UNIVERSIDADE FEDERAL DE JUIZ DE FORA
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE ESTATÍSTICA

Bárbara da Costa Campos Dias

Seleção de variáveis via Backward em Modelo Linear Normal Assimétrico.

Juiz de Fora
2014

Bárbara da Costa Campos Dias

Seleção de variáveis via Backward em Modelo Linear Normal Assimétrico.

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito parcial para obtenção do título de Bacharel em Estatística.

Orientador: Clécio da Silva Ferreira

.

Juiz de Fora

2014

Ficha catalográfica elaborada através do Programa de geração
automática da Biblioteca Universitária da UFJF,
com os dados fornecidos pelo(a) autor(a)

Dias, Bárbara da Costa Campos.
Seleção de variáveis via Backward em Modelo Linear Normal Assimétrico / Bárbara da Costa Campos Dias. -- 2014.
33 p. : il.

Orientador: Clécio da Silva Ferreira
Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Juiz de Fora, Instituto de Ciências Exatas, 2014.

1. Distribuição Normal Assimétrica. 2. Seleção de variáveis. 3. Critério de informação Akaike. I. Ferreira, Clécio da Silva, orient. II. Título.

Bárbara da Costa Campos Dias

Seleção de variáveis via Backward em Modelo Linear Normal Assimétrico.

Monografia apresentada ao Curso de Estatística da Universidade Federal de Juiz de Fora, como requisito parcial para obtenção do título de Bacharel em Estatística.

Aprovado em 31 de janeiro de 2014

BANCA EXAMINADORA

_____

Clécio da Silva Ferreira

Doutor em Estatística- Universidade de São Paulo

_____

Camila Borelli Zeller

Doutora em Estatística-Universidade Estadual de Campinas

_____

Ronaldo Rocha Bastos

Doutor em Regional Planning – Liverpool University

AGRADECIMENTOS

À Deus, pela proteção e por iluminar sempre o meu caminho.

Aos meus pais e minha irmã Alice pelo incentivo, amizade e pelo amor incondicional, me dando forças para continuar. Sem eles eu nada seria.

À minha avó Maria da Penha, por ser minha rainha e sempre me apoiar através de seu amor e suas orações.

Aos meus padrinhos, meus tios e meus primos pelo excesso de amor e carinho. À minha prima Thayene, pela irmandade e amizade de sempre.

Ao Felipe, pelo amor, companheirismo e por sempre ter paciência com os meus estudos.

Aos amigos da UFJF, Bethânia, Juliana Ladeira, Juliana Fisher, Douglas, Gabriely, Isabela, Daniel, Marcel e Vinicius. Obrigada por me ajudar e por me proporcionar muita alegria durante esses anos.

À todos os professores do departamento pela contribuição do meu conhecimento e pela dedicação. Em especial aos professores Lupércio e Camila, pelas excelentes aulas e por despertar em mim o desejo pela área acadêmica.

Ao meu orientador Clécio, pelos inúmeros ensinamentos, paciência e pelos valiosos conselhos que me ajudaram chegar até aqui.

À todas as pessoas que de alguma forma contribuíram para essa conquista.

Serei sempre grata.

RESUMO

Este trabalho tem por objetivo propor uma nova maneira de selecionar variáveis usando o método backward, via Critério de Informação Akaike (AIC), em um modelo de regressão linear com erros seguindo uma distribuição Normal Assimétrica. Investigaremos o efeito de multicolinearidade das variáveis explicativas e da falta de correlação da variável dependente com as variáveis explicativas no processo de ajuste do modelo final (parcimônia). O objetivo dessa investigação será verificar a qualidade do método de seleção proposto nesta monografia.

Para tanto, serão investigadas quatro estratégias de seleção do modelo final. A primeira estratégia será investigar a qualidade do ajuste do modelo final sem utilizar nenhuma analise exploratória anterior a seleção de variáveis usando o método Backward, via AIC. Na segunda estratégia serão retirados os problemas de falta de correlação da variável dependente com as variáveis explicativas antes de utilizar o método de seleção. Já na terceira estratégia serão retirados os problemas de multicolinearidade das variáveis explicativas antes de utilizar o método estudado. E por último, na quarta estratégia serão retirados os dois problemas citados anteriormente antes da seleção.

Este estudo de simulação será realizado para demonstrar a melhor estratégia de seleção do modelo. Por fim, a proposta desta monografia será aplicada a um conjunto de dados reais.

Palavras-Chave: Distribuição Normal Assimétrica, Critério de informação Akaike, Seleção Backward, Algoritmo EM.

ABSTRACT

This work aims to select variables using the backward method, via Akaike Information Criterion (AIC), in a linear regression model with errors following a Skew normal distribution. We investigated the effect of multicollinearity of the explanatory variables and the lack of correlation between the explanatory variables dependent on the fit of the final model (parsimony) process variable. The purpose of this research is to verify the quality of the selection method proposed in this monograph.

For this, we investigated four strategies for selecting the final model. The first strategy is to investigate the goodness of fit of the final model without using any prior exploratory analysis variable selection using the backward method, via AIC. In the second strategy will be phased out problems of lack of correlation between the dependent variable and the explanatory variables before using the selection method. In the third strategy will be removed problems of multicollinearity of the explanatory variables before using the method studied. Finally, the fourth strategy will be removed the two problems mentioned above before selection.

A simulation study is conducted to demonstrate the best strategy for model selection. Finally, the proposal of this monograph is applied to a real data set.

Key-words: Skew-Normal Distribution, Akaike Information Criterion, Selection Backward, EM Algoritmo.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1- Introdução

A distribuição Normal Assimétrica consegue modelar a assimetria dos dados, tendo como caso particular a distribuição Normal. Ela é usada em casos quando a usual suposição de normalidade não é satisfeita, devido à falta de simetria dos dados. E isso é bastante comum na prática.

A distribuição Normal Assimétrica começou a ser discutida através de Azzalini (1985), onde foram introduzidas suas principais propriedades. Posteriormente, Azzalini (2005) apresentou uma discussão sobre distribuições normais assimétricas com aplicações em modelos de regressão. Generalizações para o caso multivariado dessas ideias têm sido propostas por vários autores, por exemplo, Azzalini e Dalla-Valle (1996) e Azzalini e Capitanio (1999).

Existem várias formas de seleção de variáveis, sendo todas de grande importância para a modelagem. Seus resultados podem impactar na estimativa de predição do modelo.

A detecção de multicolinearidade em variáveis regressoras também apresenta grande importância para uma boa seleção do modelo, pois a existência da mesma pode afetar a variância, tornando a estimativa dos parâmetros imprecisa, o que não é conveniente estatisticamente (Kutner et al., 2004; Tamhane & Dunlop, 2000).

Este trabalho tem por objetivo verificar se o método de seleção Backward via Critério de Informação Akaike (AIC) é capaz de detectar algum tipo de problema que possa prejudicar a qualidade do ajuste do modelo selecionado ou se é necessária uma análise anterior à seleção, como a retirada de variáveis multicolineares e/ou com problema de baixa correlação com a variável dependente.

Com o intuito de estudar melhor esse assunto é apresentado no capítulo 2 as formas padrão e locação-escala da distribuição Normal Assimétrica, juntamente com algumas propriedades.

O capítulo 3 descreve o procedimento do método de seleção de variáveis Backward via Critério de Informação Akaike. Já o capítulo 4 apresenta uma forma de detecção da presença de multicolinearidade em variáveis regressoras.

No capitulo 5 foi feito um estudo de simulação verificando se o método estudado é capaz de detectar problemas de baixa correlação da variável dependente com as variáveis

regressoras e problemas de multicolinearidade, ou se é necessária uma análise anterior a seleção.

No capitulo 6, apresentamos uma aplicação do método de seleção estudado através de dados reais. Encontrando assim o modelo mais parcimonioso, usando resultados obtidos no estudo de simulação. Por fim será feita uma conclusão e uma descrição de perspectivas futuras.

2- Modelo Normal Assimétrico

2.1- Distribuição Normal Assimétrica Padrão

Uma variável aleatória $Z \sim SN(\lambda)$ é dita ser normal assimétrica padrão, com parâmetro de assimetria λ, se apresentar a seguinte função de densidade de probabilidade:

$$f_Z(z) = 2\phi(z)\Phi(\lambda z), \qquad z \in \mathbb{R} \tag{1}$$

onde $\phi(.)$ e $\Phi(.)$ são as funções de densidade de probabilidade e de distribuição de uma normal padrão, respectivamente.

A função de distribuição associada à densidade (1) é denotada por $F_Z(z;\ \lambda)$ e dada por

$$F_Z(z;\ \lambda) = 2\Phi_2(z, 0|\Omega), \qquad com\ \ \Omega = \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}, \qquad \rho = \frac{\lambda}{\sqrt{1 + \lambda^2}}, \qquad z \in \mathbb{R}$$

sendo $\Phi_2(.|\Omega)$ a função de distribuição de uma normal bivariada com média zero e matriz de variância Ω.

Essa distribuição é uniparamétrica, e seu parâmetro λ representa a assimetria da sua função de densidade. Valores positivos de λ indicam uma assimetria positiva e valores negativos de λ indicam uma assimetria negativa. Quando $\lambda = 0$, a densidade (1) se torna simétrica, coincidindo com a densidade da distribuição Normal Padrão.

A densidade em (1) possui algumas propriedades interessantes, cujas provas podem ser obtidas em Azzalini (1985) e Azzalini (2004). Algumas dessas propriedades são apresentadas a seguir:

Propriedades: (2)

- Se $Z \sim SN(\lambda)$ então $Z^2 \sim \chi_1^2$
- Se $Z \sim SN(\lambda)$ então $-Z \sim SN(-\lambda)$
- (Representação estocástica de Henze, 1986) Se $U, V \sim N(0,1)$, independentes, então

$$\frac{\lambda}{\sqrt{1+\lambda^2}}|U| + \frac{1}{\sqrt{1+\lambda^2}}V \sim SN(\lambda)$$

2.2- Distribuição Normal Assimétrica Locação-Escala

Estendendo o modelo citado anteriormente, é apresentada uma variável aleatória $Y \sim SN(\mu, \sigma^2, \lambda)$, com distribuição normal assimétrica locação-escala, que possui três parâmetros, de locação μ, de escala σ^2 e de assimetria λ. Sua função de densidade de probabilidade é:

$$f_Y(y) = \frac{2}{\sigma}\,\phi\left(\frac{y-\mu}{\sigma}\right)\Phi\left(\lambda\frac{y-\mu}{\sigma}\right), y \in \mathbb{R}. \quad (3)$$

A função de distribuição de (3), é denotada por $F_Y(y;\ \mu, \sigma^2, \lambda)$ e dada por

$$F_Y(y;\ \mu, \sigma^2, \lambda) = 2\Phi_2\left(\frac{y-\mu}{\sigma}, 0\middle|\Omega\right), \text{com } \Omega = \begin{bmatrix} 1 & -\rho \\ -\rho & 1 \end{bmatrix}, z \in \mathbb{R}$$

sendo $\Phi_2(.|\Omega)$ a função de distribuição de uma normal bivariada com média zero e matriz de variância Ω.

A média e a variância de uma variável aleatória $Y \sim SN(\mu, \sigma^2, \lambda)$, são expressas por,

$$E(Y) = \mu + \sigma c\rho \quad e \quad Var(Y) = \sigma^2(1 - c^2\rho^2). \quad (5)$$

Com $c = \sqrt{\frac{2}{\pi}}$.

Uma propriedade importante da densidade em (3) é a seguinte:

- Sejam $X \sim SN\ (\mu, \sigma^2, \lambda)$ e $Y = a + bX, a\ e\ b \in \mathbb{R}$. Então:

$$Y \sim SN\ (a + b\mu, b^2\sigma^2, sinal(b)\lambda)$$

Onde $sinal(x) = 1\ se\ x \geq 0\ e\ sinal(x) = -1\ se\ x < 0$. A prova desse resultado pode ser encontrada em Rodríguez (2005).

2.3-Modelo de Regressão Normal Assimétrico

Considere $y_1, \ldots, y_n$ um conjunto de *n* observações independentes. Associado à i-ésima observação, considere o preditor linear $\mu_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$, onde β é um vetor *p*-dimensional de coeficientes de regressão desconhecidos e considerando um vetor *px1* de covariáveis x_i. Temos então o seguinte modelo de regressão

$$\begin{array}{lll} y_i & = & \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i, \\ \varepsilon_i & \sim & SN\ (0, \sigma^2, \lambda), \end{array} \quad i = 1, \ldots, n, \tag{4}$$

onde ε_i são erros independentes. Logo $Y_i \sim SN\ (\boldsymbol{x}_i^T\boldsymbol{\beta}, \sigma^2, \lambda)\ \cdot, i = 1, \ldots, n$.
Note que $E(\varepsilon_i) = c\sigma\rho \neq 0$, para $\lambda \neq 0$. Então para que Y_i seja não-viesado, basta considerar $\varepsilon_i \sim SN\ (-c\sigma\rho, \sigma^2, \lambda)$.

2.4 - Estimações dos Parâmetros via Algoritmo EM

A função verossimilhança para o modelo (4), denotado por $\ell(\boldsymbol{\theta})$, pode ser escrito como:

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{m} \log\left[2\phi(y_i|\boldsymbol{x}_i{}^T\boldsymbol{\beta}, \sigma^2)\Phi\left(\frac{\lambda(y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta})}{\sigma^2}\right)\right]$$
$$= \sum_{i=1}^{m} \log\left[2\int_0^{+\infty} \phi(y_i|\boldsymbol{x}_i{}^T\boldsymbol{\beta}, \sigma^2)\phi(t_i|\lambda(y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta}), \sigma^2)dt_i\right].$$

Uma forma de encontrar as estimativas de máxima verossimilhança dos parâmetros é maximizar a função acima, porem esse método é complicado para esse determinado caso, devido à presença de integrais. Por isso, neste caso a saída é usar o método do algoritmo EM, simplificando função de probabilidade, admitindo a existência de valores adicionais (utilização de modelos hierárquicos).

Seja $\boldsymbol{y}$ o conjunto de dados observados e $\boldsymbol{s}$ denotando o conjunto de dados faltantes. O dado completo $\boldsymbol{y}_c = (\boldsymbol{y}, \boldsymbol{s})$ é $\boldsymbol{y}$ aumentado com $\boldsymbol{s}$. Denota-se por $\ell_c(\boldsymbol{\theta}|\boldsymbol{y}_c)$, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, a função log-verossimilhança dos dados completos e por $Q(\boldsymbol{\theta}|\,\widehat{\boldsymbol{\theta}}) = E[\ell_c(\boldsymbol{\theta}|\boldsymbol{y}_c)|\,\boldsymbol{y}, \widehat{\boldsymbol{\theta}}]$, o valor esperado desta função. Cada iteração do algoritmo EM envolve dois passos, um passo E (esperança) e um passo M (maximização), definidos como:

- Passo E: Calcule $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$ como uma função de $\boldsymbol{\theta}$;
- Passo M: Encontre $\boldsymbol{\theta}^{(k+1)}$ que maximiza $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(k)})$.

Considere $\widehat{\boldsymbol{\theta}}^{(k)} = \left(\beta^{(k)^T}, \sigma^{2^{(k)}}, \lambda^{(k)}\right)^T$ a estimativa de $\boldsymbol{\theta}$ na k-ésima iteração.

Passo E: Dado $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}^{(k)}$, calcule $\hat{t}_j^{\,(k)}$ e $\widehat{t^2}_j^{\,(k)}$, para $j = 1, \ldots, n$.

Passo M: Atualize $\widehat{\boldsymbol{\theta}}^{(k+1)}$ maximizando $Q(\boldsymbol{\theta}|\,\widehat{\boldsymbol{\theta}}^{(k)})$ sob $\boldsymbol{\theta}$, o que leva às seguintes soluções analíticas:

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} - \frac{\lambda^{(k)}}{1 + \lambda^{(k)^2}}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\hat{\boldsymbol{t}}^{(k)},$$

$$\hat{\sigma}^{2^{(k+1)}} = \frac{1}{2n}\left\{\left(1 + \lambda^{(k)^2}\right)\left[Q\left(\boldsymbol{\beta}^{(k)}\right)\right] + \widehat{\boldsymbol{t}^2}^{(k)^T}\mathbf{1}_n - 2\lambda^{(k)}\hat{\boldsymbol{t}}^{(k)^T}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(k)}\right)\right\},$$

$$\hat{\lambda}^{(k+1)} = \frac{\hat{\boldsymbol{t}}^{(k)^T}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(k)}\right)}{Q_\phi(\boldsymbol{\beta}^{(k)})}$$

onde $Q\left(\boldsymbol{\beta}^{(k)}\right) = \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(k)}\right)^T\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^{(k)}\right)$,

e

$\hat{t}_j = \hat{\lambda}\hat{\eta}_j + \hat{\sigma}W_\phi\left(\frac{\hat{\lambda}\hat{\eta}_j}{\hat{\sigma}}\right)$ e $\widehat{t^2}_j = \hat{\lambda}^2\hat{\eta}_j^2 + \hat{\sigma}^2 + \hat{\lambda}\hat{\sigma}\hat{\eta}_j\, W_\phi\left(\frac{\hat{\lambda}\hat{\eta}_j}{\hat{\sigma}}\right)$, $j = 1, \ldots, n$, com $W_\phi(x) = \frac{\phi(x)}{\Phi(x)}$, $\hat{\eta}_j = \hat{\lambda}\left(y_j - \boldsymbol{x}_j^T\widehat{\boldsymbol{\beta}}\right)$, $j = 1, \ldots, n, j \neq i$ e $\hat{\eta}_i = \hat{\lambda}\left(y_i - \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}\right)$, calculado no passo k.

Claramente, se $\gamma = 0$, as equações do Passo M se reduzem às equações obtidas em Ferreira (2008) para o modelo normal assimétrico. Note que, quando $\lambda = 0$, $\widehat{\boldsymbol{\beta}}^{(k+1)} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ e $\hat{\sigma}^{2\,(k+1)} = \frac{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})}{n}$ são os EMV de β e σ^2, respectivamente, do modelo normal simétrico.

Os passos E e M necessários para a implementação do algoritmo EM para encontrar os estimadores de máxima verossimilhança (EMV) dos parâmetros do modelo definido em (4) encontra-se em Ferreira (2008).

2.5 - Matriz de Informação de Fisher Observada

Considerando $Y_1, \ldots, Y_n$ um conjunto de *n* observações independentes, com distribuição $\mathrm{Y_i} \sim \mathrm{SN}\left(\mathrm{x_i^T}\beta, \sigma^2, \lambda\right) \cdot, \mathrm{i} = 1, \ldots, \mathrm{n}$. Sendo β um vetor p-dimensional de coeficientes de regressão desconhecidos.

Seja $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2, \lambda)^T$, a função de log-verossimilhança $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell_i(\boldsymbol{\theta})$ é da forma $\ell_i(\boldsymbol{\theta}) = log2 + \ell_{1i}(\boldsymbol{\theta}) + \log\left[\Phi\left(\ell_{2i}(\boldsymbol{\theta})\right)\right]$, onde $\ell_{1i}(\boldsymbol{\theta}) = \phi(y_i|\boldsymbol{x}_i{}^T\boldsymbol{\beta}, \sigma^2)$ e $\ell_{2i}(\boldsymbol{\theta}) = \lambda\frac{y_i - \boldsymbol{x}_i{}^T\boldsymbol{\beta}}{\sigma}$.

Logo, a primeira derivada de $\ell_i(\boldsymbol{\theta})$é dada por

$$\frac{\partial\, \ell_i(\boldsymbol{\theta})}{\partial \psi} = \frac{\partial\, \ell_{1i}(\boldsymbol{\theta})}{\partial \psi} + W_\Phi\left(\ell_{2i}(\boldsymbol{\theta})\right)\frac{\partial\, \ell_{2i}(\boldsymbol{\theta})}{\partial \psi}, \psi = \boldsymbol{\beta}, \sigma^2, \lambda$$

E a segunda derivada é

$$\frac{\partial^2\, \ell_i(\boldsymbol{\theta})}{\partial \gamma \partial \psi^T} = \frac{\partial^2\, \ell_{1i}(\boldsymbol{\theta})}{\partial \gamma \partial \psi^T} + W_\Phi\left(\ell_{2i}(\boldsymbol{\theta})\right)\frac{\partial^2\, \ell_{2i}(\boldsymbol{\theta})}{\partial \gamma \partial \psi^T} +, W_\Phi^1\left(\ell_{2i}(\boldsymbol{\theta})\right)\frac{\partial\, \ell_{2i}(\boldsymbol{\theta})}{\partial \gamma}\frac{\partial\, \ell_{2i}(\boldsymbol{\theta})}{\partial \psi^T}$$

Onde $W_\Phi^1(x) = -W_\Phi(x)\left(x + W_\Phi(x)\right)$ é a derivada de $W_\Phi\,(x)$.

Logo, a matriz de informação observada para $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \sigma^2, \lambda)^T$ é dada por

$$\mathbf{I}(\boldsymbol{\theta}) = \mathrm{I}_1(\boldsymbol{\theta}) + \mathrm{I}_2(\boldsymbol{\theta}) \qquad (5)$$

$$\mathrm{I_k}(\boldsymbol{\theta}) = \begin{pmatrix} \mathrm{I}^{\mathrm{k}}_{\beta\beta} & \mathrm{I}^{\mathrm{k}}_{\sigma^2\beta} & \mathrm{I}^{\mathrm{k}}_{\lambda\beta} \\ & \mathrm{I}^{\mathrm{k}}_{\sigma^2\sigma^2} & \mathrm{I}^{\mathrm{k}}_{\lambda\sigma^2} \\ & & \mathrm{I}^{\mathrm{k}}_{\lambda\lambda} \end{pmatrix} \text{ para k = 1, 2 e } \gamma, \psi = \boldsymbol{\beta}, \sigma^2, \lambda \text{ , onde}$$

$$\mathrm{I}^{1}_{\beta\beta} = \frac{1}{\sigma^2}\mathbf{X}^{\mathrm{T}}\mathbf{X}\,,$$

$$\mathrm{I}^{1}_{\sigma^2\beta} = \frac{1}{\sigma^4}\mathbf{X}^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\,,$$

$$\mathrm{I}^{1}_{\sigma^2\sigma^2} = -\frac{\mathrm{n}}{2\sigma^4} + \frac{1}{\sigma^6}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\,,$$

$$\mathrm{I}^{1}_{\lambda\boldsymbol{\Gamma}} = \mathbf{0}, \boldsymbol{\Gamma} = \boldsymbol{\beta}, \sigma^2, \lambda\,,$$

$$\mathrm{I}^{2}_{\beta\beta} = -\frac{\lambda^2}{\sigma^2}\mathbf{X}^{\mathrm{T}}\mathbf{D}\left(W^{1}_{\Phi}\left[\lambda\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma}\right]\right)\mathbf{X}\,,$$

$$\mathrm{I}^{2}_{\sigma^2\beta} = -\frac{\lambda}{2\sigma^3}\mathbf{X}^{\mathrm{T}}W_{\Phi}\left[\lambda\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma}\right] - \frac{\lambda^2}{2\sigma^4}\mathbf{X}^{\mathrm{T}}\mathbf{D}\left(W^{1}_{\Phi}\left[\lambda\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma}\right]\right)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\,,$$

$$\mathrm{I}^{2}_{\lambda\beta} = \frac{1}{\sigma}\mathbf{X}^{\mathrm{T}}W_{\Phi}\left[\lambda\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma}\right] + \frac{\lambda}{\sigma^2}\mathbf{X}^{\mathrm{T}}\mathbf{D}\left(W^{1}_{\Phi}\left[\lambda\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma}\right]\right)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\,,$$

$$\mathrm{I}^{2}_{\sigma^2\sigma^2} = -\frac{3\lambda}{4\sigma^5}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}W_{\Phi}\left[\lambda\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma}\right] - \frac{\lambda^2}{4\sigma^6}\mathrm{Q_w}(\boldsymbol{\beta})\,,$$

$$\mathrm{I}^{2}_{\lambda\sigma^2} = \frac{1}{2\sigma^3}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}W_{\Phi}\left[\lambda\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma}\right] + \frac{\lambda}{2\sigma^4}\mathrm{Q_w}(\boldsymbol{\beta}),$$

$$\mathrm{I}^{2}_{\lambda\lambda} = -\frac{1}{\sigma^2}\mathrm{Q_w}(\boldsymbol{\beta}),$$

onde $\mathbf{D}(\mathbf{a})$ é a matriz diagonal do vetor **a** e $\mathrm{Q_w}(\boldsymbol{\beta}) = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{D}\left(W^{1}_{\Phi}\left[\lambda\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\sigma}\right]\right)(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})$.

3- Seleção de variáveis

3.1- Critério de Informação Akaike (AIC)

A ideia principal desse critério é encontrar o modelo mais parcimonioso, ou seja, o modelo que melhor se ajusta aos dados, com um menor número de parâmetros. E esse critério é usado sem envolver testes estatísticos.

Essa função foi proposta por Akaike (1974), onde ele se baseia na função de log-verossimilhança acrescida de uma penalidade pelo número de parâmetros do modelo.

Entre vários modelos candidatos, deve-se escolher aquele que apresentar o menor valor da função AIC.

$$AIC = -2l(\boldsymbol{\theta}) + 2p \qquad (6)$$

onde p é o número de parâmetros e $l(\boldsymbol{\theta})$ é a função de log-verossimilhança.

Outro critério que possui o mesmo raciocínio que o AIC é o critério de Informação BIC. Neste caso, a seleção do melhor modelo também é baseada no menor BIC.

$$BIC = -2l(\boldsymbol{\theta}) + p * log(n) \qquad (7)$$

3.2- Seleção de variáveis usando o método Backward via AIC

O método Backward via AIC inicia o processo com todas as variáveis auxiliares do modelo proposto e depois, por etapas, cada uma pode ou não ser eliminada.

A decisão de retirada da variável é tomada baseando-se no método AIC, que são calculados para cada etapa, referentes àquele modelo.

- Procedimento:

Passo 1: Calcula-se o AIC para o modelo completo, que é denominado AIC_{TOTAL}.

$$AIC_{TOTAL} = -2l(\boldsymbol{\theta}) + 2p$$

Passo 2: Calcula-se o AIC retirando-se uma variável de casa vez, ou seja, teremos p-1 valores de AIC:

AIC sem a variável $x_i,\ para\ i = 2, \ldots . p$:

$$AIC_i = -2l(\boldsymbol{\theta})_i + 2_{(p_i - 1)}$$

onde pi é o número de parâmetros no i-ésimo passo.

Passo 3: Compara-se o AIC_{TOTAL} com o menor AIC_i, ou seja, se $\min(AIC_i) < AIC_{TOTAL}$, deve-se retirar a variável referente ao $\min(AIC_i)$, e $AIC_{TOTAL} = \min(AIC_i)$.

Passo 4: Deve-se repetir os passos 2 e 3 até que $\min(AIC_i) > AIC_{TOTAL}$.

4- Detecção de Multicolinearidade em Variáveis Regressoras:

A multicolinearidade é um problema comum em regressão linear múltipla, ela indica que existe uma relação de linearidade entre as variáveis regressoras, prejudicando assim a estimação dos coeficientes de regressão. O problema de multicolinearidade torna a estimativa dos parâmetros imprecisa, por conta de um alto valor do erro padrão, o que não é conveniente estatisticamente.

Existem vários métodos propostos para detectar a multicolinearidade entre as variáveis explicativas do modelo de regressão.

Quando o coeficiente de determinação R_i^2 apresenta um alto valor, mas nenhum dos coeficientes da regressão é estatisticamente significativo, existe um indício de multicolinearidade. Ou seja, uma das formas de detecção é observar o valor do coeficiente de determinação da variável regressão X_i.

Isso é feito através do calculo do *fator de inflação da variância* (VIF) (Berk, 1977). Esse fator mede o grau em que cada variável explicativa do modelo é explicada pelas demais variáveis independente, dado por:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{8}$$

Pode-se observar que, quanto maior o R_i^2, maior é o valor de VIF, indicando alta colinearidade.

Valores de VIF_i maiores que 10 correspondem a um coeficiente de determinação $R_i^2 >$ 0,90 e isso é considerado inaceitável. (Kutner et al., 2004; Tamhane & Dunlop, 2000). Por isso, usaremos esse critério para detectar a multicolinearidade nos dados.

5- Estudo de Simulação

Nesta seção será realizado um estudo de simulação para avaliar o método de seleção de variáveis Bacward via AIC.

Foi simulada uma amostra com 10 variáveis ($p = 11$) de tamanho $n = 100$, considerando o modelo de regressão linear assimétrico da forma $\boldsymbol{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, onde $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$ é um vetor de erros independentes, com $\varepsilon_i \sim SN(0, \sigma^2, \lambda)$, β é um vetor p-dimensional de coeficientes de regressão e X é uma matriz regressora $n \times p$.

Definimos na simulação $\sigma^2 = 1$, $\lambda = 4$ e o vetor de parâmetros $\boldsymbol{\beta} = (5.0,\ 1.1,\ 0.03,\ 1.3,\ 1.23,\ 0.02,\ 1.4,\ 0.01,\ 0.02,\ 0.04,\ 0.023)^T$.

A matriz X foi gerada a partir de uma distribuição normal multivariada de tamanho 100, com $X \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, onde $\boldsymbol{\mu}$ é um vetor de médias simulado a partir de uma distribuição normal padrão de tamanho 10 e $\boldsymbol{\Sigma}$ é a matriz de covariância que possui os seguintes valores:

$$\boldsymbol{\Sigma} = \begin{vmatrix} \mathbf{1} & \mathbf{-0.84} & \mathbf{0.97} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{-0.84} & \mathbf{1} & \mathbf{-0.79} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \mathbf{0.97} & \mathbf{-0.79} & \mathbf{1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{1} & \mathbf{-0.84} & \mathbf{0.97} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{-0.84} & \mathbf{1} & \mathbf{-0.79} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{0.97} & \mathbf{-0.79} & \mathbf{1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{0} & \mathbf{2.1} & \mathbf{0} & \mathbf{0} \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{2.1} \end{vmatrix}$$

Propositalmente a matriz de variáveis regressoras (**X**) foi simulada com problemas de multicolinearidade e os valores do vetor de parâmetros $\boldsymbol{\beta}$ foram simulados bem próximos de zero, para que algumas variáveis regressoras apresentem uma baixa correlação com a variável dependente Y.

O objetivo do estudo de simulação é verificar se o método de seleção estudado nesta monografia é capaz de detectar este tipo de problema ou se é necessária uma analise anterior à seleção, como a retirada de variáveis multicolineares e/ou com problemas de baixa correlação com a variável dependente.

Em seguida são apresentadas as tabelas de correlação entre as variáveis regressoras e a variável dependente e o calculo do *fator de inflação da variância* (VIF) das variáveis regressoras.

Tabela 1: Correlação das variáveis regressoras com a variável dependente, no estudo de simulação.

Variáveis Regressoras (X_i,)	$\boldsymbol{Correla}$ção$\boldsymbol{(X_i, Y)}$
X_1	0.6695
X_2	-0.4517
X_3	0.6544
X_4	0.7249
X_5	-0.6013
X_6	0.7123
X_7	**0.0510**
X_8	**0.0144**

X_9	0.1869
X_{10}	**-0.0293**

Tabela 2: Cálculo do *fator de inflação da variância* das variáveis regressoras, no estudo de simulação.

Variáveis Regressoras (X_i)	$\boldsymbol{VIF}$
X_1	**16.164**
X_2	2.837
X_3	**13.509**
X_4	**27.621**
X_5	3.747
X_6	**21.676**
X_7	1.076
X_8	1.035
X_9	1.153
X_{10}	1.091

Nosso estudo de simulação ficou dividido em quatro estratégias de seleção, que estão descritas a seguir:

Estratégia 1:

Nesta estratégia o método de seleção de variáveis Backward via AIC será aplicado em todas as variáveis da matriz de regressoras $\boldsymbol{X}$, sem considerar nenhuma análise anterior dos dados.

O modelo selecionado pelo método backward via AIC foi:

Tabela 3: Resultados da seleção de variáveis pelo método Backward via AIC (Estratégia 1), com $z = \frac{\ddot{\theta}}{\widehat{EP}}$.

Parâmetros	Simulado	Estimativas	Erro Padrão	z
β_0	5.00	5.023	0.539	9.324
β_1	1.10	0.961	0.231	4.152
β_2	0.03	0.158	0.084	1.865
β_3	1.30	1.545	0.216	7.162
β_4	1.23	1.241	0.234	5.297
β_6	1.40	1.342	0.212	6.327
β_7	0.01	0.171	0.049	3.475
β_{10}	0.023	0.064	0.038	1.664
σ^2	1.00	0.831	0.167	4.963
λ	4.00	4.506	2.304	1.956

Podemos observar que foram selecionadas variáveis não significativas ($|z| < 2$), considerando um nível de significância de 4,56% e variáveis que possuem baixa correlação com a variável dependente (X_7 *e* X_{10}).

Outro problema foi a presença de multicolinearidade entre as variáveis explicativas. Como o valor do *fator de inflação da variância* (VIF) muda a cada retirada de alguma variável regressora, deve-se calcular o mesmo a cada estratégia.

Podemos observar na Tabela 4 que existem valores do VIF maiores do que 10 para o modelo selecionado.

Tabela 5: Cálculo do *fator de inflação da variância* das variáveis regressoras do modelo selecionado pela estratégia 1.

Variáveis Regressoras (X_i)	$\boldsymbol{VIF}$
X_1	**15.868**
X_2	2.776
X_3	**13.141**
X_4	**20.466**
X_6	**20.257**
X_7	1.044
X_{10}	1.054

Estratégia 2:

Nesta estratégia, antes da aplicação do método de seleção de variáveis Backward via AIC, serão eliminadas as variáveis regressoras que possuem baixa correlação com a variável dependente. Foram retiradas as variáveis correspondentes a correlações menores que 10 %.

As variáveis retiradas foram as X_7, X_8 e X_{10}. Então a matriz regressora considerada no método de seleção não contém essas variáveis.

O modelo selecionado pelo método bacward via AIC foi:

Tabela 6: Resultados da seleção de variáveis pelo método Backward via AIC (Estratégia 2), com $z = \frac{\widehat{\theta}}{\widehat{EP}}$.

Parâmetros	Simulado	Estimativas	Erro Padrão	z
β_0	5.00	4.470	0.544	8.214
β_1	1.10	1.077	0.217	4.987
β_2	0.03	0.146	0.094	1.553
β_3	1.30	1.413	0.214	6.604
β_4	1.23	1.474	0.231	6.375
β_6	1.40	1.139	0.204	5.571
σ^2	1.00	0.931	0.200	4.648
λ	4.00	4.187	2.284	1.834

Podemos observar que foram selecionadas variáveis não significativas ($|z| < 2$), considerando um nível de significância de 4,56% e o problema de multicolinearidade persistiu, ou seja, ele não eliminou as variáveis que possivelmente podem causar algum problema para o modelo.

Tabela 7: Cálculo do *fator de inflação da variância* das variáveis regressoras do modelo selecionado pela estratégia 2.

.

Variáveis Regressoras (X_i,)	$\boldsymbol{VIF}$
X_1	**15.656**
X_2	2.757
X_3	**12.806**
X_4	**20.002**
X_6	**19.995**

Estratégia 3:

Nesta estratégia, antes da aplicação do método de seleção de variáveis Backward via AIC, serão eliminadas as variáveis regressoras que possuem problemas de multicolinearidade, ou seja, serão analisados os valores do *fator de inflação da variância* (VIF).

Essa eliminação será da seguinte forma:

- Calcula-se o VIF para todas as variáveis explicativas, e retira-se aquela que possui o maior valor, necessariamente superior a 10.
- Repete-se o passo anterior até que todos os VIF's estejam menores do que 10.

Depois de aplicar esse procedimento, permaneceram as seguintes variáveis com VIF's menores do que 10.

Tabela 8: Variáveis Regressoras que apresentam *VIF's* menores que 10.

Variáveis Regressoras (X_i,)	$\boldsymbol{VIF}$
X_2	2.322
X_3	2.364
X_5	2.912
X_6	2.939
X_7	1.066
X_8	1.026
X_9	1.084
X_{10}	1.047

Então a matriz regressora considerada no método de seleção contém somente essas variáveis citadas na tabela anterior.

O modelo selecionado pelo método bacward via AIC foi:

Tabela 9: Resultados da seleção de variáveis pelo método Backward via AIC (Estratégia 3), com $z = \frac{\widehat{\theta}}{\widehat{EP}}$.

Parâmetros	Simulado	Estimativas	Erro Padrão	Z
β_0	5.00	7.101	0.106	66.943
β_3	1.30	2.263	0.061	37.193
β_5	0.02	-0.145	0.094	-1.547
β_6	1.40	2.398	0.089	27.026
β_7	0.01	0.230	0.057	4.013
β_9	0.04	0.121	0.067	1.812
β_{10}	0.023	0.068	0.043	1.575
σ^2	1.00	1.326	0.221	5.99
λ	4.00	6.586	3.049	2.159

Podemos observar que foram selecionadas variáveis não significativas ($|z| < 2$), considerando um nível de significância de 4,56% e também foram selecionadas variáveis com problema de baixa correlação entre a variável dependente e a variável explicativa. A seleção não eliminou as variáveis X_7 e X_{10}, que na tabela 10 mostraram baixa correlação.

Estratégia 4:

Nesta estratégia, antes da aplicação do método de seleção de variáveis Backward via AIC, serão eliminadas as variáveis regressoras que possuem baixa correlação com a variável dependente e as que possuem problemas de multicolinearidade, nesta ordem.

As variáveis retiradas por baixa correlação foram as X_7, X_8 e X_{10}. Logo, as demais variáveis regressoras foram consideradas na análise dos valores do *fator de inflação da variância* (VIF). O procedimento foi o mesmo da estratégia 3, e as variáveis eliminadas por apresentarem problemas de multicolinearidade foram X_1 e X_4.

Tabela 11: Cálculo do *fator de inflação da variância* das variáveis regressoras usadas na estratégia 4.

Variáveis Regressoras (X_i,)	***VIF***
X_2	2.252
X_3	2.267
X_5	2.882
X_6	2.873
X_9	1.041

Logo, a matriz regressora considerada no método de seleção contém as variáveis X_2, X_3, X_5, X_6 e X_9.

O modelo selecionado pelo método bacward via AIC foi:

Tabela 12: Resultados da seleção de variáveis pelo método Backward via AIC (Estratégia 4), com $z = \frac{\widehat{\theta}}{\widehat{EP}}$.

Parâmetros	Simulado	Estimativas	Erro Padrão	Z
β_0	5.00	7.264	0.183	39.711
β_3	1.30	2.329	0.071	32.699
β_5	0.02	-0.265	0.129	-2.055
β_6	1.40	2.342	0.122	19.157
σ^2	1.00	1.174	0.315	3.724
λ	4.00	2.336	1.143	2.044

Neste caso todas variáveis selecionadas foram significativamente diferentes de zero ($|z| > 2$) e o problema de multicolinearidade e baixa correlação foram eliminados anteriormente. Portanto este modelo é considerado o mais parcimonioso.

Observando o gráfico de envelope (Figura 1) a seguir, pode-se dizer que o modelo final selecionado pela estratégia 4 realmente se ajusta bem aos dados.

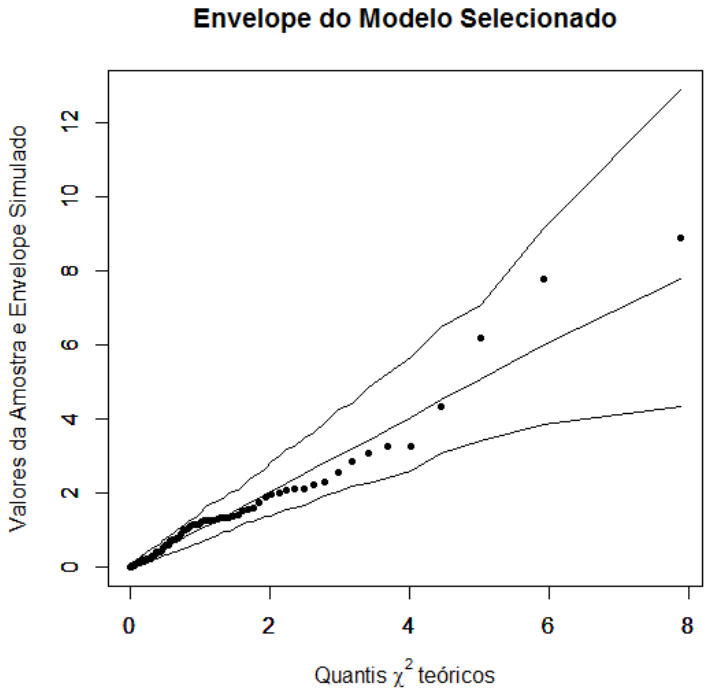

Figura 1: Envelope do modelo selecionado na simulação pela estratégia 4

Tabela 13: Valores da função de log-verossimilhanças e do Critério de Informação Akaike referentes as quatro estratégias de seleção do estudo de simulação.

Estratégias	$l(\widehat{\boldsymbol{\theta}})$	AIC	BIC
1	-79.163	178.325	204.377
2	-85.917	187.833	208.675
3	-98.061	214.122	237.569
4	-108.994	229.987	245.618

Analisando a tabela anterior, pode-se notar que o AIC e o BIC não foram coerente com a nossa percepção do melhor modelo. O modelo final (Estratégia 4) não apresentou o menor AIC ou BIC, porém ele foi selecionado considerando os nossos critérios de seleção.

6- Aplicação em Dados Reais

Nesta seção, o método de seleção de variáveis Backward via AIC para modelos de regressão lineares assimétricos (Estratégia 4) será aplicado às variáveis do banco de dados *Australian Institute of Sport data (AIS)*. Esse banco tem informações coletadas sobre 202 atletas (102 homens e 100 mulheres), no quais foram coletadas diversas medidas referentes a peso, altura, gordura corporal, substancia no sangue, soma das dobras cutâneas, dentre outras. Os dados podem ser baixados no endereço http://azzalini.stat.unipd.it/index-en.html.

A variável resposta considerada foi a soma das medidas das dobras cutâneas (ssf), a escolha foi baseada na sua assimetria e na sua boa interpretação. As variáveis explicativas utilizadas foram:

- sex: Feminino ou Masculino
- rcc: Números de células vermelhas
- wcc: Números de células brancas
- Hc: Hematócritos
- Hg: Hemoglobina
- Fe: Concentração de ferro no plasma
- bmi: Índice de Massa Corporal $\left({Peso}/{Altura^2}\right)$
- Bfat: Porcentagem de gordura corporal
- lbm: Massa corporal magra
- Ht: Altura (cm)
- Wt: Peso (Kg)

Como foi mostrado no estudo de simulação da seção anterior, é necessário que haja uma análise do banco de dados antes de usar o método de seleção de variáveis Backward via AIC, pois só ele não é capaz de eliminar os problemas de multicolinearidade das variáveis explicativas e o problema de baixa correlação da variável dependente com as explicativas.

Por isso foi feita uma análise exploratória do banco de dados, com o objetivo de eliminar esses principais problemas.

Analisando somente a correlação, e utilizando o mesmo critério do estudo de simulação (Capítulo 5-), pode-se observar na Tabela 14 que as variáveis explicativas Fe e Ht devem ser excluídas, e as demais devem permanecer na seleção de variáveis.

Tabela 12: Correlação das variáveis regressoras com a variável dependente, do bando de dados ais.

Variáveis Explicativas (xi)	Corr(xi,Y)
sex	-0.547
rcc	-0.403
wcc	0.137
Hc	-0.449
Hg	-0.435
Fe	**-0.108**
bmi	0.321
Bfat	0.963
lbm	-0.208
Ht	**-0.071**
Wt	0.154

O próximo passo é detectar o problema de multicolinearidade entres as variáveis explicativas, através do calculo do *fator de inflação da variância* (VIF) (Berk, 1977). Será considerada a mesma referência do estudo de simulação, ou seja, serão retiradas as variáveis que apresentarem VIF maior do que 10, retirando uma de cada vez, partindo do maior valor. O processo se encerrará quando todas as variáveis apresentarem valores menores que 10. As variáveis retiradas por apresentarem baixa correlação não entram nesse procedimento.

Pode-se observar na Tabela 13 que as variáveis explicativas Hc e lbm foram excluídas por apresentarem problemas de multicolinearidade.

Logo, as variáveis selecionadas por apresentarem VIF menores do que 10, e por isso não considerada com presença de multicolinearidade, são:

Tabela 13: Variáveis regressoras do bando ais que apresentam *Fator de Inflação da Variância* menor que 10.

Variáveis Regressoras (X_i)	VIF
sex	5.318
rcc	4.944
wcc	1.079
Hg	6.207
bmi	4.467
Bfat	3.645
Wt	5.282

Seleção de variáveis via Bacward- AIC

O método de seleção de variáveis Backward via AIC será aplicado somente nas variáveis explicativas Sexo (sex), *Números de células vermelhas (rcc), Números de células brancas (wcc), Hemoglobina (Hg), Índice de Massa Corporal (bmi), Porcentagem de gordura corporal (Bfat) e Altura (Wt).*

O modelo selecionado pelo método Bacward via AIC foi:

Tabela 14: Resultados da seleção de variáveis do bando ais pelo método Backward via AIC, com $z = \frac{\hat{\theta}}{\widehat{EP}}$.

Parâmetros	Estimativas	Erro Padrão	Z
β_0	-18.869	7.273	-2.594
sex	13.419	1.854	7.237
Hg	-1.114	0.525	-2.121
bmi	0.766	0.231	3.316
Bfat	5.546	0.174	31.873
σ^2	66.655	19.610	3.399
λ	1.121	0.599	1.900

A verossimilhança do modelo selecionado foi $\ell(\hat{\theta}) = -665.9709$ e AIC=1345.942. Podemos observar pelos gráficos a seguir que realmente existe uma forte relação entre as variáveis regressoras selecionadas com a variável dependente ssf. Além disso, é apresentado o envelope do modelo final selecionado, mostrando que o modelo se ajusta bem aos dados.

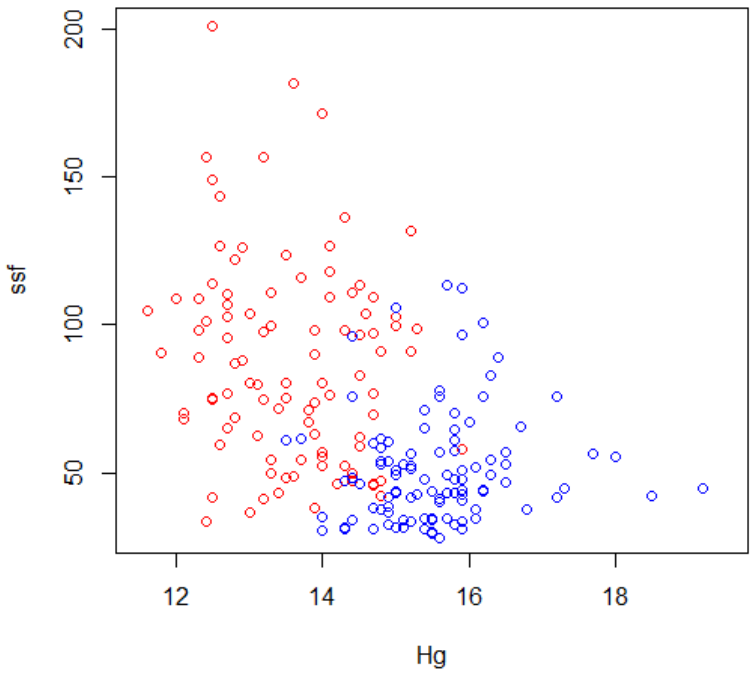

Figura 2: Gráfico de Dispersão da Soma das medidas das dobras cutâneas (ssf) versus Hemoglobina (Hg).

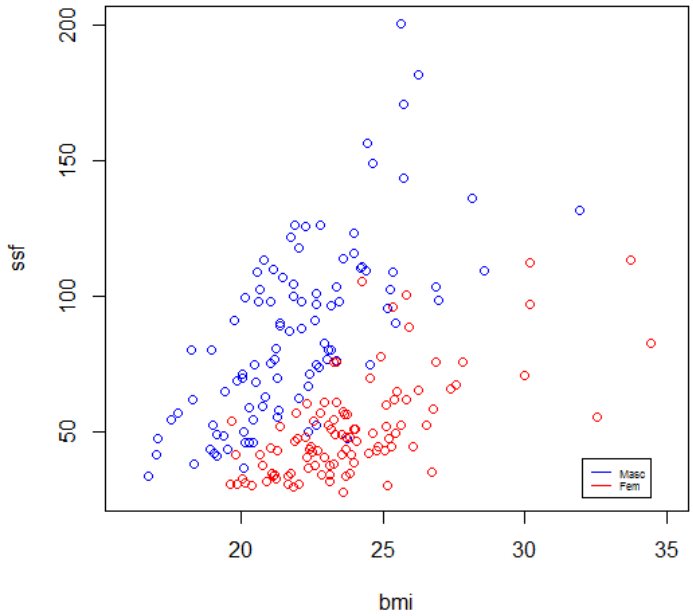

Figura 3: Gráfico de Dispersão da Soma das medidas das dobras cutâneas (ssf) versus Índice de Massa Corporal (bmi).

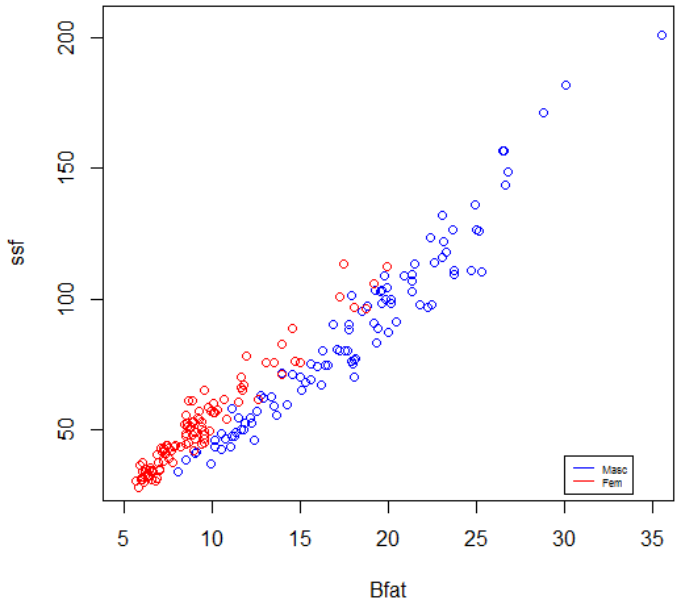

Figura 4: Gráfico de Dispersão da Soma das medidas das dobras cutâneas (ssf) versus Porcentagem de gordura corporal (Bfat).

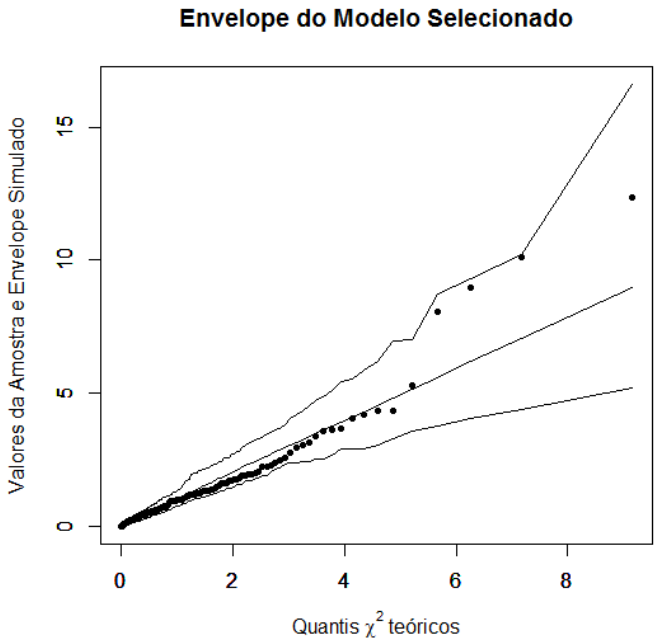

Figura 5: Envelope do modelo selecionado do banco de dados ais.

7- Conclusão

Neste trabalho, apresentamos o método de seleção de variáveis backward, via Critério de Informação Akaike, em um modelo de regressão linear com erros seguindo uma distribuição Normal Assimétrica.

Concluímos que o método de seleção de variáveis Backward via AIC não é capaz de eliminar todos os problemas relacionados as variáveis , prejudicando assim a qualidade do ajuste do modelo. Baseado no estudo de simulação é recomendável fazer uma análise exploratória de dados, como verificar a correlação da variável dependente com as variáveis regressoras e em seguida eliminar o problema de multicolinearidade, para então aplicar o método de seleção de variáveis Backward via AIC.

8- Perspectivas Futuras

A partir do resultado desta monografia seria interessante construir um Intervalo de Confiança de Predição para novas observações. Espera-se que na presença de multicolinearidade esses intervalos fiquem maiores. Para tanto, a distribuição do Erro de Predição deveria ser calculada.

9- Referências Bibliográficas

1 Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19, 716-723.

2 Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal Statistics*, 12, 171-178.

3 Azzalini, A. (2005). The skew-normal distribution and related multivariate families (with discussion). *Scandinavian Journal Statistics*, 32, 159-188.

4 Azzalini, A. & Dalla-Valle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, 83, 715-726.

5 Azzalini, A., Capitanio, A. (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society*, ser. B, 61, 579-602.

6 Azzalini, A. (2004). The skew-normal distribution and related multivariate families.

7 Berk, K. N. (1977). "Tolerance and condition in regression computations", *Journal of the American Statistical Association*, 72 (360), 863-866.

8 Conover, W.J. (1998). Practical Nonparametric Statistics, John Wiley & Sons.

9 Kutner, M. H. et al. Applied linear models. 5th ed. New York: McGraw-Hill Irwin, 2004

10 Tamhane, A. C. & Dunlop D. D. Statistics and Data Analysis – from elementary to, intermediate. Upper Saddle River: Prentice-Hall, 2000.

11 Ferreira. C.S. (2008). *Inferência e diagnóstico em modelos assimétricos.* Tese de Doutorado. Departamento de Estatística. IME-USP. São Paulo.

12 Cook. R.D. e Weisberg. S. (1994). *An Introduction to Regression Graphics*. London: Chapman and Hall.

13 Rodríguez. C.L.B. (2005). *Inferência bayesiana no modelo normal assimétrico*. Dissertação de mestrado. Departamento de Estatística. IME-USP. São Paulo.