmo da razão de chances é definido por:

$$\ln(\varphi) = \ln[exp(\beta_1)] = \beta_1$$

2.6 SELEÇÃO DA VARIÁVEIS

Um modelo de regressão ótimo busca otimizar a relação da variável resposta com as variáveis explicativas, ou seja, podem conseguir que o menor número de variáveis explicativas reflitam o máximo da variável resposta. Existem diversas formas de se fazer a seleção, dentre as quais definiremos dois métodos: o *stepwise* e o modo de 6 passos não automatizados adaptados das recomendações de Collett (2002) e Hosmer e Lemeshow (1989).

O método de seleção *Stepwise procedure* é a junção de dois procedimentos automáticos de seleção, a *Forward selection* e a *Backward elimination,* nos quais, a partir do modelo saturado, somam-se as variáveis que minimizam o valor AIC e, em sequência, retiram-se uma a uma as variáveis do modelo de menor AIC até que seja encontrado o modelo definitivo, ou seja, aquele com menor AIC. É um procedimento

automatizado, em que comparam-se o *Deviance* de cada modelo. Inicialmente o modelo utilizado é o sem variáveis explicativas como modelo de referência e compara-se com modelos que possuem uma variável explicativa. O modelo com menor *AIC* passa a ser o modelo de referência e se refaz as comparações. Agora o novo modelo de referência é comparado com modelos em que se integra mais uma variável explicativa. Esse procedimento vai se repetindo até que o modelo de referência possua o menor valor de *AIC.* Parando a seleção aqui, temos o método *Forward selection.* Entrando no modo de escolha *Backward elimination*, compara-se o AIC do modelo de referência retido no passo anterior com modelos em que se retira 1 variável explicativa. Se o valor da estatística da referência for o menor, ele será o modelo final, caso contrário, o modelo com menor *AIC* passa a ser o novo modelo de referência e repete-se o procedimento. O procedimento *stepwise* é automatizado

Entretanto, como salientado por Collett (2002), rotinas automatizadas não levam em consideração o conhecimento do pesquisador, podendo assim, excluir modelos igualmente bons que para a pesquisa podem apresentar maior relevância argumentativa. Sendo assim, seguindo as recomendações de Collett (2002) e Hosmer e Lemeshow (1989), são realizados os seis passos definidos abaixo para a obtenção do modelo de regressão logística final:

Passo 1: comparação entre o modelo com cada variável em relação ao modelo que não contém nenhuma variável. Todos os valores que apresentaram p-valor significativo passam a compor o novo modelo de referência.

Passo 2: com o novo modelo de referência, comparamos com modelos em que são retirados uma variável por vez. As variáveis que quando retiradas o modelo possui p-valor significativo, são retiradas da referência. Caso o pesquisador considere-as importantes, as mesmas podem permanecer.

Passo 3: comparando o novo modelo de referência com modelos onde serão incluídas as variáveis excluídas do passo 2 uma por vez, vamos certificar que essas variáveis não são significativas. Em caso de apresentarem p-valor significativo, elas voltam ao modelo de referência, caso contrário, são descartadas definitivamente.

Passo 4: de posse do modelo referência criado no passo 3, incluiremos as variáveis excluídas no passo 1 uma por vez. Todas variáveis que mostrarem p-valor significativo em relação ao modelo referência voltam para modelo, as que não mostrarem, são excluídas definitivamente.

Passo 5: compara-se o modelo referência com uma variável retirada. O elemento que ao ser excluído do modelo, tiver p-valor significativo, é retirado de vez. Caso o pesquisador sinta a necessidade de a mesma continuar, ela permanece. Também é possível que o pesquisador entenda que existem duas variáveis muito correlacionadas e decida pela retirada de uma delas, pois mesmo se mostrando estatisticamente significativa, ela pode ser retirada.

Passo 6: nesse passo incluímos interações entre as variáveis que ainda estão no modelo de referência. Caso alguma interação se mostre significativa ela é incluída no modelo, e assim temos o modelo final.

Em todo o procedimento descrito acima, são realizados testes da razão de verossimilhança. Salienta-se que, de acordo Allison (2014), o p-valor gerado associado a hipótese de verificação permite inferir se o modelo em análise é ou não um bom preditor.

Ambos os procedimentos resultam em modelos muito explicativos. Os passos não automatizados permitem ao pesquisador ser mais participativo, incluindo interações entre as variáveis e não excluindo variáveis que possuem relevância conceitual no modelo. Porém, em casos onde a quantidade de variáveis explicativas é muito grande, adotar uma rotina automatizada como a *stepwise procedure* pode ser mais vantajosa, pois a influência do pesquisador passa a ser menos relevante e a velocidade e a menor complexidade dos passos se mostra talvez a melhor escolha.

3 APLICAÇÂO

3.1 O BANCO DE DADOS PENGUINS

Penguins é um conjunto de dados disponibilizado pelo pacote palmerpenguins, da linguagem estatística R (HORST, 2020). Os dados do penguins foram coletados de 2007 a 2009 pela Dra Kristen Gorman na estação americana Palmer, contém 344 observações com 8 variáveis, sendo *Species* a variável resposta. *Adelie, Chinstrap e Gentoo* são as três espécies a serem estudadas. Tendo uma variável categórica nominal com três categorias. Utilizaremos o modelo de regressão logística multinominal para predição da variável resposta (*Species*).

3.2 ANALISE EXPLORATÓRIA DOS DADOS

O banco de dados possui 8 variáveis com 344 animais estudados, porém desse total, 2 animais foram excluídos por não ter sido possível coletar mais do que 3 variáveis. As variáveis são:

1 - *species*: É a variável resposta. É uma variável categórica nominal que representa as espécies de pinguins que foram coletadas. *Adelie, Chinstrap e Gentoo* são as espécies, contendo 152, 68 e 124 indivíduos, respectivamente;
2 – *island*.: É a variável que representa qual ilha do arquipélago de Palmer os pinguins foram coletados. É uma variável categórica nominal com três elementos: *Biscoe, Dream e Torgersen*, com 168,124 e 52 animais coletados, respectivamente;
3 – *sex*: A variável que indica o sexo de cada animal. Foram 168 machos, 165 fêmeas e 11 que não foi possível determinar o sexo;
4 – *year*: Representa qual ano o animal foi pesquisado. A pesquisa aconteceu nos anos 2007, 2008 e 2009 e foram pesquisados 110, 114 e 120 animais, respectivamente;
5 - *bill_length_mm*: representa o tamanho do bico do animal. É uma variável quantitativa contínua medida em milímetros, com média 43,92 mm, mediana 44,45 mm, mínimo 32,10 mm e máximo 56,90 mm;
6 - *bill_depth_mm*: explicita a profundidade do bico de cada pinguim. É uma variável quantitativa contínua medida em milímetros. Sua média é 17,15 mm, mediana 17,30 mm, mínimo e máximo 13,10 mm e 21,50 mm respectivamente;
7 - *flípper_length_mm*: O tamanho da asa também é uma variável quantitativa contínua medida em milímetros, com media 200,90 mm, mediana 197,00 mm e mínimo e máximo 172,00 mm e 231,00 mm;
8 - *body_mass_g*; define o peso de cada animal. Variável quantitativa contínua medida em gramas, possui média 4202 g, mediana 4050 g, mínimo 2700 g e máximo 6300 g

É importante ressaltar algumas interações entre nossas variáveis explicativas com a variável resposta, *species*. O banco possui três variáveis explicativas qualitativas *island., sex e year* e as interações são:

Tabela 1 - Interação espécie e ilhas

Espécie	Ilhas		
	Biscoe	*Dream*	*Torgersen*
Adelie	44	56	52
Chinstrap	0	68	0
Gentoo	124	0	0

Fonte: Elaborado pelo autor (2021).

Tabela 2 - Interação espécie e sexo

Espécie	Sexo	
	Feminino	Masculino
Adelie	73	73
Chinstrap	34	34
Gentoo	58	61

Fonte: Elaborado pelo autor (2021).

Tabela 3 - Interação espécie e ano

Espécie	Anos		
	2007	2008	2009
Adelie	50	50	52
Chinstrap	26	18	24
Gentoo	34	46	44

Fonte: Elaborado pelo autor (2021).

Por meio do teste de qui-quadrado nas tabelas de contingência, o teste de independência é realizado, o p-valor dos testes foram:

Tabela 4 - Resultados Teste Qui Quadrado

Interação	P-valor
Espécie x Ano	0,5224
Espécie x Sexo	0,976
Espécie x Ilha	2.2e-16

Fonte: Elaborado pelo autor (2021).

Com esses resultados, verificamos que não há evidências amostrais para rejeitar a hipótese de independência tanto da interação espécie e sexo, como da interação espécie e anos.há forte evidência amostral para rejeitar a hipótese de independência da interação espécie e ilha. Esses resultados indicam que a ilha em que o animal foi selecionado possui muita influência na variável *Species*, podendo assim ser uma variável explicativa importante ao estudo.

As variáveis quantitativas do banco, que medem o tamanho da asa, tamanho do bico, largura do bico e peso corporal, são variáveis que tratam da anatomia do animal. Naturalmente é importante verificar se existe alguma correlação entre elas. A matriz de correlação é dada abaixo:

Tabela 5 - Matriz de correlação das variáveis quantitativa

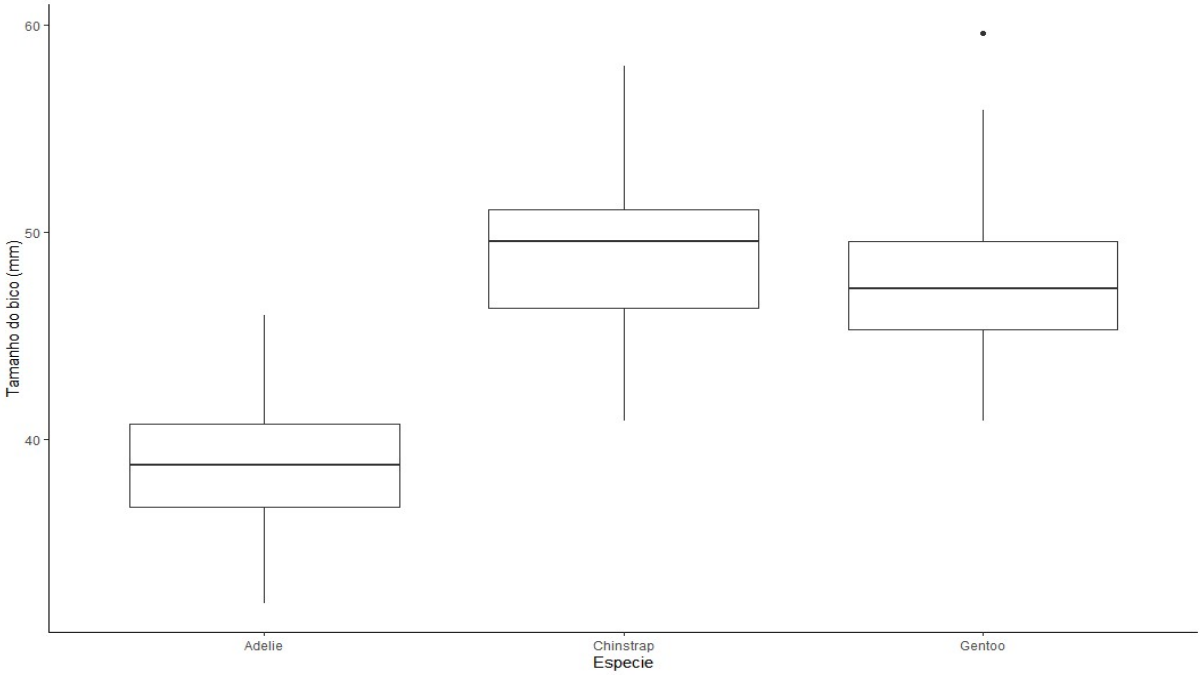
	bill_length_mm	bill_depth_mm	flípper_length_mm	body_mass_g
bill_length_mm	1,0000	-0,2351	0,6562	0,5951
bill_depth_mm	-0,2351	1,0000	-0,5839	-0,4749
flípper_length_mm	0,6562	-0,5839	1,0000	0,8712
body_mass_g	0,5951	-0,4749	0,8712	1,0000

Fonte: Elaborado pelo autor (2021).

Seguindo a definição de Mukaka (2012), uma correlação é considerada forte quando o seu valor em módulo é maior que 0,7. A única correlação forte apontada foi entre as variáveis tamanho da asa e peso corporal. Uma correlação entre -0,3 e 0,3 é considerada desprezível. A relação entre tamanho do bico e profundidade do bico, mesmo se tratando de medidas da mesma parte do corpo, obtiveram tal classificação.

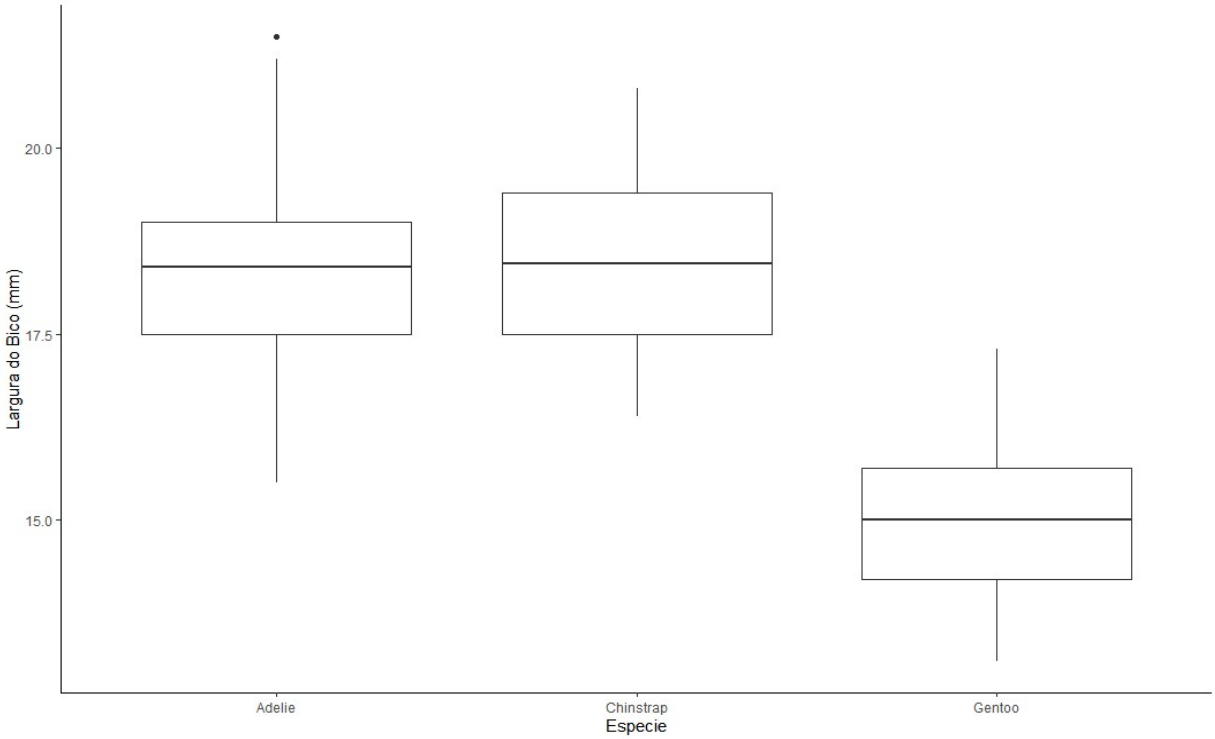
Quanto a interação das variáveis quantitativas com a variável resposta podemos observar através dos gráficos boxplot:

Figura 1 - Boxplot Tamanho do Bico x Espécie



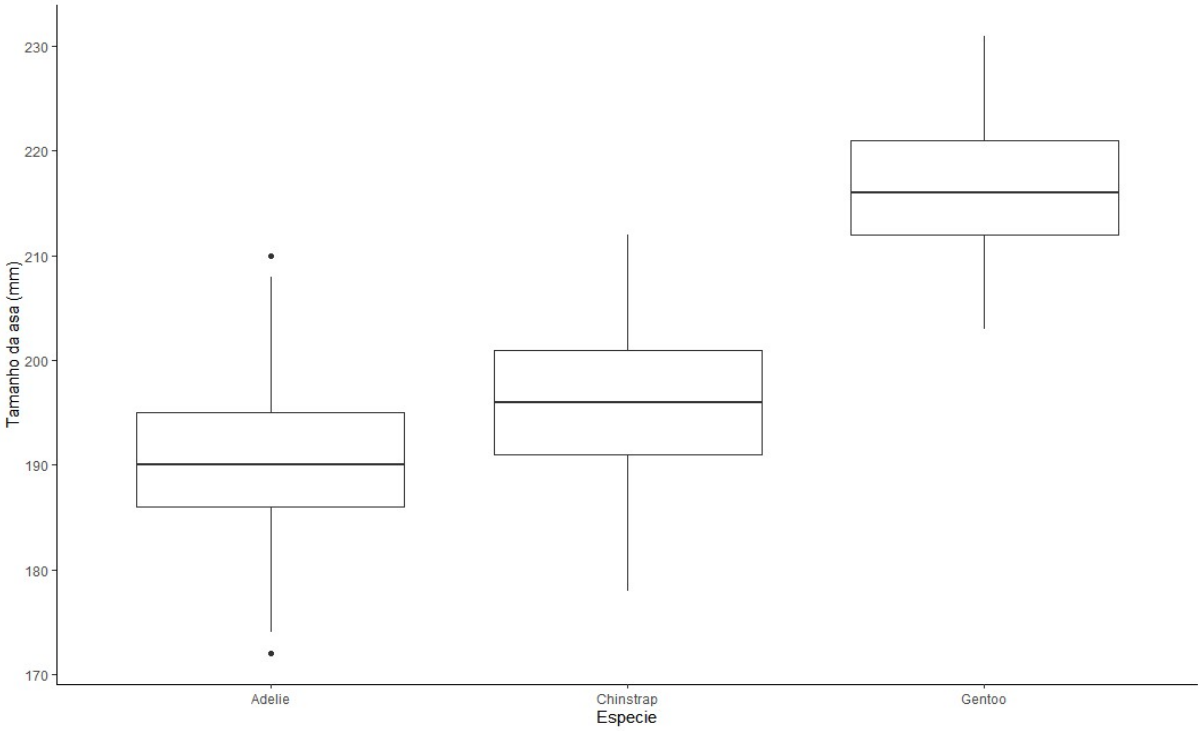
Fonte: Elaborado pelo autor (2021).

Figura 2 - Boxplot Largura do Bico x Espécie



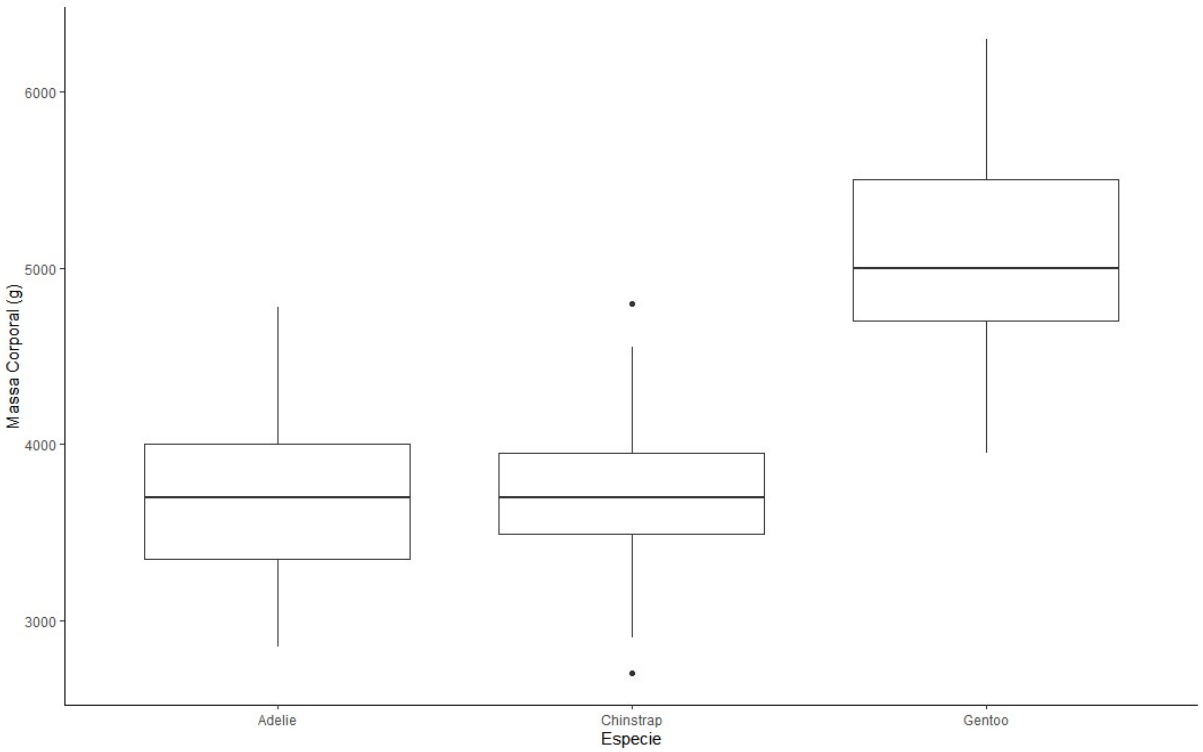
Fonte: Elaborado pelo autor (2021).

Figura 3 - Boxplot Tamanho da Asa x Espécie



Fonte: Elaborado pelo autor (2021).

Figura 4 - Boxplot Peso Corporal x Espécie



Fonte: Elaborado pelo autor (2021).

Os gráficos apresentam comportamentos distintos das variáveis em relação a espécie, exceto massa corporal e tamanho da asa, onde *Adelie* e *Chinstrap* apresentam medidas semelhante e menores que as de *Gentoo*. Essa característica, associada ao grau de correlação forte, pode ser um indicativo de multicolinearidade, como Ferrari (1989) definiu. Porém, como a correlação é menor que 0,90, a multicolinearidade pode ser considerada branda, sem produzir um grande impacto no desempenho do modelo, conforme apontado por Nakamura (2013).

3.3 CONSTRUÇÃO DO MODELO

Para a construção do modelo separamos o banco de dados em duas partes. O banco possui 344 indivíduos pesquisados, foram excluídos 2 animais, pois foi possível coletar apenas a ilha, o ano e a espécie desses animais. Dos 342 elementos restantes, ficaram 227 para a construção do modelo e 115 para validarmos o modelo. Essa divisão foi realizada através de sorteio aleatório e usando o método de *Holdaut* alocando 1/3 dos dados para validação como Devroye e Wagner (1979) indicou como uma das formas desse procedimento.

Para a seleção das variáveis, são usados os dois métodos descritos na seção anterior: o método *stepwise* e os 6 passos do procedimento recomendado por Hosmer e Lemeshow (1989). Foram usadas como medida de desempenho do modelo, o AIC, no procedimento *stepwise* (estatística de saída no código em R, disponível no Apêncice A). Por outro lado, o procedimento não automatizado proposto por Hosmer e Lemeshow (1989) utiliza o p-valor do teste da razão de verossimilhanças como medida de desempenho na escolha do modelo.

3.3.1 SELEÇÃO *STEPWISE*

O método *stepwise* é a junção de dois métodos: o método *Forward selection* e *Backward elimination.* Ou seja, partimos do modelo sem variável explicativa e adicionamos conforme as medidas de desempenho vão melhorando e retiramos até chegar na melhor medida de desempenho. O passo a passo foi:

Na comparação entre o modelo sem variáveis explicativas contra a adição de cada variável o resultado do AIC foi:

Tabela 6 - Resultado do AIC Passo 1

Modelo	AIC
Referencia	480,5917
+ Ilha	258,7493
+ Tamanho do Bico	233,1139
+ Largura do Bico	264,0083
+ Tamanho da Asa	194,1700
+ Massa Corporal	277,7699
+ Sexo	478,5826
+ Ano	483,6762

Fonte: Elaborado pelo autor (2021).

Com esse resultado, a variável Tamanho da asa é incorporada ao modelo de referência e passamos para o próximo conjunto de comparações.

Tabela 7 - Resultado do AIC Passo 2

Modelo	AIC
Referencia	194,1700
+ Ilha	125,1289
+ Tamanho do Bico	69,7356
+ Largura do Bico	164,1128
+ Massa Corporal	185,5134
+ Sexo	174,9927
+ Ano	196,6290

Fonte: Elaborado pelo autor (2021).

Com esse resultado a variável tamanho do bico é incorporada ao modelo para se fazer novas comparações.

Tabela 8 - Resultado do AIC Passo 3

Modelo	AIC
Referencia	69,7356
+ Ilha	40,8002
+ Largura do Bico	31,2750
+ Massa Corporal	43,0003
+ Sexo	41,2485
+ Ano	92,3131

Fonte: Elaborado pelo autor (2021).

A variável Largura do Bico apresentou menor AIC e passa a fazer parte do modelo de referência, e vai em sequência para o próximo passo.

Tabela 9 - Resultado do AIC Passo 4

Modelo	AIC
Referencia	31,2750
+ Ilha	24,0133
+ Massa Corporal	20,0012
+ Sexo	20,0023
+ Ano	35,2829

Fonte: Elaborado pelo autor (2021).

É incorporado a variável massa corporal à referência e seguimos ao próximo passo.

Tabela 10 - Resultado do AIC Passo 5

Modelo	AIC
Referencia	20,0012
+ Ilha	28,0008
+ Sexo	24,0026
+ Ano	24,0023

Fonte: Elaborado pelo autor (2021).

Como o modelo de referência possui o menor valor de AIC, não é alocado mais nenhuma variável ao modelo. Agora é realizado a retirada de uma variável da referência para identificar qual possui melhor ajuste.

Tabela 11 - Resultado do AIC Passo 6

Modelo	AIC
Referencia	20,0012
- Tamanho da Asa	16,4808
- Tamanho do Bico	245,5448
- Largura do Bico	43,0003
- Massa Corporal	31,2750

Fonte: Elaborado pelo autor (2021).

Conforme analise do AIC, a variável Tamanho da Asa é retirada do modelo, e seguimos com a análise.

Tabela 12 - Resultado do AIC Passo 7

Modelo	AIC
Referencia	16,4808
- Tamanho do Bico	283,0642
- Largura do Bico	89,5547
- Massa Corporal	59,8967

Fonte: Elaborado pelo autor (2021).

Quando o modelo de referência atinge o menor valor de AIC, ele passa a ser o modelo definitivo para o método *stepwise*, o modelo é:

$$species \sim bill\_length\_mm + bill\_depth\_mm + body\_mass\_g$$

3.3.2 MÉTODO NÃO AUTOMATIZADO

O passo a passo do modelo é demostrado abaixo com as tabelas do respectivo p-valor. No passo 1, é feita a comparação do modelo sem variáveis explicativas com os modelos com 1 variável explicativa. Todos que possuírem p-valor significativo são agregados ao modelo de referência. Os p-valores foram:

Tabela 13 - Resultado p-valor passo1

Modelo	P-valor
+ Ilha	0,0000
+ Tamanho do Bico	0,0000
+ Largura do Bico	0,0000
+ Tamanho da Asa	0,0000
+ Massa Corporal	0,0000
+ Sexo	0,0142
+ Ano	0,3387

Fonte: Elaborado pelo autor (2021).

No passo 1, apenas a variável Ano não é incluída no modelo de referência. No passo 2, retira-se da referência uma variável e faz a comparação. O modelo em que o p-valor deixar de ser significativo permanece na referência, caso contrário, é retirado, então o resultado foi:

Tabela 14 - Resultado p-valor passo 2

Modelo	P-valor
- Ilha	0,9524
- Tamanho do Bico	0,0000
- Largura do Bico	0,1266
- Tamanho da Asa	0,9830
- Massa Corporal	0,9972
- Sexo	0,9784

Fonte: Elaborado pelo autor (2021).

A variável que sendo retirada não afetou na significância foi o Tamanho do Bico, por isso ela é retirada do modelo de referência. No passo 3, comparamos o modelo de referência com o modelo incluindo as variáveis que foram retiradas no passo 2. No passo 2 foi retirada apenas a variável Tamanho da Bico e o p-valor do teste dessa comparação foi de 0,0000, então voltamos com a variável para o modelo de referência.

O passo 4 retorna a variável que foi retirada no passo 1 para testar sua significância. No modelo foi retirada apenas a variável Ano, o seu p-valor foi de 0,9961, portanto, essa variável não retorna ao modelo.

O passo 5 desse estudo terá o mesmo resultado do passo 2. Pelo fato de o banco de dados possuir poucas variáveis explicativas, o modelo de referência desse passo é similar ao passo 2. Entretanto, ao contrário do passo 2, não será retirada a variável Tamanho do Bico, pois ao decorrer dos passos a mesma já se provou relevante. Nesse passo também é o momento do pesquisador retirar alguma variável caso deseje. As variáveis Tamanho da Asa e Massa Corporal apresentam correlação alta, porém os p-valores dos modelos sem suas presenças não são significativos, então a retirada dessas variáveis não é feita.

O passo 6 inclui possíveis interações entre as variáveis em comparação ao modelo de referência, porém todas as interações não foram significativas, não sendo incorporado ao modelo final, que ficou da forma:

$$species \sim island + bill\_length\_mm + bill\_depth\_mm + flipper\_length\_mm + body\_mass\_g + sex$$

3.4 RESULTADOS

Seja o modelo A o obtido através do método *stepwise* e o modelo B referente ao obtido pelo método não automático, eles apresentaram bastante diferenças. O modelo A apresenta três variáveis explicativas a menos. Os resultados dos modelos foram:

Tabela 15 - Resumo das estatísticas do Modelo A

	Chinstrap			*Gentoo*		
	Coeficientes	Erro Padrão	Odds Ration	Coeficientes	Erro Padrão	Odds Ration
(*Intercept*)	-180,037	6,06e-02	6,47e-79	9,711	4,09e-06	16498,09
Tamanho do Bico	24,64	1,6116	5,02e10	15,97	0,0002	8,62e6
Profundidade do Bico	-34,33	8,66e-01	1,23e-15	-54,68	5,97e-05	1,79e-24
Massa Corporal	-0,0730	0,0219	0,93	0,0492	0,0172	1,05

Fonte: Elaborado pelo autor (2021).

Tabela 16 - Resumo das estatísticas do Modelo B

	Chinstrap			*Gentoo*		
	Coeficientes	Erro Padrão	Odds Ration	Coeficientes	Erro Padrão	Odds Ration
(*Intercept*)	-108,126	0,0633	1,10 e-47	-6,147	0,0281	2,10 e-3
Ilha *Dream*	21,02	5,43e-01	1,35 e9	-28,51	1,11e-05	4,15 e-13
Ilha *Torgersen*	-7,668	7,58e-12	4,68 e-4	-22,522	3,39e-01	1,66 e-10
Tamanho do Bico	15,457	0,5060	5,16 e6	9,881	3,8680	1,96 e4
Profundidade do Bico	-21,83	3,0310	3,31 e-10	-19,87	1,3740	2,35 e-9
Tamanho da Asa	-0,3814	3,0600	0,68	-0,7142	12,2400	0,49
Massa Corporal	-0,0307	0,1670	0,97	0,0195	0,5567	1,02
Sexo Masculino	-13,852	0,0007	9,64 e-7	-7,702	0,4530	4,52 e-4

Fonte: Elaborado pelo autor (2021).

Parte do banco de dados ficou reservado para testar a acurácia dos modelos. As matrizes de confusão dos modelos foram:

Tabela 17 - Matrix de confusão Modelo A

		Predito		
		Adelie	*Chinstrap*	*Gentoo*
Real	*Adelie*	51	0	0
	Chinstrap	0	23	0
	Gentoo	0	0	41

Fonte: Elaborado pelo autor (2021).

Tabela 18 - Matriz de confusão do Modelo B

		Predito		
		Adelie	*Chinstrap*	*Gentoo*
Real	*Adelie*	48	0	0
	Chinstrap	0	23	0
	Gentoo	0	0	38

Fonte: Elaborado pelo autor (2021).

O Modelo A obteve acurácia de 100% enquanto o Modelo B de 95% pois 6 indivíduos não tiveram definido qual espécie, ou seja, em ambos os modelos não ocorreu indicação falsa da espécie. Resultado final muito positivo de ambos os modelos.

4 CONCLUSÃO

O volume de informações devido ao avanço da tecnologia dos dias atuais é gigantesco, portanto, saber extrair informações desse quantitativo é essencial. Assim, modelos preditivos tornam-se cada vez mais importantes. Neste trabalho a regressão logística multinomial, foi apresentada e se mostrou como uma alternativa para variáveis respostas categóricas como evidenciou os resultados da nossa aplicação.

Em relação a comparação nos métodos de seleção de variável, o banco penguins Horst (2020), embora com um número pequeno de variáveis explicativas, a partir dele foi possível identificar a principal diferença entre modos automatizados e não automatizados. O método *stepwise*, que é automatizado, penaliza de certa forma modelos com muitas variáveis respostas, enquanto o método não automatizado privilegia o conhecimento do pesquisador, podendo agregar variáveis explicativas baseando-se no conhecimento do problema. Na aplicação, isto se refletiu no Modelo A, obtido através do *stepwise*, contendo três variáveis explicativas, enquanto o Modelo B, obtido com o método não automatizado, possuía seis.

Na aplicação estudada, embora distintos, os modelos de regressão logística multinomial selecionados apresentaram boa acurácia, mostrando serem boas ferramentas para predição de dados. Percebe-se assim que, distintos em sua concepção, ambos os métodos apresentam resultados satisfatórios, ficando a cargo do pesquisador escolher qual o método se encaixa mais adequadamente à área de conhecimento de seu estudo.

Existem diversos dados categóricos, que são respostas e estão abertos a estudos de predição, como o resultado de uma partida de futebol, o bem estar social, o bem estar financeiro, local para abrir uma franquia entre outros. A regressão logística multinomial pode ser utilizada para a predição por apresentar uma excelente acurácia.

REFERÊNCIAS BIBLIOGRÁFICAS

AKAIKE, H. A new look at the statistical model identification. **IEEE transactions on automatic control**, 19, n. 6, p. 716-723, 1974.

ALLISON, P. D., 2014, **Measures of fit for logistic regression**. 1-13.

ALVARENGA, A. M. T. **Modelos lineares generalizados: aplicação a dados de acidentes rodoviários**. 2015. -.

BARBOZA, F.; KIMURA, H.; ALTMAN, E. Machine learning models and bankruptcy prediction. **Expert Systems with Applications**, 83, p. 405-417, 2017.

COLLETT, D. **Modelling binary data**. CRC press, 2002. 1584883243.

COX, D.; SNELL, E. **Analysis of Binary Data**. CRC Press, 1989. 0412306204.

DANTAS, R. F.; DESOUZA, S. A. Modelo de risco e decisão de crédito baseado em estrutura de capital com informação assimétrica. **Pesquisa Operacional**, 28, p. 263-284, 2008.

NUNES, L. L. Aplicaçao do modelo de Regressao Logıstica para apoioa decisao de crédito.

DE SOUSA, G. C.; ALVES, J. M. S. A regressão linear de Galton: atividades históricas para função afim e estatística básica usando planilhas eletrônicas. **Conexões-Ciência e Tecnologia**, 9, n. 4, p. 26-36, 2016.

DEVROYE, L.; WAGNER, T. Distribution-free performance bounds for potential function rules. **IEEE Transactions on Information Theory**, 25, n. 5, p. 601-604, 1979.

DINIZ, C.; LOUZADA, F. Modelagem Estatística para risco de crédito. **ABE, São Paulo-SP**, 2012.

FERRARI, F. **Estimadores viesados para modelos de regressão em presença de multicolinearidade**. 1989. -, Universidade de São Paulo.

GALTON, F. **Natural inheritance**. Macmillan and Company, 1889.

HORST AM, HILL AP, GORMAN KB. **palmerpenguins: Palmer**, 2020. Dados de pinguins do arquipélago (Antártica). Pacote R versão 0.1.0. Disponível em <https://allisonhorst.github.io/palmerpenguins/>. Acesso em 14 de ago. de 2021

HOSMER DW, LEMESHOW S, Applied logistic regression. Wiley, New York: Wiley & Sons, Inc 1989.

LATTIN, J.; CARROLL, J. D.; GREEN, P. E. Análise de dados multivariados. **São Paulo: Cengage Learning**, 475, 2011.

MINUSSI, J. A.; DAMACENA, C.; NESS JR, W. L. Um modelo de previsão de solvência utilizando regressão logística. **Revista de Administração Contemporânea**, 6, p. 109-128, 2002.

MUKAKA, M. M. A guide to appropriate use of correlation coefficient in medical research. **Malawi medical journal**, 24, n. 3, p. 69-71, 2012.

NAKAMURA, K. G. **Multicolinearidade em modelos de regressão logística**. 2013. -, Universidade de São Paulo.

RIBEIRO, A. M.; GUIMARÃES, M. J.; LIMA, M. D. C.; SARINHO, S. W. *et al.* Fatores de risco para mortalidade neonatal em crianças com baixo peso ao nascer. **Revista de Saúde Pública**, 43, p. 246-255, 2009.

SOUZA, É. C. D. **Análise de influência local no modelo de regressão logística**. 2006. -, Universidade de São Paulo.

SZUMILAS, M. Explaining odds ratios. **Journal of the Canadian academy of child and adolescent psychiatry**, 19, n. 3, p. 227, 2010.

ZELTERMAN, D. **Applied multivariate statistics with R**. Springer, 2015. 3319140930.

APÊNDICE A – Códigos das Aplicações em R

```
library(tidyverse)
library(palmerpenguins)
library(tidyverse)
library(palmerpenguins)
library(dbplyr)
library(dplyr)
library(ggplot2)
require(nnet)
library(caret)

plot(penguins)
penguins<-penguins

## analise das variaveis categoricas
table(penguins$species,penguins$island)
table(penguins$species,penguins$sex)
table(penguins$species,penguins$year)

## teste de distribuição das variaveis categoricas
chisq.test(penguins$species,penguins$year)
chisq.test(penguins$species,penguins$sex)
chisq.test(penguins$species,penguins$island)



## Resumo das variaveis continuas
summary(penguins$bill_length_mm)
summary(penguins$bill_depth_mm)
summary(penguins$flipper_length_mm)
summary(penguins$body_mass_g)


## Matriz de correlação
penguins %>%
```

```
  select_if(is.numeric) %>%
  drop_na() %>%
  cor()

## Boxplot das variaveis continuas

ggplot(data = penguins,
       aes(x = species,
           y = bill_length_mm)) +
  geom_boxplot() +
   labs(x = "Especie",
       y = "Tamanho do bico (mm)",
       color = "Penguin species",
       shape = "Penguin species") +
  theme_classic()


ggplot(data = penguins,
       aes(x = species,
           y = bill_depth_mm)) +
  geom_boxplot() +
  labs(x = "Especie",
       y = "Largura do Bico (mm)",
       color = "Penguin species",
       shape = "Penguin species") +
  theme_classic()


ggplot(data = penguins,
       aes(x = species,
           y = flipper_length_mm)) +
  geom_boxplot() +
  labs(x = "Especie",
       y = "Tamanho da asa (mm)",
       color = "Penguin species",
       shape = "Penguin species") +
```

```
  theme_classic()


ggplot(data = penguins,
       aes(x = species,
           y = body_mass_g)) +
  geom_boxplot() +
  labs(x = "Especie",
       y = "Massa Corporal (g)",
       color = "Penguin species",
       shape = "Penguin species") +
  theme_classic()


#########################################################################
###################################

## Regressão Logistica multinomial

p<-penguins[c(-4),]
c<-p[c(-271),]

training.samples <- createDataPartition(penguins$species, p = 0.66, list = FALSE)
train.data  <- penguins[training.samples, ]
test.data <- penguins[-training.samples, ]

train.data$species <- as.factor(train.data$species) # create factor categories
train.data$species <- relevel(train.data$species, ref = 1) # set reference category


################### Stepwase ##################

inicio = multinom(species ~ 1, data = train.data, maxit = 200)

1 island
2 bill_length_mm
```

```
3 bill_depth_mm
4 flipper_length_mm
5 body_mass_g
6 sex
7 year

P1.1 = multinom(species ~ island, data = train.data, maxit = 200)
P1.2 = multinom(species ~ bill_length_mm, data = train.data, maxit = 200)
P1.3 = multinom(species ~ bill_depth_mm, data = train.data, maxit = 200)
P1.4 = multinom(species ~ flipper_length_mm, data = train.data, maxit = 200)
P1.5 = multinom(species ~ body_mass_g, data = train.data, maxit = 200)
P1.6 = multinom(species ~ sex, data = train.data, maxit = 200)
P1.7 = multinom(species ~ year, data = train.data, maxit = 200)

inicio$AIC
P1.1$AIC
P1.2$AIC
P1.3$AIC
P1.4$AIC
P1.5$AIC
P1.6$AIC
P1.7$AIC
# adicionado a variavel 4 flipper_length_mm; species ~ flipper_length_mm
P2<- P1.4
P2.1 = multinom(species ~ flipper_length_mm + island, data = train.data, maxit = 200)
P2.2 = multinom(species ~ flipper_length_mm + bill_length_mm, data = train.data, maxit =
200)
P2.3 = multinom(species ~ flipper_length_mm + bill_depth_mm, data = train.data, maxit = 200)
P2.5 = multinom(species ~ flipper_length_mm + body_mass_g, data = train.data, maxit = 200)
P2.6 = multinom(species ~ flipper_length_mm + sex, data = train.data, maxit = 200)
P2.7 = multinom(species ~ flipper_length_mm + year, data = train.data, maxit = 200)
T2.4 = multinom(species ~ 1, data = train.data, maxit = 200)

P2$AIC
P2.1$AIC
P2.2$AIC
```

```
P2.3$AIC
P2.5$AIC
P2.6$AIC
P2.7$AIC
T2.4$AIC

# adicionado a variavel 2 bill_length_mm; species ~ flipper_length_mm + bill_length_mm
P3 <- P2.2
P3.1 = multinom(species ~ flipper_length_mm + bill_length_mm + island, data = train.data, maxit = 200)
P3.3 = multinom(species ~ flipper_length_mm + bill_length_mm + bill_depth_mm, data = train.data, maxit = 200)
P3.5 = multinom(species ~ flipper_length_mm + bill_length_mm + body_mass_g, data = train.data, maxit = 200)
P3.6 = multinom(species ~ flipper_length_mm + bill_length_mm + sex, data = train.data, maxit = 200)
P3.7 = multinom(species ~ flipper_length_mm + bill_length_mm + year, data = train.data, maxit = 200)
T3.4 = multinom(species ~ bill_length_mm, data = train.data, maxit = 200)
T3.2 = multinom(species ~ flipper_length_mm, data = train.data, maxit = 200)



P3$AIC
P3.1$AIC
P3.3$AIC
P3.5$AIC
P3.6$AIC
P3.7$AIC
T3.4$AIC
T3.2$AIC

# Adicionando a variavel  5 body_mass_g; species ~ flipper_length_mm + bill_length_mm + body_mass_g
P4 <- P3.5
```

```
P4.1 = multinom(species ~ flipper_length_mm + bill_length_mm + body_mass_g + island, data = train.data, maxit = 200)
P4.3 = multinom(species ~ flipper_length_mm + bill_length_mm + bill_depth_mm + body_mass_g, data = train.data, maxit = 200)
P4.6 = multinom(species ~ flipper_length_mm + bill_length_mm + body_mass_g + sex, data = train.data, maxit = 200)
P4.7 = multinom(species ~ flipper_length_mm + bill_length_mm + body_mass_g + year, data = train.data, maxit = 200)
T4.4 = multinom(species ~ bill_length_mm + body_mass_g, data = train.data, maxit = 200)
T4.2 = multinom(species ~ flipper_length_mm + body_mass_g, data = train.data, maxit = 200)
T4.5 = multinom(species ~ flipper_length_mm + bill_length_mm, data = train.data, maxit = 200)


P4$AIC
P4.1$AIC
P4.3$AIC
P4.6$AIC
P4.7$AIC
T4.4$AIC
T4.2$AIC
T4.5$AIC

#Adicionando a variavel 3 bill_depth_mm, species ~ flipper_length_mm + bill_length_mm + bill_depth_mm + body_mass_g

P5 <- P4.3
P5.1 = multinom(species ~ flipper_length_mm + bill_length_mm + bill_depth_mm + body_mass_g + island, data = train.data, maxit = 200)
P5.6 = multinom(species ~ flipper_length_mm + bill_length_mm + bill_depth_mm + body_mass_g + sex, data = train.data, maxit = 200)
P5.7 = multinom(species ~ flipper_length_mm + bill_length_mm + bill_depth_mm + body_mass_g + year, data = train.data, maxit = 200)
T5.4 = multinom(species ~ bill_length_mm + bill_depth_mm + body_mass_g, data = train.data, maxit = 200)
```

```
T5.2 = multinom(species ~ flipper_length_mm + bill_depth_mm + body_mass_g, data = train.data, maxit = 200)
T5.3 = multinom(species ~ flipper_length_mm + bill_length_mm + body_mass_g, data = train.data, maxit = 200)
T5.5 = multinom(species ~ flipper_length_mm + bill_length_mm + bill_depth_mm, data = train.data, maxit = 200)


P5$AIC
P5.1$AIC
P5.6$AIC
P5.7$AIC
T5.4$AIC
T5.2$AIC
T5.3$AIC
T5.5$AIC

# Retirar a variavel 4 flipper_length_mm;  species ~ bill_length_mm + bill_depth_mm + body_mass_g

T6 <- T5.4
P6.1 = multinom(species ~ bill_length_mm + bill_depth_mm + body_mass_g + island, data = train.data, maxit = 200)
P6.6 = multinom(species ~ bill_length_mm + bill_depth_mm + body_mass_g + sex, data = train.data, maxit = 200)
P6.7 = multinom(species ~ bill_length_mm + bill_depth_mm + body_mass_g + year, data = train.data, maxit = 200)
S6.4 = multinom(species ~ flipper_length_mm + bill_length_mm + bill_depth_mm + body_mass_g, data = train.data, maxit = 200)
T6.2 = multinom(species ~ bill_depth_mm + body_mass_g, data = train.data, maxit = 200)
T6.3 = multinom(species ~ bill_length_mm + body_mass_g, data = train.data, maxit = 200)
T6.5 = multinom(species ~ bill_length_mm + bill_depth_mm, data = train.data, maxit = 200)

T6$AIC
P6.1$AIC
P6.6$AIC
```

```
P6.7$AIC
S6.4$AIC
T6.2$AIC
T6.3$AIC
T6.5$AIC

# Com o menor AIC a equação escolhida é species ~ bill_length_mm + bill_depth_mm + body_mass_g
##############################
# Escolha 1  1
1 island
2 bill_length_mm
3 bill_depth_mm
4 flipper_length_mm
5 body_mass_g
6 sex
7 year

Var <- c("island"  , "bill_length_mm" , "bill_depth_mm"  , "flipper_length_mm"  , "body_mass_g" , "sex" , "year")
# Passo 1
inicio = multinom(species ~ 1, data = train.data, maxit = 200)
P1.1 = multinom(species ~ island, data = train.data, maxit = 200)
P1.2 = multinom(species ~ bill_length_mm, data = train.data, maxit = 200)
P1.3 = multinom(species ~ bill_depth_mm, data = train.data, maxit = 200)
P1.4 = multinom(species ~ flipper_length_mm, data = train.data, maxit = 200)
P1.5 = multinom(species ~ body_mass_g, data = train.data, maxit = 200)
P1.6 = multinom(species ~ sex, data = train.data, maxit = 200)
P1.7 = multinom(species ~ year, data = train.data, maxit = 200)

Pval1.1<- 1 - pchisq(inicio$deviance-P1.1$deviance,1)
Pval1.2<- 1 - pchisq(inicio$deviance-P1.2$deviance,1)
Pval1.3<- 1 - pchisq(inicio$deviance-P1.3$deviance,1)
Pval1.4<- 1 - pchisq(inicio$deviance-P1.4$deviance,1)
Pval1.5<- 1 - pchisq(inicio$deviance-P1.5$deviance,1)
Pval1.6<- 1 - pchisq(inicio$deviance-P1.6$deviance,1)
```

```
Pval1.7<- 1 - pchisq(inicio$deviance-P1.7$deviance,1)

Pass1 <- c(Pval1.1,Pval1.2,Pval1.3,Pval1.4,Pval1.5,Pval1.6,Pval1.7)
ResP1 <- data.frame(Var,Pass1)
View(ResP1)


# não incluimos a variavel 7 Year
# Passo 2

P2 <- multinom(species ~ island + bill_length_mm + bill_depth_mm + flipper_length_mm + body_mass_g + sex , data = train.data, maxit = 200)
P2.1 <-multinom(species ~ bill_length_mm + bill_depth_mm + flipper_length_mm + body_mass_g + sex , data = train.data, maxit = 200)
P2.2 <-multinom(species ~ island + bill_depth_mm + flipper_length_mm + body_mass_g + sex , data = train.data, maxit = 200)
P2.3 <-multinom(species ~ island + bill_length_mm + flipper_length_mm + body_mass_g + sex , data = train.data, maxit = 200)
P2.4 <-multinom(species ~ island + bill_length_mm + bill_depth_mm + body_mass_g + sex , data = train.data, maxit = 200)
P2.5 <-multinom(species ~ island + bill_length_mm + bill_depth_mm + flipper_length_mm + sex , data = train.data, maxit = 200)
P2.6 <-multinom(species ~ island + bill_length_mm + bill_depth_mm + flipper_length_mm + body_mass_g , data = train.data, maxit = 200)

Pval2.1<- 1 - pchisq(abs(P2.1$deviance-P2$deviance),1)
Pval2.2<- 1 - pchisq(abs(P2.2$deviance-P2$deviance),1)
Pval2.3<- 1 - pchisq(abs(P2.3$deviance-P2$deviance),1)
Pval2.4<- 1 - pchisq(abs(P2.4$deviance-P2$deviance),1)
Pval2.5<- 1 - pchisq(abs(P2.5$deviance-P2$deviance),1)
Pval2.6<- 1 - pchisq(abs(P2.6$deviance-P2$deviance),1)

Var2 <- c("island" , "bill_length_mm" , "bill_depth_mm" , "flipper_length_mm" , "body_mass_g" , "sex")
Pass2 <- c(Pval2.1,Pval2.2,Pval2.3,Pval2.4,Pval2.5,Pval2.6)
ResP2 <- data.frame(Var2,Pass2)
```

```
View(ResP2)

# Retiramos a variavel 2 bill_length_mm

#Passo 3
P3<- multinom(species ~ island  + bill_depth_mm + flipper_length_mm + body_mass_g + sex , data = train.data, maxit = 200)
P3.2<- multinom(species ~ island + bill_length_mm + bill_depth_mm + flipper_length_mm + body_mass_g + sex , data = train.data, maxit = 200)


Pval3.2<- 1 - pchisq(abs(P3.2$deviance-P3$deviance),1)

Pval3.2

# Voltamos com a variavel excluida pois a mesma mostrou significancia

#Passo 4
P4<- multinom(species ~ island + bill_length_mm + bill_depth_mm + flipper_length_mm + body_mass_g + sex , data = train.data, maxit = 200)
P4.7 <- multinom(species ~ island + bill_length_mm + bill_depth_mm + flipper_length_mm + body_mass_g + sex + year, data = train.data, maxit = 200)

Pval4.7<- 1 - pchisq(abs(P4.7$deviance-P4$deviance),1)
Pval4.7

# Não retorna a variavel 7 year

#Passo 5

# como nosso modelo possui poucas variaveis esse passo será semelhante ao passo 2, onde apenas a variavel 2 bill_length_mm foi indicada pra ser retirada, porem seguindo o preceito do passo 3 ela não é retirada do modelo
# Porem devido a correlação alta entre tamanho da asa e massa corporal será importante retirar 1 variavel do modelo e ver se ainda possui significancia
```

```
P5<- multinom(species ~ island + bill_length_mm + bill_depth_mm + flipper_length_mm + body_mass_g + sex , data = train.data, maxit = 200)
P5.4 <- multinom(species ~ island + bill_length_mm + bill_depth_mm + body_mass_g + sex, data = train.data, maxit = 200)
P5.5 <- multinom(species ~ island + bill_length_mm + bill_depth_mm + flipper_length_mm + sex, data = train.data, maxit = 200)

Pval5.4<- 1 - pchisq(abs(P5.4$deviance-P5$deviance),1)
Pval5.5<- 1 - pchisq(abs(P5.5$deviance-P5$deviance),1)
Pval5.4
Pval5.5

#Passo 6

P6  <- multinom(species ~ island + bill_length_mm + bill_depth_mm + body_mass_g + flipper_length_mm + sex , data = train.data, maxit = 200)
P6.1<- multinom(species ~ island + bill_length_mm + bill_depth_mm + body_mass_g + flipper_length_mm + sex + island*bill_length_mm , data = train.data, maxit = 200)
P6.2<- multinom(species ~ island + bill_length_mm + bill_depth_mm + body_mass_g + flipper_length_mm + sex + island*bill_depth_mm, data = train.data, maxit = 200)
P6.3<- multinom(species ~ island + bill_length_mm + bill_depth_mm + body_mass_g + flipper_length_mm + sex + island*body_mass_g, data = train.data, maxit = 200)
P6.4<- multinom(species ~ island + bill_length_mm + bill_depth_mm + body_mass_g + flipper_length_mm + sex + bill_length_mm*bill_depth_mm, data = train.data, maxit = 200)
P6.5<- multinom(species ~ island + bill_length_mm + bill_depth_mm + body_mass_g + flipper_length_mm + sex + bill_length_mm*body_mass_g, data = train.data, maxit = 200)
P6.6<- multinom(species ~ island + bill_length_mm + bill_depth_mm + body_mass_g + flipper_length_mm + sex + bill_depth_mm*body_mass_g, data = train.data, maxit = 200)
P6.7  <- multinom(species ~ island + bill_length_mm + bill_depth_mm + body_mass_g + flipper_length_mm + sex + flipper_length_mm*island, data = train.data, maxit = 200)
P6.8  <- multinom(species ~ island + bill_length_mm + bill_depth_mm + body_mass_g + flipper_length_mm + sex + flipper_length_mm*bill_length_mm, data = train.data, maxit = 200)
P6.9  <- multinom(species ~ island + bill_length_mm + bill_depth_mm + body_mass_g + flipper_length_mm + sex + flipper_length_mm*bill_depth_mm, data = train.data, maxit = 200)
P6.10  <- multinom(species ~ island + bill_length_mm + bill_depth_mm + body_mass_g + flipper_length_mm + sex + flipper_length_mm*body_mass_g, data = train.data, maxit = 200)
```

```
P6.11 <- multinom(species ~ island + bill_length_mm + bill_depth_mm + body_mass_g + flipper_length_mm + sex + flipper_length_mm*sex, data = train.data, maxit = 200)
P6.12 <- multinom(species ~ island + bill_length_mm + bill_depth_mm + body_mass_g + flipper_length_mm + sex + sex*island, data = train.data, maxit = 200)
P6.13 <- multinom(species ~ island + bill_length_mm + bill_depth_mm + body_mass_g + flipper_length_mm + sex + sex*bill_length_mm, data = train.data, maxit = 200)
P6.14 <- multinom(species ~ island + bill_length_mm + bill_depth_mm + body_mass_g + flipper_length_mm + sex + sex*bill_depth_mm, data = train.data, maxit = 200)
P6.15 <- multinom(species ~ island + bill_length_mm + bill_depth_mm + body_mass_g + flipper_length_mm + sex + sex*body_mass_g, data = train.data, maxit = 200)



Pval6.1<- 1 - pchisq(abs(P6.1$deviance-P6$deviance),1)
Pval6.2<- 1 - pchisq(abs(P6.2$deviance-P6$deviance),1)
Pval6.3<- 1 - pchisq(abs(P6.3$deviance-P6$deviance),1)
Pval6.4<- 1 - pchisq(abs(P6.4$deviance-P6$deviance),1)
Pval6.5<- 1 - pchisq(abs(P6.5$deviance-P6$deviance),1)
Pval6.6<- 1 - pchisq(abs(P6.6$deviance-P6$deviance),1)
Pval6.7<- 1 - pchisq(abs(P6.7$deviance-P6$deviance),1)
Pval6.8<- 1 - pchisq(abs(P6.8$deviance-P6$deviance),1)
Pval6.9<- 1 - pchisq(abs(P6.9$deviance-P6$deviance),1)
Pval6.10<- 1 - pchisq(abs(P6.10$deviance-P6$deviance),1)
Pval6.11<- 1 - pchisq(abs(P6.11$deviance-P6$deviance),1)
Pval6.12<- 1 - pchisq(abs(P6.12$deviance-P6$deviance),1)
Pval6.13<- 1 - pchisq(abs(P6.13$deviance-P6$deviance),1)
Pval6.14<- 1 - pchisq(abs(P6.14$deviance-P6$deviance),1)
Pval6.15<- 1 - pchisq(abs(P6.15$deviance-P6$deviance),1)

Pass6 <- c(Pval6.1,Pval6.2,Pval6.3,Pval6.4,Pval6.5,Pval6.6,Pval6.7,Pval6.8,Pval6.9,Pval6.10,Pval6.11,Pval6.12,Pval6.13,Pval6.14,Pval6.15)
Pass6

# Nenhuma interação deu significativa, sendo assim nosso modelo será: species ~ island + bill_length_mm + bill_depth_mm + flipper_length_mm + body_mass_g + sex
```

```
ModeloA<- multinom(species ~ bill_length_mm + bill_depth_mm + body_mass_g , data = penguins, maxit = 200)
ModeloB<- multinom(species ~ island + bill_length_mm + bill_depth_mm + flipper_length_mm + body_mass_g + sex , data = penguins, maxit = 200)

print(psA <- summary(ModeloA), digits = 4)
print(psB <- summary(ModeloB), digits = 4)

pA <- predict(ModeloA, test.data)
pB <- predict(ModeloB, test.data)

exp(psA$coefficients)
exp(psB$coefficients)

table(pA,test.data$species)
table(pB,test.data$species)
```