

MODELOS DE CONJUNTO *FUZZY* E DE REGRESSÃO
LOGÍSTICA PARA DADOS DE OBESIDADE

Diego Augusto Queijo

Monografia apresentada à Universidade Estadual Paulista “Júlio de Mesquita Filho” para a obtenção do título de Bacharel em Física Médica.

BOTUCATU
São Paulo
Dezembro – 2010

MODELOS DE CONJUNTO *FUZZY* E DE REGRESSÃO
LOGÍSTICA PARA DADOS DE OBESIDADE

Diego Augusto Queijo

Orientadora: Prof. Dr. **Lician Vaz de Arruda Silveira**

Monografia apresentada à Universidade Estadual Paulista “Júlio de Mesquita Filho” para a obtenção do título de Bacharel em Física Médica.

BOTUCATU
São Paulo
Dezembro – 2010

FICHA CATALOGRÁFICA ELABORADA PELA SEÇÃO TÉC. AQUIS. TRATAMENTO DA INFORM.
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - CAMPUS DE BOTUCATU - UNESP
BIBLIOTECÁRIA RESPONSÁVEL: *ROSEMEIRE APARECIDA VICENTE*

Queijo, Diego Augusto.

Modelo de conjuntos fuzzy e de regressão logística para dados de
obesidade / Diego Augusto Queijo. - Botucatu, 2010

Trabalho de conclusão de curso (bacharelado - Física Médica) - Instituto de
Biociências de Botucatu, Universidade Estadual Paulista, 2010

Orientador: Liciania Vaz de Arruda Silveira

Capes: 40600009

1. Saúde coletiva. 2. Epidemiologia – Pesquisa. 3. Conjuntos difusos.
4. Nutrição.

Palavras-chave: Composição corporal; Epidemiologia; Índice de massa
corporal; Modelos de conjuntos fuzzy; Modelo de regressão logística; Nutrição.

Dedicatória

A minha mãe, meu irmão e minha cachorrinha,

Marta, Gabriel e Nina

Por todos os momentos de confiança, força e sabedoria.

Agradecimentos

Agradeço primeiramente e acima de tudo a Deus, pois nada seria possível se não fosse pela vontade Dele.

Em especial, a minha mãe e meu irmão, que sempre agiram como verdadeiros alicerces na construção do ser humano que me tornei. Agradeço por estarem sempre ao meu lado, independentemente da situação, seja boa ou ruim. Também agradeço à meu pai que, apesar do distanciamento, sempre esteve presente em meu coração.

À Profa. Dra. Liciane Vaz de Arruda Silveira pela orientação, sabedoria, paciência e acima de qualquer coisa, pela amizade criada ao longo do estágio, e que com certeza permanecerá para sempre.

Ao Prof. Dr. José Raimundo, pela disposição em sempre me receber em sua sala, mesmo quando ocupado, e escutar alguns dos pensamentos, dúvidas e problemas que permeavam minha mente.

Aos Departamentos de Bioestatística e Biofísica, sem os quais não teria tido a estrutura necessária para ter desenvolvido a capacidade de avaliar, pensar e criar situações favoráveis a solução de problemas, das mais diversas áreas associadas a Física Médica e Ciências em geral.

À V Turma do curso de Física Médica da UNESP de Botucatu; muitos tiveram participação em momentos excepcionais, com suas amizades, parcerias nos estudos e por todos os momentos agradáveis que passamos juntos. Eles sabem quem são.

À Lucas Francisco Carmelo Guimarães, colega da V Turma do curso de Física Médica, o qual foi imprescindível, juntamente com sua família, desde meu

acolhimento no início do curso até o final do mesmo.

À Lucas Massaroppe, que em tantos momentos me ajudou com seu brilhantismo, dividindo conhecimento, materiais, conversas e sua amizade.

À Victor Alessandro Abib Pastore por ter sido temporariamente minha família em Botucatu, por todos os conselhos, risadas, conversas e pela bondade que sempre transpareceu de seu coração.

À Lorena Carla Oliveira da Costa por ter feito crescer em mim sonhos, planos e sentimentos que eu jamais havia pensado em conceber; por todos os momentos juntos, pelas conversas, apoio, sorrisos, fé...

Porfim, agradeço o suporte financeiro cedido pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP).

Sumário

	Página
LISTA DE FIGURAS	viii
LISTA DE TABELAS	ix
RESUMO	x
1 INTRODUÇÃO	1
2 MODELO DE REGRESSÃO LOGÍSTICA	3
2.1 Linhas Gerais	3
2.2 Estimativa dos Parâmetros	8
2.3 Teste de significância para os parâmetros.	11
3 TEORIA DOS CONJUNTOS FUZZY	14
3.1 Introdução	14
3.1.1 O que é a lógica <i>fuzzy</i> ?	14
3.1.2 Porque usar a lógica <i>fuzzy</i> na Epidemiologia?	15
3.2 Conjuntos Fuzzy	16
3.3 Variáveis Linguísticas	22
3.4 Regras Fuzzy	22
3.5 Relações Fuzzy	23
3.6 Composição de Relações Fuzzy	24
3.6.1 Modelo linguístico tipo Mamdani	25
3.7 Métodos de defuzificação	26

	vii
4 BREVE INTRODUÇÃO A CURVA ROC	28
5 ANÁLISE DOS DADOS	30
5.1 Regressão Logística	33
5.2 Modelo Fuzzy	34
5.3 Curva ROC	44
6 DISCUSSÃO	47
REFERÊNCIAS BIBLIOGRÁFICAS	49

Lista de Figuras

	Página
1 <i>Gráfico de coluna para representar a porcentagem de indivíduos segundo estado nutricional e categoria de idade. Teste qui-quadrado com valor de $p = 0,0027$.</i>	31
2 <i>Gráfico de coluna para representar a porcentagem de indivíduos segundo estado nutricional e categoria de ano de escolaridade. Teste qui-quadrado com valor de $p = 0,3575$.</i>	31
3 <i>Gráfico de coluna para representar a porcentagem de indivíduos segundo estado nutricional e categoria de porcentagem de gordura corporal. Como a pressuposição do teste qui-quadrado não foi satisfeita, usou-se o teste exato de Fisher que forneceu um valor de $p < 0,0001$.</i>	32
4 <i>Gráfico de coluna para representar a porcentagem de indivíduos segundo estado nutricional e categoria de razão cintura quadril. Teste qui-quadrado com valor de $p < 0,0001$.</i>	32
5 <i>Gráfico da análise do IMC levando em consideração as variáveis idade e ano de escolaridade.</i>	43
6 <i>Superfícies de indicação de obesidade pelo modelo lingüístico fuzzy.</i>	43
7 <i>Gráfico da curva ROC comparando o modelo de regressão logística e o modelo fuzzy para as variáveis idade e ano de escolaridade.</i>	44
8 <i>Gráfico da curva ROC comparando o modelo de regressão logística e o modelo fuzzy para as variáveis razão cintura quadril e porcentagem de gordura corporal.</i>	45

Lista de Tabelas

	Página
1 <i>Análise descritiva inicial das variáveis utilizadas usando o software SAS</i>	30
2 <i>Parâmetros, grau de liberdade (g), razão das chances (OR), estimativas, erro-padrão, estatística de Wald (Wald) e valor p para os parâmetros segundo o ajuste do modelo de regressão logística (1).</i>	33
3 <i>Parâmetros, grau de liberdade (g), razão das chances (OR), estimativas, erro-padrão, estatística de Wald (Wald) e valor p para os parâmetros segundo o ajuste do modelo de regressão logística (2).</i>	33
4 <i>Modelos testados, área abaixo da curva, erro padrão (SE) e significância assintótica (SA) obtidos através do software SPSS, levando em conta as variáveis idade e ano de escolaridade.</i>	44
5 <i>Modelos testados, área abaixo da curva, erro padrão (SE) e significância assintótica (SA) obtidos através do software SPSS, levando em conta as variáveis RCQ e GC.</i>	45

MODELOS DE CONJUNTO *FUZZY* E DE REGRESSÃO LOGÍSTICA PARA DADOS DE OBESIDADE

Autor: DIEGO AUGUSTO QUEIJO

Orientadora: Prof. Dr. LICIANA VAZ DE ARRUDA SILVEIRA

RESUMO

Obesidade é comumente definida como um excesso de gordura corporal, porém diante da dificuldade em mensurar tal gordura diretamente, esta tem sido definida como um excesso de peso mais do que um excesso de gordura corporal, que tem como desdobramento a ocorrência de doenças associadas e/ou prejuízos à saúde do indivíduo (4). Atualmente, o excesso de massa corpórea é verificado por meio de um Índice de Massa Corpórea (IMC), que considera o quociente entre o peso corporal (kg) e a estatura elevada ao quadrado (m^2). Indivíduos com sobrepeso apresentam IMC de 25 até 29,9 kg/m^2 , e com obesidade apresentam IMC de 30 kg/m^2 ou mais, de acordo com a OMS (1).

A identificação das causas da obesidade não é trivial e objetiva. Especialistas reconhecem que a obesidade é uma doença crônica, de difícil tratamento, denominada multifatorial, envolvendo em sua gênese diversos aspectos, entre eles: o consumo alimentar, aspectos ambientais, genéticos, psicossociais, entre outros.

O objetivo deste trabalho foi desenvolver um modelo baseado em teoria dos conjuntos Fuzzy para a classificação de obesidade levando em consideração as suas causas, e compará-lo com um modelo de regressão logística através da curva ROC.

Para estudar as causas da obesidade na população de moradores da região do Distrito Sul de Campinas, foram coletados dados de uma amostra aleatória de 651 indivíduos, por meio de entrevista domiciliar. No primeiro estágio amostral, a partir do cadastro de domicílios residenciais dos agentes comunitários de saúde, foram aleatoriamente sorteados 920 domicílios (15% a mais do inicialmente previsto para cobrir perdas). Foram coletados dados de identificação geral, como: nome, idade, sexo, anos de escolaridade, tipo de ocupação e dados de consumo alimentar. O diagnóstico foi observado através do IMC.

Num estudo preliminar, no modelo fuzzy foram consideradas como variáveis de entrada, a idade e o grau de escolaridade dos indivíduos e, como saída o IMC, classificando-os como não-obesos ou obesos, levando em consideração dados da OMS. Observando os resultados pudemos constatar que a obesidade atinge em maior número os idosos e, através desses dados alguns "pesos" foram dados as variáveis para que se obtivesse um resultado melhor. Pelo modelo de regressão logística pudemos concluir que indivíduos com 40 anos ou mais têm 2 vezes mais chance de obesidade.

As variáveis razão cintura quadril e porcentagem de gordura corporal também foram utilizadas como entrada do sistema de modelagem.

Por fim, os dois modelos foram comparados com a curva ROC, separadamente para cada par de variáveis, avaliando qual deles apresenta um poder de tomada de decisão mais confiável.

Através deste estudo, foi possível dizer que o modelo de regressão tem um poder de decisão mais confiável do que o modelo fuzzy para as variáveis idade e ano de escolaridade. Por outro lado, o modelo fuzzy aparenta ser superior ao de regressão logística na tomada de decisão para a obesidade quando as variáveis razão cintura quadril e porcentagem de gordura corporal estão sendo estudadas.

Palavras-chave: Composição corporal; Epidemiologia; Índice de massa corporal; Modelos de conjuntos fuzzy; Modelo de regressão logística; Nutrição.

1 INTRODUÇÃO

A obesidade tem sido considerada uma epidemia global pela Organização Mundial da Saúde (OMS) devido ao aumento alarmante de sua prevalência em diversas partes do mundo (1). Este aumento está distribuído em quase todas as raças e sexo, e atinge tanto crianças como adultos (2), (3) .

Obesidade é comumente definida como um excesso de gordura corporal, porém diante da dificuldade em mensurar tal gordura diretamente, esta tem sido definida como um excesso de massa mais do que um excesso de gordura corporal, que tem como desdobramento a ocorrência de doenças associadas e/ou prejuízos à saúde do indivíduo (4). Atualmente, o excesso de massa é verificado por meio de um Índice de Massa Corpórea (IMC), de 25 até 29,9 kg/m^2 , e a obesidade de um IMC de 30 kg/m^2 ou mais, de acordo com a OMS (1). O IMC é calculado considerando o quociente entre a massa corporal (kg) e a estatura elevada ao quadrado (m^2).

A identificação das causas da obesidade não é trivial e objetiva. Especialistas reconhecem que a obesidade é uma doença crônica, de difícil tratamento, denominada multifatorial, envolvendo em sua gênese diversos aspectos, entre eles: o consumo alimentar, aspectos ambientais, genéticos, psicossociais, entre outros.

Para estudar as causas da obesidade na população de moradores da região do Distrito Sul de Campinas, a orientanda de doutorado do Prof. Dr. Ricardo Carlos Cordeiro, Lia Thieme Oikawa Zangirolani, coletou dados, por meio de entrevista domiciliar, de uma amostra aleatória de 651 indivíduos. No primeiro estágio amostral, a partir do cadastro de domicílios residenciais dos agentes comunitários de saúde, foram aleatoriamente sorteados 920 domicílios (15% a mais do inicialmente previsto para cobrir perdas). Foram coletados 2 dados de identificação

geral, como: nome, idade, sexo, anos de escolaridade, tipo de ocupação, dados de consumo alimentar e de composição corporal. O diagnóstico foi observado através do Índice de Massa Corporal (IMC), sendo este dividido em 2 categorias (uma: baixo peso e estrófico, outra: sobrepeso, obesidade tipo I e obesidade tipo II).

Com base nos dados acima, pretende-se comparar uma técnica matemática, teoria dos conjuntos *fuzzy* e uma técnica estatística, modelo de regressão logística, afim de verificar qual das técnicas utilizadas é a mais adequada para uma possível previsão do problema epidemiológico em questão, ou seja, a presença, ou não, de obesidade.

2 MODELO DE REGRESSÃO LOGÍSTICA

2.1 Linhas Gerais

Em Epidemiologia, a idéia de regressão logística é freqüentemente encontrada em problemas que envolvem a análise da relação entre uma variável resposta e uma ou mais variáveis explicativas. Uma questão típica dos pesquisadores é (5): Qual a relação de uma ou mais variáveis de exposição x (ou estudo) com uma doença ou estado patológico adquirido (Y)?

Para ilustrar, consideremos a variável dicotômica obesidade, a qual é representada pelo resultado 0 quando o paciente não é obeso e 1 quando o paciente é obeso (Lembrando que existem n conjuntos de doenças que podem ser representadas de tal maneira).

Em seguida, suponha que o interesse seja em uma variável dicotômica de exposição, como por exemplo, o fato do indivíduo fumar ou não, sendo este classificado então como fumante ou não fumante.

Com isto, a pergunta do pesquisador seria: Qual a relação do ato de fumar com a obesidade?

Outras variáveis também devem ser levadas em conta na obtenção do modelo logístico, e que apesar de não serem de principal interesse, fazem parte fundamental da análise, pois também exercem certa influência no resultado de relação final obtido. Tais variáveis são denominadas “controle”, as quais normalmente são definidas como idade, sexo, ano de escolaridade, entre outras.

Generalizando, a variável de exposição juntamente com as variáveis de controle formam o conjunto das variáveis independentes $x_1, x_2, \dots, x_k$, onde k é o número de variáveis a serem consideradas e que utilizamos para descrever ou prever a variável dependente Y , ou seja, a relação de um determinado resultado dados alguns antecedentes.

Em qualquer problema de regressão logística (6) a quantidade chave é o valor médio da variável de saída, dado os valores das variáveis independentes. Essa quantidade é denominada *média condicional* e será expressa por $E(Y|x)$ onde Y denota a variável de saída ou resultado e x denota o valor da variável independente, como já comentado anteriormente. Esta quantidade $E(Y|x)$ é lida como o valor esperado de Y dado um valor de x . Como pode ser visto na literatura, (7) um modelo razoável para $E(Y|x)$ é definido por:

$$E(Y|x) = \beta_0 + \beta_1 x \quad (1)$$

em que β_0 e β_1 são parâmetros desconhecidos e que tal definição vem dos princípios básicos da regressão linear. A expressão (1) implica que $E(Y|x)$ pode adquirir qualquer valor dado x pertencente ao intervalo $]-\infty, \infty[$.

Para iniciar o desenvolvimento teórico da regressão logística, é necessário descrever primeiramente como a função logística é representada, começando pelo modelo mais simples, onde há somente uma variável preditora:

$$E(Y) = \frac{1}{1 + e^{-x}} \quad (2)$$

em que x nada mais é que uma soma linear simples em função de uma variável independente, ou seja, o próprio $E(Y|x)$.

Utilizando o intervalo aplicado na equação (1) em (2), percebe-se que

a função logística adquire um intervalo de 0 a 1, sendo assim destinada a descrever uma probabilidade que esteja sempre enquadrada em tal intervalo. Em termos Epidemiológicos (8), tal probabilidade dá o risco de um indivíduo adquirir uma doença ou condição de saúde.

A função que dá base para o modelo de regressão logística ainda pode ser escrita como:

$$E(Y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \quad (3)$$

Portanto, o modelo de regressão pode ser denotado pela notação de $P(Y = 1|x)$, que representa uma sentença de probabilidade condicional e que pode ser descrita, neste caso, por:

$$P(X) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} \quad (4)$$

Uma propriedade interessante da função logística, é que ela pode ser linearizada.

Denotando-se $E(Y)$ por π (pela definição de valor esperado), pois a resposta média é a probabilidade quando a variável resposta é binária, fazemos a seguinte transformação (6):

$$\pi'(x) = \ln \left(\frac{\pi(x)}{1 - \pi(x)} \right) \quad (5)$$

em que,

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}, \quad (6)$$

obtendo:

$$\pi'(x) = \beta_0 + \beta_1 x \quad (7)$$

Esta transformação é chamada de transformação *logit* da probabilidade π . A razão $\pi/(1 - \pi)$ na transformação logit é chamada de **Odds** (chance).

Quando tratamos de problemas que fazem alusão a realidade, é difícil efetuar uma comparação adequada de um resultado obtido dado uma única variável, tendo em vista que “nada existe inteiramente sozinho, tudo está relacionado com tudo” (The Buddha).

Por esta razão, quando tratamos de problemas Epidemiológicos, os quais possuem inúmeras variáveis preditoras, é essencial que quanto mais variáveis forem levadas em consideração, mais próximo da realidade será o resultado obtido, o que é muito importante para a realização de um estudo mais sério do problema em questão.

Para tal, é necessário que o modelo preliminar seja estendido. Esta extensão consiste em levar a variável preditora x à um conjunto de variáveis preditoras (ou independentes), que são ilustradas por $x_1, x_2, \dots, x_k$, como visto anteriormente, de maneira que:

$$x_1, x_2, \dots, x_k \implies Y$$

em que os $x's$ podem representar qualquer coleção de variáveis de exposição, variáveis controle ou até mesmo uma combinação das variáveis de interesse.

Sempre que for desejado relacionar um conjunto de $x's$ a uma variável dependente, como Y , estamos tratando de um problema.

Com isto, (5) podemos perceber que a regressão logística é uma aproximação de um modelo matemático que pode ser usado para descrever a relação de diversos $x's$ a uma variável dependente dicotômica, como Y .

Dado este novo panorama, retomemos a função logística (2).

Desta vez, a variável x será uma extensão da soma linear $E(Y|x)$, sendo reformulada para uma soma linear de β_0 mais β_1 vezes x_1 mais β_2 vezes x_2 , assim por diante, até β_k vezes x_k , onde os β_j são novamente parâmetros a serem determinados e que serão discutidos mais adiante.

Portanto, (9) o processo de regressão pode ser visto como um método para estimar $E(Y|x_1, x_2, \dots, x_k)$. Sendo assim,

$$E(Y|x_1, x_2, \dots, x_k) = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k \quad (8)$$

e usando a mesma linha de raciocínio do início da discussão deste modelo, obtemos:

$$P(X) = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^k \beta_j x_j)}}, \quad (9)$$

ou ainda

$$P(X) = \frac{e^{\beta_0 + \sum_{j=1}^k \beta_j x_j}}{1 + e^{\beta_0 + \sum_{j=1}^k \beta_j x_j}} \quad (10)$$

em que o somatório adquire um valor de $j = 1, \dots, k$, β_0 é dito intercepto e o β_j são os parâmetros de regressão.

Dessa maneira, a transformação logit adquire um novo formato, cujo desenvolvimento resulta em:

$$\pi' = \beta_0 + \sum_{j=1}^k \beta_j x_j$$

Com a melhor formulação do modelo de regressão é possível, agora, obter-se na seção seguinte as estimativas dos parâmetros que envolvem o cálculo da expressão fornecedora do resultado quantitativo referente a análise estatística que o

modelo logístico desempenha.

2.2 Estimativa dos Parâmetros

Supondo que é desejado fazer a análise de um conjunto de dados através da equação (11); além da substituição dos valores de cada variável independente, devem ser obtidas as estimativas dos parâmetros que multiplicam tais variáveis.

Quando são utilizados os dados da pesquisa para estimar os parâmetros desconhecidos, diz-se que o modelo está sofrendo um *ajuste*.

O método que fornece a estimação desses parâmetros é denominado **maximum likelihood** ou ainda método dos estimadores de máxima verossimilhança.

Para aplicação deste método é necessário, inicialmente, construir a função de verossimilhança, a qual expressa a probabilidade dos dados observados como uma função dos parâmetros desconhecidos.

Considerando Y_i uma variável de Bernoulli, tem-se a seguinte distribuição de probabilidade:

$$Y_i = 1 \rightarrow P(Y_i = 1) = \pi_i$$

$$Y_i = 0 \rightarrow P(Y_i = 0) = 1 - \pi_i,$$

ou seja, a função de verossimilhança propriamente dita, em que $i = 1, \dots, n$.

Feita esta consideração, a distribuição de probabilidade (Bernoulli) é dada por:

$$f_i(Y_i) = \pi_i^{Y_i} (1 - \pi_i)^{1-Y_i} \quad (11)$$

sendo $Y_i = 0, 1$ e $i = 1, 2, \dots, n$.

Como as observações Y_i são independentes, a conjunta fica:

$$g(Y_i, \dots, Y_n) = \prod_{i=1}^n f_i(Y_i) = \prod_{i=1}^n \pi_i^{Y_i} (1 - \pi_i)^{1-Y_i} \quad (12)$$

em que o produtório está no intervalo de $i = 1, \dots, n$.

O princípio da máxima verossimilhança postula que deve-se usar como estimadores de β_i , o valor que maximiza a equação (12). No entanto, é mais fácil matematicamente trabalhar com o log desta expressão. Essa expressão, o **log likelihood** ou verossimilhança logarítmica, é definida como:

$$L(\beta_0, \beta_1) = \ln[g(Y_i, \dots, Y_n)] = \sum_{i=1}^n (Y_i \ln[\pi_i(x)] + (1 - Y_i) \ln[1 - \pi_i(x)]) \quad (13)$$

Sabendo que $E(Y_i) = \pi_i$ para uma variável binária, de (3) temos:

$$1 - \pi = (1 + e^{\beta_0 + \beta_1 x_i})^{-1} \quad (14)$$

Além disso, considerando (5) e (6), obteríamos a seguinte função de verossimilhança:

$$L(\beta_0, \beta_1) = \sum_{i=1}^n Y_i (\beta_0 + \beta_1 x_i) - \sum_{i=1}^n \ln(1 + e^{\beta_0 + \beta_1 x_i}) \quad (15)$$

Para achar o valor de (β_0, β_1) que maximiza $L(\beta_0, \beta_1)$ diferencia-se, então, a equação (15) em β_0 e β_1 , igualando o resultado obtido a zero. As equações encontradas são:

$$\sum_{i=1}^n [Y_i - \pi_i(x)] = 0 \quad (16)$$

$$\sum_{i=1}^n x_i [Y_i - \pi_i(x)] = 0 \quad (17)$$

Os valores encontrados de β_0 e β_1 pela solução das equações anteriores são denotados, respectivamente, por $\hat{\beta}_0$ e $\hat{\beta}_1$, os quais representam os valores dos parâmetros estimados. Não existe uma solução analítica para os valores β_0 e β_1 que maximizam a função de verossimilhança (15). Métodos numéricos são necessários para encontrar as estimativas de máxima verossimilhança, $\hat{\beta}_0$ e $\hat{\beta}_1$.

A princípio, o problema estatístico para achar os estimadores dos parâmetros esta resolvido: o resto é apenas uma questão de desenvolver um método numérico para achar os valores desejados. Tais métodos estão, atualmente, embutidos em inúmeros programas estatísticos.

Encontradas as estimativas $\hat{\beta}_0$ e $\hat{\beta}_1$, substituem-se esses valores em (3) para encontrar os valores ajustados.

A função de resposta ajustada é dada por:

$$\hat{\pi} = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x}} \quad (18)$$

Por conseguinte, usando a transformação logit, tem-se que:

$$\hat{\pi}' = \hat{\beta}_0 + \hat{\beta}_1 x \quad (19)$$

É importante notar, que todos os cálculos foram realizados para um problema com uma única variável preditora. Isto foi feito, pois facilita a manipulação das fórmulas e porque obtêm-se de imediato as equações que fazem partes da estimação dos parâmetros do modelo de regressão linear múltiplo.

Resumidamente, a função log-verossimilhança (15) estende-se diretamente para o modelo de regressão logística múltipla, dada por:

$$L(\beta') = \sum_{i=1}^n Y_i(\beta' \mathbf{x}_i) - \sum_{i=1}^n \ln(1 + e^{\beta' \mathbf{x}_i}) \quad (20)$$

em que $\beta' = (\beta_0, \beta_1, \beta_2, \dots, \beta_{k-1})$, $\mathbf{x}$ representa uma matriz das variáveis independentes com o mesmo índice de β' e, novamente, métodos numéricos devem ser utilizados para encontrar os valores de β' que maximizem (20). As estimativas de máxima verossimilhança serão denotadas por $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_{k-1}$ (ou ainda por $\hat{\beta}'$).

Por fim, a função resposta logística ajustada será:

$$\hat{\pi} = \frac{e^{\hat{\beta}'x}}{1 + e^{\hat{\beta}'x}} = (1 + e^{-\hat{\beta}'x})^{-1} \quad (21)$$

Como a maioria dos problemas em Epidemiologia e nas ciências em geral envolvem inúmeras variáveis, as considerações a seguir serão todas feitas para o problema de regressão logística com mais de uma variável independente.

2.3 Teste de significância para os parâmetros.

Uma aproximação para testar a significância dos coeficientes de uma variável está diretamente relacionado com o efeito causado sobre a variável resposta quando tem-se um modelo que leva em consideração a variável independente de interesse e um modelo que não inclui tal variável. O teste em questão diz respeito a comparação destes dois modelos, pelo resultado obtido da variável resposta. Assim

como na regressão linear, quanto maior for o valor encontrado nos cálculos, maior a importância da variável preditora, e quanto menor for o valor encontrado, menor será sua significância.

Na regressão logística, a comparação destes dois modelos é feita sobre a função log verossimilhança.

A equação que dá o valor desejado está baseada num valor de desvio, dado por:

$$D = -2\ln(\text{verossimilhança do modelo ajustado})$$

Agora, com o propósito de realizar a comparação entre os dois modelos, com a variável e sem a variável, obtêm-se:

$$G = (\text{Modelo sem a variável}) - (\text{Modelo com a variável}) = D_{sv} - D_{cv}$$

ou ainda, transportando o conceito para a área matemática, e depois de alguma álgebra, têm-se que:

$$G = 2\left\{\sum_{i=1}^n [y_i \ln(\hat{\pi}_i) + (1 - y_i) \ln(1 - \hat{\pi}_i)] - [n_1 \ln(n_1) + n_0 \ln(n_0) - n \ln(n)]\right\} \quad (22)$$

em que, $n_1 = \sum y_i$, $n_0 = \sum (1 - y_i)$ e n é o número total de casos em estudo.

Sobre a hipótese que β_1 é igual a zero, o valor de G segue uma distribuição *qui-quadrado* com grau de liberdade igual a 1.

O cálculo da log verossimilhança e do teste da razão de verossimilhança é padrão em qualquer *software* de regressão logística, tornando o trabalho com manipulações algébricas mais simples para o pesquisador.

Outro teste, também muito importante e que ajuda na avaliação da significância, é o teste de *Wald*. Esse teste (6), é obtido pela comparação da estimação

da máxima verossimilhança do parâmetro $\hat{\beta}_j$ a uma estimativa do seu erro padrão ($\hat{EP}$), ou seja,

$$W = \frac{\hat{\beta}_j}{\hat{EP}(\hat{\beta}_j)} \quad (23)$$

em que o erro padrão estimado é calculado em função da variância (também estimada).

Portanto,

$$\hat{EP}(\hat{\beta}_j) = [V\hat{AR}(\hat{\beta}_j)]^{1/2} \quad (24)$$

em que a variância vem da teoria de máxima estimação de verossimilhança (11). Essa teoria postula que os estimadores são obtidos a partir da matriz das derivadas parciais de segunda ordem da função log verossimilhança. A matriz obtida é chamada de *matriz de informação observada* (6) e o seu inverso fornece $\text{VAR}(\beta_j)$, porém como os valores de β_j são estimados, a sua variância também é.

3 TEORIA DOS CONJUNTOS FUZZY

3.1 Introdução

3.1.1 O que é a lógica *fuzzy*?

Nas ciências em geral, nos confrontamos freqüentemente com diversos problemas que envolvem incertezas e imprecisão, dos quais o processo de tomada de decisão acaba apoiando-se em conceitos vagos estranhos à lógica clássica e em parâmetros de natureza subjetiva (12).

A teoria fuzzy foi apresentada em 1964 por Lotfi A. Zadeh, professor no departamento de engenharia elétrica e ciências da computação da Universidade da Califórnia, em Berkeley, quando ele trabalhava com problemas de classificações de conjuntos que não possuíam fronteiras bem definidas (ou seja, a transição entre os conjuntos é suave e não abrupta). É importante, porém, ressaltar que no decurso da ciência outros pesquisadores demonstraram seu desconforto com relação a lógica binária, relatando sua fragilidade para lidar com situações mais realistas (13).

O intuito de Zadeh (14) foi flexibilizar a pertinência de elementos aos conjuntos criando a idéia de *grau de pertinência*, possibilitando que um elemento pertencesse parcialmente a um determinado conjunto.

Segundo (15), conforme o grau de incerteza de um problema aumenta, a capacidade de descrição de um modelo para resolução do mesmo decresce. Assim sendo, fez-se necessário o surgimento de uma teoria que fornecesse subsídios para a resolução de problemas com alto grau de incerteza, sem que informações importantes se perdessem durante a manipulação dos dados por incapacidade do modelo matemático em lidar com a incerteza inerente ao mesmo.

O termo *fuzzy* significa nebuloso, difuso, e se refere ao fato de, em muitos casos, não conhecermos completamente os sistemas que estamos analisando (16), e é por isso que a lógica *fuzzy* tem como ponto fundamental (17) a representação da lógica e da racionalidade humana na resolução de problemas complexos.

Por isso, em áreas onde a imprecisão e a subjetividade andam lado a lado, como a biologia, a medicina, a epidemiologia (18) e a saúde pública (19), que são as áreas relacionadas a este trabalho, a aplicação da teoria dos conjuntos *fuzzy* tem ajudado os pesquisadores a desenvolver modelos que se enquadrem de melhor maneira a realidade e possibilitem a obtenção de um resultado mais preciso do problema estudado.

3.1.2 Porque usar a lógica *fuzzy* na Epidemiologia?

Em epidemiologia a incerteza não se restringe apenas a variações aleatórias. Nessa área podemos agrupar as incertezas em duas classes: a variabilidade, originada da heterogeneidade da população ou da estocacidade; e a ignorância parcial, que resulta de erros sistemáticos de medida ou do desconhecimento de parte do processo considerado (subjetividade) (13). Portanto, variabilidade e ignorância devem ser tratados com métodos diferentes e mais apropriados (20). No caso da variabilidade a Teoria de Probabilidades (estatística) é, em geral, o método mais indicado, porém, ela não consegue na maioria das vezes abordar o problema da ignorância e da subjetividade. Esses últimos podem ser tratados, entre outros métodos, com a Teoria da Lógica Fuzzy.

A teoria da lógica *fuzzy* tem sua estrutura baseada no conceito de verdade parcial, ou seja, entre os valores de 0 a 1 da Teoria Clássica dos Conjuntos. Não é preciso uma vasta gama de exemplos para perceber que dificilmente o cotidiano de qualquer indivíduo é preenchido em sua maior parte pela incerteza, e apenas uma pequena parcela é composta de verdades absolutas. Com isto, fica inviável pensar, trazendo para o campo biológico/médico, num indivíduo que seja completamente saudável.

É neste sentido que a lógica fuzzy difere da lógica convencional, pois ela nos permite considerar afirmações com valores entre *Falso* e *Verdadeiro*, nos possibilitando inclusive trabalhar com variáveis linguísticas (13), e é a partir daí que se torna claro a força da lógica fuzzy para entender e modelar problemas relacionaddos a incerteza, imprecisões e verdades parciais, ou seja, ao mundo-real.

Consequentemente, com os fatos apresentados acima, percebe-se que a lógica fuzzy tem grande aplicabilidade em problemas relacionados a área da saúde e que ela é uma ferramenta de grande valor para a evolução das áreas de pesquisa, equipamentos, diagnóstico entre outros.

Abaixo, segue a estrutura, na qual a lógica fuzzy tem suas raízes firmadas.

3.2 Conjuntos Fuzzy

Como já foi dito, a teoria dos conjuntos fuzzy é uma extensão da teoria dos conjuntos clássicos e a função característica dos conjuntos clássicos, nada mais é que um caso particular da teoria em estudo.

A teoria dos conjuntos clássicos está baseada na função característica clássica (21), dada por

$$\mu_A(x) = \begin{cases} 1 & \text{se e somente se } x \text{ pertence a } A \\ 0 & \text{se e somente se } x \text{ não pertence a } A \end{cases} \quad (25)$$

onde U é o conjunto Universo, A é um subconjunto de U e x é um elemento de U , ou seja, a função característica é um mapeamento do conjunto universo no conjunto $\{0, 1\}$. Ainda, as operações básicas dos conjuntos clássicos são dadas

pela união, intersecção e o complemento, as quais podem ser expressas pela função característica:

$$\mu_{A \cup B} = \max(\mu_A(x), \mu_B(x))$$

$$\mu_{A \cap B} = \min(\mu_A(x), \mu_B(x))$$

$$\mu_{\bar{A}}(x) = 1 - \mu_A(x)$$

em que A e B são subconjuntos de U, *max* representa o operador disjunção (OU) e *min* representa o operador conjunção (E).

Para obtermos os conjuntos fuzzy e suas operações basta generalizarmos a função característica da lógica clássica para o intervalo $[0, 1]$, ou seja, $\mu_A(x) : U \longrightarrow [0, 1]$, o que implica em considerarmos um contínuo de valores de pertinência e não apenas *pertence* e *não-pertence*. O elemento x pertencerá ao subconjunto A com um grau de pertinência que é um valor no intervalo $[0, 1]$ (13).

O primeiro conceito está associado a definição de conjuntos fuzzy. Um subconjunto A de um conjunto U é considerado subconjunto fuzzy de U se for descrito como um conjunto de pares ordenados segundo o que segue (22):

$$A = \{ (x, \mu_A(x)); x \in U; \mu_A(x) \in [0, 1] \}$$

onde $\mu_A(x)$ é uma função de pertinência que determina com que grau x está em A:

$$\mu_A(x) = 1 \rightarrow x \text{ pertence totalmente ao conjunto A;}$$

$$0 < \mu_A(x) < 1 \rightarrow x \text{ pertence parcialmente ao conjunto A;}$$

$$\mu_A(x) = 0 \rightarrow x \text{ não pertence ao conjunto A.}$$

Assim, as principais operações e relações entre Conjuntos *Fuzzy* são definidas como extensão das operações e relações tradicionais, como pode ser visto

na Tabela 1, onde A e B denotam Conjuntos *Fuzzy* sobre um conjunto base X e $\mu_A(x)$ e $\mu_B(x)$ representam os graus de pertinência de x nos Conjuntos *Fuzzy* A e B respectivamente.

N°	Operação	Representação	Natureza
1	Complemento	$\neg A(x) = 1 - A(x)$	Operação
2	Diferença	$(A \neq B)$ se $A(x) \neq B(x)$ para pelo menos um elemento de $x \in X$	Relação
3	Igualdade	$(A = B)$ se $A(x) = B(x) \forall x \in X$	Relação
4	Inclusão	$(A \subseteq B)$ se $A(x) \leq B(x) \forall x \in X$	Relação
5	Intersecção	$A \cap B = A(x) \cap B(x) = \min[A(x), B(x)]$	Operação
6	União	$A \cup B = A(x) \cup B(x) = \max[A(x), B(x)]$	Operação

Em última análise, um conjunto fuzzy é caracterizado por uma função de pertinência, e o grau de pertinência pode ser considerado como uma medida que expressa a possibilidade de que um dado elemento seja membro de um conjunto fuzzy.

Quanto à representação de um conjunto fuzzy temos que, se ele é discreto podemos simplesmente enumerar os seus elementos juntamente com seus graus de pertinência, na forma:

$$A = \sum_i \mu_A(x_i)/x_i$$

onde a somatória se refere a operação união (disjunção) e a notação $\mu_A(x_i)/x_i$ se refere ao elemento x_i que pertence ao conjunto fuzzy A com grau $\mu_A(x_i)$. Em geral, por simplicidade, somente é listado no conjunto A aqueles elementos cujo grau de pertinência é diferente de zero.

Como foi visto, um conjunto fuzzy é caracterizado por uma função de pertinência que assume valores dentro do intervalo $[0, 1]$, portanto, a princípio, qualquer função que associe valores entre zero e um a elementos de um dado conjunto, pode ser tomada como função de pertinência. Entretanto, tanto o número quanto o formato das funções de pertinência devem ser escolhidos de acordo com o conhecimento sobre o processo que se quer estudar. As funções mais utilizadas são (22):

- **Triangular**

As funções de pertinência triangulares são caracterizadas por uma terna (**a**, **b**, **c**), onde **a** e **c** determinam o intervalo dentro do qual a função de pertinência assume valores diferentes de zero, e **b** é o ponto onde a função de pertinência é máxima.

$$\mu_A(x) = \begin{cases} 0 & \text{se } x \leq a; \\ \frac{x-a}{b-a} & \text{se } a < x \leq b; \\ \frac{c-x}{c-b} & \text{se } b < x \leq c; \\ 0 & \text{se } x > c. \end{cases} \quad (26)$$

- **Trapezoidal**

As funções de pertinência trapezoidais são caracterizadas por um conjunto de quatro valores de **a**, **b**, **c** e **d**, onde **a** e **d** determinam o intervalo dentro do qual a função de pertinência assume valores diferentes de zero, e **b** e **c** determinam o intervalo dentro do qual a função de pertinência é máxima e igual a 1.

$$\mu_A(x) = \begin{cases} 0 & \text{se } x \leq a; \\ \frac{x-a}{b-a} & \text{se } a < x < \leq b; \\ 1 & \text{se } b \leq x \leq c; \\ \frac{d-x}{d-c} & \text{se } c < x \leq d; \\ 0 & \text{se } x > d. \end{cases} \quad (27)$$

- **Gaussiana**

As funções de pertinência Gaussianas são caracterizadas pela sua média (μ) e seu desvio padrão (σ). Este tipo de função de pertinência tem um decaimento suave e tem valores diferentes de zero para todo domínio da variável estudada.

$$\mu_A(x) = \begin{cases} 0 & \text{se } x \text{ está fora do domínio;} \\ e^{-\frac{(x-\mu)^2}{2\sigma^2}} & \end{cases} \quad (28)$$

No entanto, o problema da escolha das funções de pertinência não foi ainda resolvido teoricamente e elas permanecem sendo escolhidas a depender da aplicação e do contexto do problema abordado (23).

Além dessas considerações, os conjuntos *fuzzy* possuem algumas características especiais. Entre tais características encontram-se: α -*nível*, *suporte*, *cardinalidade* e *altura*.

- **α -nível:** O α -*nível* de um conjunto *fuzzy* A corresponde ao conjunto tradicional que contém todos os elementos do conjunto universo U com grau de pertinência em A maior ou igual a α .

$$A_\alpha = \{x \in U \mid \mu_A(x) \geq \alpha\}$$

- **suporte:** O suporte $S(A)$ de um conjunto *fuzzy* A é o conjunto clássico de

todos os elementos $x \in U$ cuja função de pertinência tem valor diferente de zero.

$$S(A) = \{x \in U \mid \mu_A(x) > 0\}$$

- **cardinalidade:** A cardinalidade de um conjunto é o número total de elementos no conjunto. Uma vez que os elementos podem pertencer parcialmente a um conjunto fuzzy, uma generalização natural da noção clássica de cardinalidade consiste em pesar cada elemento pelo seu grau de pertinência. Sendo assim, a cardinalidade de um conjunto fuzzy é definida por (13) :

$$Card(A) = \sum_{x_i} \mu_A(x_i)$$

onde A é um conjunto fuzzy e os x_i 's são os elementos do conjunto Universo.

O conceito de cardinalidade torna-se importante quando estamos interessados no tamanho dos conjuntos, como acontece em sistemas de informação e análise de banco de dados fuzzy, e quando precisamos definir fatores de normalização para os conjuntos fuzzy, como acontece em alguns processos de defuzzificação (24).

- **altura:** A altura (h) de um conjunto *fuzzy* A corresponde ao seu maior grau de pertinência, entre todos os elementos do conjunto.

$$h(A) = \max_{x_i} \mu_A(x_i)$$

Os conjuntos fuzzy com altura igual a 1 são chamados normais e aqueles cuja altura é inferior a 1 são chamados subnormais.

3.3 Variáveis Linguísticas

Uma variável linguística *fuzzy* é uma variável cujo valor é expresso qualitativamente por um termo lingüístico (que fornece um conceito à variável) e quantitativamente por uma função de pertinência. De fato, uma variável lingüística é caracterizada por $\{n, T, X, m(n)\}$ onde n é o nome da variável (por exemplo, Índice de Massa Corpórea - IMC), T é o conjunto de termos lingüísticos de n (baixo peso, estrófico, sobrepeso, obesidade tipo I, obesidade tipo II, etc.), X é o domínio (Universo) de valores de n sobre o qual o significado do termo lingüístico é determinado (obesidade tipo I, por exemplo, entre 30 e 34,9 IMC) e m é uma função semântica que assinala para cada termo lingüístico $t \in T$ o seu significado, que é um conjunto *fuzzy* em X (ou seja, $m: T \rightarrow (X)$ onde (X) é o espaço dos conjuntos *fuzzy*).

É importante notar que a variável linguística é expressa em termos de uma variável básica, que denota a sua medida.

Os termos lingüísticos são usados para expressar conceitos e conhecimentos da transmissão de informações do ser humano, e em muitas campos de estudo eles são a forma mais importante de quantificar os dados obtidos. No universo médico o uso de termos lingüísticos permeia todas as áreas, inclusive exames laboratoriais. São usados frequentemente os termos *normalidade*, *levemente aumentado/diminuído*, *debilitado*, *bom estado*, etc (13).

As variáveis linguísticas são expressas dentro de um certo domínio de valores. Em geral é o pesquisador que estuda a determina área de interesse que define esse domínio e realiza sua partição fuzzy. Nesse contexto, o papel do especialista torna-se indispensável na modelagem fuzzy, particularmente em alguns modelos de epidemiologia e sistemas diagnóstico.

3.4 Regras Fuzzy

Conceitualmente, as regras fuzzy descrevem situações específicas que podem ser submetidas a análise de um painel de especialistas, e cuja inferência nos

conduz a algum resultado desejado.

A regra fuzzy é uma unidade capaz de capturar algum conhecimento específico, e um conjunto de regras é capaz de descrever um sistema em suas várias possibilidades. Cada regra fuzzy, da mesma forma que uma afirmação clássica, é composta por uma parte antecedente (a parte *Se*) e uma parte conseqüente (a parte *Então*), resultando em uma estrutura do tipo:

$$\{Se\} \{\text{antecedentes}\} \{Então\} \{\text{conseqüentes}\}.$$

Os antecedentes descrevem uma condição (premissas), enquanto a parte conseqüente descreve uma conclusão ou uma ação que pode ser esboçada quando as premissas se verificam. A diferença entre os antecedentes de uma regra fuzzy e uma regra clássica é que os primeiros descrevem uma condição elástica, ou seja, uma condição que pode ser parcialmente satisfeita, enquanto os últimos descrevem uma condição rígida.

Uma vez construído o conjunto de regras fuzzy necessitaremos de uma “máquina” de inferência para extrair dela a resposta final. Existem vários métodos de inferência possíveis e a escolha por um deles depende do sistema que está sendo analisado. No entanto, a inferência mais comum, e amplamente utilizada no controle de sistemas, é o Método de Mamdani, que será o utilizado no trabalho.

As regras são processadas em paralelo, ou seja, todas as regras são consideradas ao mesmo tempo, e ao final obtemos uma resposta que pode ser tanto um valor numérico clássico, quanto um conjunto fuzzy ou um funcional, a depender do tipo de conseqüente utilizado.

3.5 Relações Fuzzy

As relações fuzzy são, assim como os conjuntos fuzzy, uma generalização das relações clássicas. Uma relação clássica descreve a inter-relação entre dois ou mais objetos.

Uma relação clássica segue a função característica clássica. Sendo assim, uma relação de amizade entre duas pessoas, por exemplo, designada como “amigos” considera que nas relações humanas ou alguém **é** seu amigo ou **não** o **é**, o que é uma simplificação da realidade. Por outro lado, uma relação de “amizade” fuzzy entre duas pessoas pode considerar o grau de amizade entre elas. Sendo assim, dois ou mais indivíduos podem se relacionar com diferentes graus de amizade, desde 1 (são certamente amigos) até 0 (não são amigos). Uma relação binária sobre as variáveis x e y , cujos domínios são respectivamente X e Y , pode ser definida como um conjunto de pares ordenados em $X \times Y$ (13).

Uma relação fuzzy R sobre n objetos $x_1, x_2, \dots, x_n$, cujos domínios são $X_1, X_2, \dots, X_n$ respectivamente, é definida por uma função que mapeia o ponto $(x_1, x_2, \dots, x_n)$ em $X_1 \times X_2 \times \dots \times X_n$ para um número no intervalo $[0, 1]$, ou seja, $R : X_1 \times X_2 \times \dots \times X_n \rightarrow [0, 1]$. Este mapeamento é equivalente a uma função de pertinência de um conjunto fuzzy multidimensional (21).

3.6 Composição de Relações Fuzzy

Seja X e Y o universo de discurso para as variáveis x e y , respectivamente, e x_i e y_i os elementos de X e Y . Considere R a relação fuzzy que mapeia $X \times Y$ em $[0, 1]$ e a distribuição de possibilidade de X ; em termos das funções de pertinência de X e Y , pode-se escrever (25):

$$\mu_Y(y) = \oplus [\mu_X(x) \otimes \mu_R(x, y)]$$

onde os símbolos $\oplus$ e $\otimes$ denotam operadores de disjunção fuzzy e conjunção fuzzy, e que para a equação anterior representam os passos de multiplicação e soma na multiplicação de matrizes, respectivamente.

Na prática as duas regras composicionais mais utilizadas são a composição **max-min** e a composição **max-produto**. Utilizar-se-a, para a modelagem do problema epidemiológico em questão, somente a composição **max-min**, que é dada por:

$$\mu_Y(y) = \max[\min(\mu_X(x_i), \mu_R(x_i, y_i))]$$

Ainda, existem alguns conceitos que não serão aqui abordados, pois tratam de tópicos muito específicos e de alta complexidade de desenvolvimento matemático, mas que podem ser consultados em diversas referências da literatura, como (14), (13), (21), e que dizem respeito ao princípio de extensão, processos de decisão fuzzy e probabilidade de eventos fuzzy.

Contudo, para finalizar esta explanação teórica, é importante cometar brevemente sobre o modelo linguístico a ser utilizado (modelo tipo Mamdani) e sobre o método de *defuzificação*.

3.6.1 Modelo linguístico tipo Mamdani

Antes de iniciar esta seção, é importante avaliar o conceito de modelo linguístico. Os modelos linguísticos podem ser entendidos como sistemas especialistas que descrevem linguisticamente um determinado objeto complexo. O método de inferência determina a forma operacional do modelo linguístico. Ele é um mapeamento que define uma transformação do valor fuzzy de entrada em um valor de saída. Existem diversos métodos de inferência, porém o mais simples e que melhor se adapta a problemas epidemiológicos em geral é o modelo de Mamdani (13), (25).

Nesse modelo (25), a saída é construída pela superposição dos consequentes das regras individuais. Desta maneira, cada regra,

$$\text{SE } U \text{ é } A_i \text{ ENTÃO } V \text{ é } B_i$$

é expressa como um relação fuzzy R_i que é interpretada como o produto cartesiano dos conjuntos fuzzy A_i e B_i ,

$$R_i = A_i \times B_i,$$

isto é, R_i é um subconjunto de $X \times Y$, e cuja função de pertinência é

$$R_i(x, y) = A_i(x) \wedge B_i(y)$$

onde $\wedge$ é o operador de conjunção fuzzy, *min*.

Neste modelo, a agregação do conjunto de regras é realizada através do operador união sobre todas as relações individuais, o que caracteriza o caráter construtivo desse tipo de modelo. Assim, a função de pertinência $R(x,y)$ da relação fuzzy R é

$$R(x, y) = \vee_i R_i(x, y) = \vee_{i=1}^m (A_i(x) \wedge B_i(y))$$

onde $\vee$ é o operador de disjunção fuzzy, *max*.

Este modelo pode ser extendido para o caso de múltiplas variáveis de entrada e até mesmo de saída. Adiante será possível constatar a utilização do modelo lingüístico de múltiplas variáveis, com duas variáveis de entrada e apenas uma de saída.

Os modelos lingüísticos fuzzy fornecem uma saída para o sistema que é também um conjunto fuzzy. Todavia, muitas vezes estamos interessados em obter um número como resposta, ou seja, desejamos obter um resultado clássico a partir da saída fuzzy obtida. Este objetivo pode ser alcançado através dos métodos de defuzificação.

3.7 Métodos de defuzificação

A defuzificação, por simplicidade, é um procedimento que traduz o estado da variável de saída fuzzy para um valor numérico. Existem muitas técnicas de defuzificação e entre as mais utilizadas estão a *média dos máximos*, *centro de área* e *métodos das alturas*.

O método de centro de área (CA) é a técnica de defuzificação mais comumente usada, e que será aplicada no desenvolvimento do trabalho; ele (CA) pode ser compreendido como uma média ponderada, onde $\mu_A(x)$ funciona como o peso do valor x . Se x é discreto, então a defuzificação da conclusão fuzzy A é dada por (13):

$$y_0 = \frac{\sum_x \mu_A(x)x}{\sum_x \mu_A(x)}. \quad (29)$$

Da mesma forma, se x é contínuo, então,

$$y_0 = \frac{\int \mu_A(x)xdx}{\int \mu_A(x)dx}. \quad (30)$$

4 BREVE INTRODUÇÃO A CURVA ROC

Apesar de necessária, em parte, para a proposta do estudo deste trabalho, a teoria matemática/estatística por trás da curva ROC não é objeto de interesse e, portanto, não será apresentada. Contudo, um breve estudo sobre ela foi feito antes de realizar os passos computacionais para sua obtenção.

Na medicina, diversos métodos estatísticos estão voltados ao problema de classificar indivíduos em grupos, sendo os testes de diagnósticos o principal exemplo. Estes testes são descritos como métodos teoricamente capazes de indicar a presença ou a ausência de uma determinada doença, com uma determinada chance de erro. A quantificação destas chances é, essencialmente, o objetivo destes métodos estatísticos (26).

Para qualquer teste de diagnóstico é fixado um valor de corte para a variável em estudo, valor que determina a classificação dos indivíduos como anormais ou normais. Assim, qualquer teste é avaliado pela comparação relativa da fração de verdadeiros positivos (FVP), fração de falsos positivos (FFP), fração de verdadeiros negativos (FVN) e fração de falsos negativos (FFN) (27).

A probabilidade de um teste diagnóstico produzir um resultado positivo, dado que o indivíduo é portador da doença, é chamada *sensibilidade* (verdadeiro positivo) do teste; e a probabilidade do teste produzir um resultado negativo, dado que o indivíduo não porta a doença, é chamada *especificidade* (verdadeiro negativo) (28).

A curva **ROC** (*receiver operating characteristic*) tem por objetivo descrever a situação onde o teste fornece uma saída (resposta) que não é descrita somente como “positivo” ou “negativo”, mas por um resultado que pode ser expresso

por uma variável categorizada ordinalmente ou por uma variável contínua.

Por definição, uma curva ROC é a representação gráfica dos pares *sensibilidade* ou FVP (ordenadas) e $1 - \textit{especificidade}$ ou FFP (abcissas), resultantes da variação do valor de corte ao longo de um eixo de decisão, x , a representação gráfica assim resultante é designada por curva ROC no plano unitário.

Existe uma variedade de índices que foram propostos para especificar e/ou resumir as curvas ROC empíricas (27). Como mencionado por vários autores ((28), (29), (30)), a área abaixo da curva ROC é um dos índices mais utilizados para sumariar a “qualidade” da curva. De entre os métodos para cálculo de áreas abaixo de uma curva ROC, usar-se-à o método de máxima verossimilhança (31).

Como resultado do método proposto para o cálculo da área abaixo da curva ROC, os erros padrão associados a esta, podem ser obtidos como resultado da estimativa de máxima verossimilhança (31).

5 ANÁLISE DOS DADOS

Antes de iniciar a análise dos dados, um especialista foi consultado, afim de indicar quais são as variáveis com certa relevância para realizar o estudo desejado. Segundo a especialista, Profa. Dra. Lia Thieme Oikawa Zangirolani, o ano de escolaridade (anoesc), a idade, a porcentagem de gordura corporal (GC) e a razão cintura quadril (RCQ) poderiam ser variáveis potencialmente preditoras da obesidade. A partir deste indício, iniciou-se uma análise descritiva dessas variáveis.

Tabela 1: *Análise descritiva inicial das variáveis utilizadas usando o software SAS*

gênero	n° obs	var	media	SD	mediana	25p	75p	min	max
0	356	GC	35,56	5,41	35,90	32,10	39,60	19,50	48,70
		RCQ	0,81	0,09	0,80	0,76	0,87	0,43	1,08
1	287	GC	25,32	5,92	25,90	21,20	29,30	10,50	44,60
		RCQ	0,89	0,07	0,88	0,86	0,94	0,51	1,14
		idade	41,64	12,53	43,00	31,00	53,00	17,00	68,00
0 e 1		anoesc	8,58	4,23	9,00	5,00	11,00	0	24,00
		IMC	26,45	5,22	25,60	23,00	29,10	16,80	54,80

onde 0 e 1 representam o gênero feminino e masculino, respectivamente.

Por conseguinte, fez-se um estudo da porcentagem de indivíduos obesos e não-obesos para as categorias aferidas pelo especialista a cada uma das variáveis, apresentados na forma de gráfico de colunas (*Figuras 1 a 4*). A classificação da variável razão cintura quadril foi dada, pelo especialista, pelas seguintes categorias:

baixo, moderado e alto; para a variável porcentagem de gordura corporal: abaixo da média, média, acima da média e obeso; para a variável idade: jovem e adulto; e para a variável ano de escolaridade: primeiro grau, segundo grau e superior.

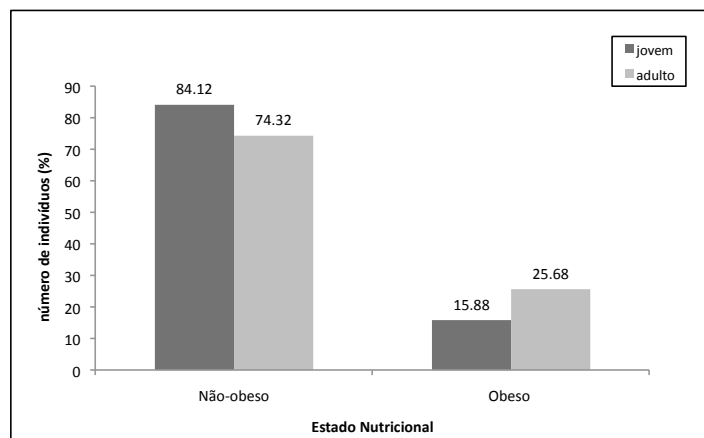


Figura 1: Gráfico de coluna para representar a porcentagem de indivíduos segundo estado nutricional e categoria de idade. Teste qui-quadrado com valor de $p = 0,0027$.

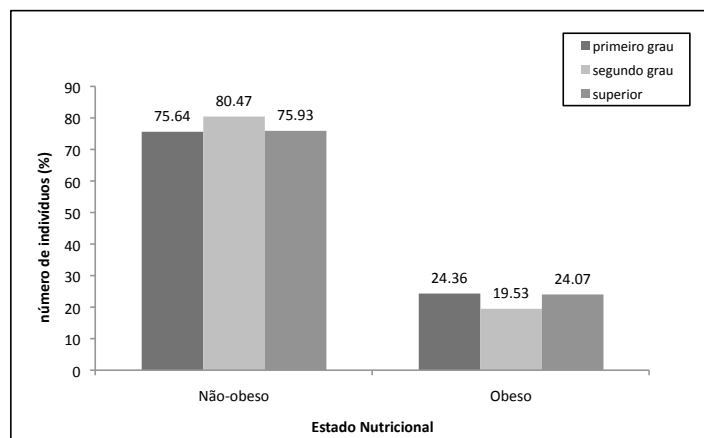


Figura 2: Gráfico de coluna para representar a porcentagem de indivíduos segundo estado nutricional e categoria de ano de escolaridade. Teste qui-quadrado com valor de $p = 0,3575$.

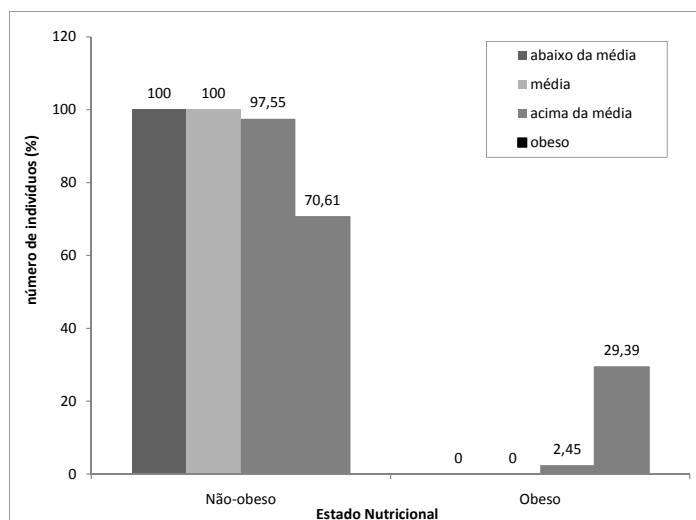


Figura 3: Gráfico de coluna para representar a porcentagem de indivíduos segundo estado nutricional e categoria de porcentagem de gordura corporal. Como a pressuposição do teste qui-quadrado não foi satisfeita, usou-se o teste exato de Fisher que forneceu um valor de $p < 0,0001$.

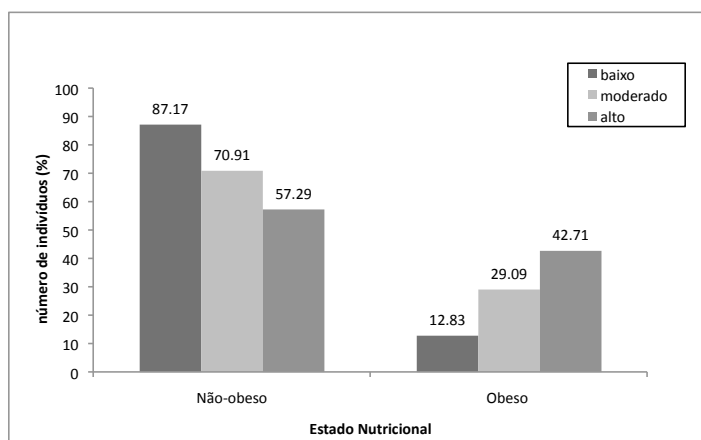


Figura 4: Gráfico de coluna para representar a porcentagem de indivíduos segundo estado nutricional e categoria de razão cintura quadril. Teste qui-quadrado com valor de $p < 0,0001$.

Dentre as variáveis estudadas, *RCQ*, *GC* e *idade* apresentaram associação significativa com o estado nutricional.

5.1 Regressão Logística

Primeiramente ajustou-se um modelo de regressão logística para (1) idade e ano de escolaridade e outro (2) para porcentagem de gordura corporal e razão cintura quadril, conforme a tabela 2 e 3, respectivamente.

Tabela 2: *Parâmetros, grau de liberdade (g), razão das chances (OR), estimativas, erro-padrão, estatística de Wald (Wald) e valor p para os parâmetros segundo o ajuste do modelo de regressão logística (1).*

parâmetro	g	OR	estimativas	erro-padrão	Wald	valor p
intercepto	1		-1,3106	0,1129	134,7904	< 0,001
anoesc 1 vs. 3	1	0,854	-0,0128	0,1568	0,0066	0,9351
anoesc 2 vs. 3	1	0,759	-0,1318	0,1321	0,9950	0,3185
adulto vs. jovem	1	1,812	0,2973	0,1049	8,0296	0,0046

Observando a tabela 2, ou seja, o modelo com as variáveis idade e ano de escolaridade, nota-se que somente o efeito da idade foi significativo para a presença de obesidade, sendo que os indivíduos com mais de 40 anos tem aproximadamente 2 vezes mais chance de serem obesos em relação aos mais jovens (valor $p = 0,0046$).

Tabela 3: *Parâmetros, grau de liberdade (g), razão das chances (OR), estimativas, erro-padrão, estatística de Wald (Wald) e valor p para os parâmetros segundo o ajuste do modelo de regressão logística (2).*

parâmetro	g	OR	estimativas	erro-padrão	Wald	valor p
intercepto	1		-7,0702	61,1853	0,0134	0,9080
GC abaixo vs obeso	1	*	-5,0428	117,4	0,0018	0,9657
GC acima vs obeso	1	0,0790	3,7992	61,1863	0,0039	0,9505
GC média vs obeso	1	*	-5,0906	153,8	0,0011	0,9736
RCQ alto vs baixo	1	3,3220	0,5764	0,1620	12,6616	0,0004
RCQ moderado vs baixo	1	1,9590	0,0479	0,1471	0,1062	0,7445

*não foi possível obter o OR.

Na tabela 3, somente o RCQ foi significativo, sendo que indivíduos com RCQ *alto* apresentaram aproximadamente 3,3 vezes mais chance de serem obesos em relação aos que têm RCQ *baixo*. Embora o especialista tivesse interesse em prever a obesidade com um modelo que contenha as variáveis GC e RCQ, para estes dados o ajuste com a presença da variável GC fica comprometida, pois entre os dados não se tem indivíduos com GC baixa e média que sejam obesos.

Para verificar a qualidade do ajuste baseou-se no critério da *deviance* (valor da *deviance*/grau de liberdade). Para o modelo (1), obteve-se o valor 1,0302 e para o modelo (2) 0,8922.

5.2 Modelo Fuzzy

Neste trabalho são propostos dois modelos lingüísticos fuzzy para avaliar o risco de obesidade baseando-se na fuzzificação das variáveis: (1) *idade* (I) e *ano de escolaridade* (ANOESC); (2) *razão cintura quadril* (RCQ) e *porcentagem de gordura corporal* (GC). O especialista, como citado anteriormente, considerou 2 conjuntos fuzzy para a variável idade: jovem (J) e adulto (AD); 3 para a variável ano de escolaridade (ANOESC): primeiro grau (PG), segundo grau (SG) e superior (S); 3 para a variável razão cintura quadril: baixo (B), moderado (MO) e alto (A); e 4 para a variável porcentagem de gordura corporal: abaixo da média (ABM), média (M), acima da média (ACM) e obeso (O).

Usou-se, para a construção de todos os conjuntos, a função de pertinência triangular.

Para o primeiro modelo (1), segundo o especialista, não há necessidade da diferenciação das variáveis idade e ano de escolaridade por gênero. Os conjuntos obtidos foram os seguintes:

Idade

$$I_{jovem} \begin{cases} \mu_{I_J}(x) = 0 & \text{se } x \leq 0; \\ \mu_{I_J}(x) = \frac{41-x}{41} & \text{se } 0 < x \leq 41; \\ \mu_{I_J}(x) = 0 & \text{se } x > 41. \end{cases} \quad (31)$$

$$I_{adulto} \begin{cases} \mu_{I_{AD}}(x) = 0 & \text{se } x \leq 39; \\ \mu_{I_{AD}}(x) = \frac{x-39}{29} & \text{se } 39 < x \leq 68; \\ \mu_{I_{AD}}(x) = \frac{79-x}{11} & \text{se } 68 < x \leq 79; \\ \mu_{I_{AD}}(x) = 0 & \text{se } x > 79. \end{cases} \quad (32)$$

Ano de escolaridade

$$ANOESC_{primeiro-grau} \begin{cases} \mu_{ANOESC_{PG}}(x) = 0 & \text{se } x \leq 0; \\ \mu_{ANOESC_{PG}}(x) = \frac{5-x}{5} & \text{se } 0 < x \leq 5; \\ \mu_{ANOESC_{PG}}(x) = 0 & \text{se } x > 5. \end{cases} \quad (33)$$

$$ANOESC_{segundo-grau} \begin{cases} \mu_{ANOESC_{SG}}(x) = 0 & \text{se } x \leq 4,5; \\ \mu_{ANOESC_{SG}}(x) = \frac{x-4,5}{4} & \text{se } 4,5 < x \leq 8,5; \\ \mu_{ANOESC_{SG}}(x) = \frac{12-x}{4,5} & \text{se } 8,5 < x \leq 12; \\ \mu_{ANOESC_{SG}}(x) = 0 & \text{se } x > 12. \end{cases} \quad (34)$$

$$ANOESC_{superior} \begin{cases} \mu_{ANOESC_S}(x) = 0 & \text{se } x \leq 11; \\ \mu_{ANOESC_S}(x) = \frac{x-11}{13} & \text{se } 11 < x \leq 24; \\ \mu_{ANOESC_S}(x) = \frac{33-x}{9} & \text{se } 24 < x \leq 33; \\ \mu_{ANOESC_S}(x) = 0 & \text{se } x > 33. \end{cases} \quad (35)$$

Para o segundo modelo (2), diferentes conjuntos fuzzy foram construídos de acordo com o gênero, já que segundo o especialista homens e mulheres possuem algumas especificidades quanto ao sistema fisiológico que influenciam diretamente nas variáveis RCQ e GC e, portanto, obtêm-se os seguintes conjuntos:

CONJUNTOS FUZZY PARA O GÊNERO FEMININO

Razão cintura quadril

$$RCQ_{baixo} \begin{cases} \mu_{RCQ_B}(x) = 0 & \text{se } x \leq 0; \\ \mu_{RCQ_B}(x) = \frac{0,76-x}{0,76} & \text{se } 0 < x \leq 0,76; \\ \mu_{RCQ_B}(x) = 0 & \text{se } x > 0,76. \end{cases} \quad (36)$$

$$RCQ_{moderado} \begin{cases} \mu_{RCQ_{MO}}(x) = 0 & \text{se } x \leq 0,74; \\ \mu_{RCQ_{MO}}(x) = \frac{x-0,74}{0,05} & \text{se } 0,74 < x \leq 0,79; \\ \mu_{RCQ_{MO}}(x) = \frac{0,84-x}{0,05} & \text{se } 0,79 < x \leq 0,84; \\ \mu_{RCQ_{MO}}(x) = 0 & \text{se } x > 0,84. \end{cases} \quad (37)$$

$$RCQ_{alto} \begin{cases} \mu_{RCQ_A}(x) = 0 & \text{se } x \leq 0,82; \\ \mu_{RCQ_A}(x) = \frac{x-0,82}{0,74} & \text{se } 0,82 < x \leq 1,08; \\ \mu_{RCQ_A}(x) = \frac{1,97-x}{0,89} & \text{se } 1,08 < x \leq 1,97; \\ \mu_{RCQ_A}(x) = 0 & \text{se } x > 1,97. \end{cases} \quad (38)$$

Porcentagem de gordura corporal

$$GC_{abaixo-da-media} \begin{cases} \mu_{GC_{ABM}}(x) = 0 & \text{se } x \leq 22,5; \\ \mu_{GC_{ABM}}(x) = \frac{22,5-x}{22,5} & \text{se } 0 < x \leq 22,5; \\ \mu_{GC_{ABM}}(x) = 0 & \text{se } x > 22,5. \end{cases} \quad (39)$$

$$GC_{media} \begin{cases} \mu_{GC_M}(x) = 0 & \text{se } x \leq 22; \\ \mu_{GC_M}(x) = \frac{x-22}{0,75} & \text{se } 22 < x \leq 22,75; \\ \mu_{GC_M}(x) = \frac{23,5-x}{0,75} & \text{se } 22,75 < x \leq 23,5; \\ \mu_{GC_M}(x) = 0 & \text{se } x > 23,5. \end{cases} \quad (40)$$

$$GC_{acima-da-media} \begin{cases} \mu_{GC_{ACM}}(x) = 0 & \text{se } x \leq 23; \\ \mu_{GC_{ACM}}(x) = \frac{x-23}{4,5} & \text{se } 23 < x \leq 27,5; \\ \mu_{GC_{ACM}}(x) = \frac{31,5-x}{4} & \text{se } 27,5 < x \leq 31,5; \\ \mu_{GC_{ACM}}(x) = 0 & \text{se } x > 31,5. \end{cases} \quad (41)$$

$$GC_{obeso} \begin{cases} \mu_{GC_O}(x) = 0 & \text{se } x \leq 31; \\ \mu_{GC_O}(x) = \frac{x-31}{18} & \text{se } 31 < x \leq 49; \\ \mu_{GC_O}(x) = \frac{62-x}{13} & \text{se } 49 < x \leq 62; \\ \mu_{GC_O}(x) = 0 & \text{se } x > 62. \end{cases} \quad (42)$$

CONJUNTOS FUZZY PARA O GÊNERO MASCULINO

Razão cintura quadril

$$RCQ_{baixo} \begin{cases} \mu_{RCQ_B}(x) = 0 & \text{se } x \leq 0; \\ \mu_{RCQ_B}(x) = \frac{0,94-x}{0,94} & \text{se } 0 < x \leq 0,94; \\ \mu_{RCQ_B}(x) = 0 & \text{se } x > 0,94. \end{cases} \quad (43)$$

$$RCQ_{moderado} \begin{cases} \mu_{RCQ_{MO}}(x) = 0 & \text{se } x \leq 0,093; \\ \mu_{RCQ_{MO}}(x) = \frac{x-0,97}{0,04} & \text{se } 0,93 < x \leq 0,97; \\ \mu_{RCQ_{MO}}(x) = \frac{1,02-x}{0,05} & \text{se } 0,97 < x \leq 1,02; \\ \mu_{RCQ_{MO}}(x) = 0 & \text{se } x > 1,02. \end{cases} \quad (44)$$

$$RCQ_{alto} \begin{cases} \mu_{RCQ_A}(x) = 0 & \text{se } x \leq 1,01; \\ \mu_{RCQ_A}(x) = \frac{x-1,01}{0,15} & \text{se } 1,01 < x \leq 1,14; \\ \mu_{RCQ_A}(x) = \frac{2,15-x}{1,01} & \text{se } 1,14 < x \leq 2,15; \\ \mu_{RCQ_A}(x) = 0 & \text{se } x > 2,15. \end{cases} \quad (45)$$

Porcentagem de gordura corporal

$$GC_{abaixo-da-media} \begin{cases} \mu_{GC_{ABM}}(x) = 0 & \text{se } x \leq 0; \\ \mu_{GC_{ABM}}(x) = \frac{14,5-x}{14,5} & \text{se } 0 < x \leq 14,5; \\ \mu_{GC_{ABM}}(x) = 0 & \text{se } x > 14,5. \end{cases} \quad (46)$$

$$GC_{media} \begin{cases} \mu_{GC_M}(x) = 0 & \text{se } x \leq 14; \\ \mu_{GC_M}(x) = \frac{x-14}{0,75} & \text{se } 14 < x \leq 14,75; \\ \mu_{GC_M}(x) = \frac{15,5-x}{0,75} & \text{se } 14,75 < x \leq 15,5; \\ \mu_{GC_M}(x) = 0 & \text{se } x > 15,5. \end{cases} \quad (47)$$

$$GC_{acima-da-media} \begin{cases} \mu_{GC_{ACM}}(x) = 0 & \text{se } x \leq 15; \\ \mu_{GC_{ACM}}(x) = \frac{x-23}{4,5} & \text{se } 15 < x \leq 19,75; \\ \mu_{GC_{ACM}}(x) = \frac{24,5-x}{4,75} & \text{se } 19,75 < x \leq 24,5; \\ \mu_{GC_{ACM}}(x) = 0 & \text{se } x > 24,5. \end{cases} \quad (48)$$

$$GC_{obeso} \begin{cases} \mu_{GC_O}(x) = 0 & \text{se } x \leq 24; \\ \mu_{GC_O}(x) = \frac{x-24}{20,5} & \text{se } 24 < x \leq 44,5; \\ \mu_{GC_O}(x) = \frac{62-x}{17,5} & \text{se } 44,5 < x \leq 62; \\ \mu_{GC_O}(x) = 0 & \text{se } x > 62. \end{cases} \quad (49)$$

O conjunto fuzzy de saída dos modelos é o risco de obesidade (IMC), tendo sido considerados 2 conjuntos fuzzy para esta variável: não-obeso (NO) e obeso (O).

$$IMC_{nao-obeso} \begin{cases} \mu_{IMC_{NO}}(x) = 0 & \text{se } x \leq 0; \\ \mu_{IMC_{NO}}(x) = \frac{33-x}{33} & \text{se } 0 < x \leq 33; \\ \mu_{IMC_{NO}}(x) = 0 & \text{se } x > 33. \end{cases} \quad (50)$$

$$IMC_{obeso} \begin{cases} \mu_{IMC_O}(x) = 0 & \text{se } x \leq 29; \\ \mu_{IMC_O}(x) = \frac{x-29}{26} & \text{se } 29 < x \leq 55; \\ \mu_{IMC_O}(x) = \frac{90-x}{35} & \text{se } 55 < x \leq 90; \\ \mu_{IMC_O}(x) = 0 & \text{se } x > 90. \end{cases} \quad (51)$$

O especialista considerou relevante as seguintes regras:

Modelo Lingüístico (2): $I \times ANOESC$

- 1) Se Idade é jovem E ANOESC é primeiro-grau ENTÃO IMC é não-obeso
- 2) Se Idade é jovem E ANOESC é segundo-grau ENTÃO IMC é não-obeso

- 3) Se Idade é jovem E ANOESC é superior ENTÃO IMC é não-obeso
- 4) Se Idade é jovem E ANOESC é primeiro-grau ENTÃO IMC é obeso
- 5) Se Idade é jovem E ANOESC é segundo-grau ENTÃO IMC é obeso
- 6) Se Idade é jovem E ANOESC é superior ENTÃO IMC é obeso
- 7) Se Idade é adulto E ANOESC é primeiro-grau ENTÃO IMC é não-obeso
- 8) Se Idade é adulto E ANOESC é segundo-grau ENTÃO IMC é não-obeso
- 9) Se Idade é adulto E ANOESC é superior ENTÃO IMC é não-obeso
- 10) Se Idade é adulto E ANOESC é primeiro-grau ENTÃO IMC é obeso
- 11) Se Idade é adulto E ANOESC é segundo-grau ENTÃO IMC é obeso
- 12) Se Idade é adulto E ANOESC é superior ENTÃO IMC é obeso

Modelo Lingüístico (1): $RCQ \times GC$

- 1) Se GC é abaixo da média E RCQ é baixo ENTÃO IMC é não-obeso
- 2) Se GC é média E RCQ é baixo ENTÃO IMC é não-obeso
- 3) Se GC é acima da média E RCQ é baixo ENTÃO IMC é não-obeso
- 4) Se GC é obeso E RCQ é baixo ENTÃO IMC é não-obeso
- 5) Se GC é obeso E RCQ é baixo ENTÃO IMC é obeso
- 6) Se GC é abaixo da média E RCQ é moderado ENTÃO IMC é não-obeso
- 7) Se GC é média E RCQ é moderado ENTÃO IMC é não-obeso
- 8) Se GC é acima da média E RCQ é moderado ENTÃO IMC é obeso

- 9) Se GC é obeso E RCQ é moderado ENTÃO IMC é obeso
- 10) Se GC é abaixo da média E RCQ é alto ENTÃO IMC é obeso
- 11) Se GC é média E RCQ é alto ENTÃO IMC é obeso
- 12) Se GC é acima da média E RCQ é alto ENTÃO IMC é obeso
- 13) Se GC é obeso E RCQ é alto ENTÃO IMC é obeso

É importante notar que apesar dos conjuntos fuzzy serem diferentes para cada gênero, quando trabalha-se com as variáveis RCQ e GC, o modelo lingüístico para ambos os sexos é praticamente idêntico, sendo acrescentada duas regras complementares aos homens e removida somente a décima das mulheres:

- 13) Se GC é acima da média E RCQ é moderado ENTÃO IMC é não-obeso
- 14) Se GC é obeso E RCQ é alto ENTÃO IMC é não-obeso

A partir dos conjuntos de regras e do modelo lingüístico fornecido pelo especialista, foi possível fazer a modelagem matemática do problema epidemiológico com o auxílio do software **MATLAB**, através do *Fuzzy Logical ToolBox*. Como resultado, obtivemos três superfícies:

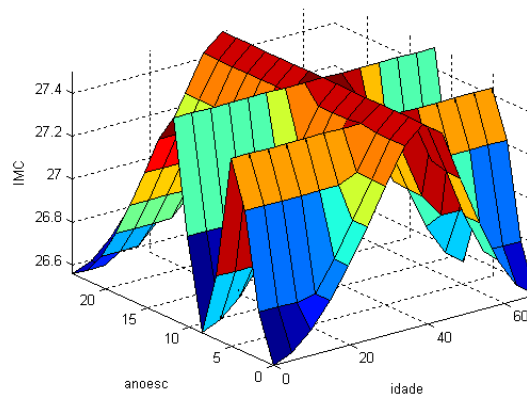
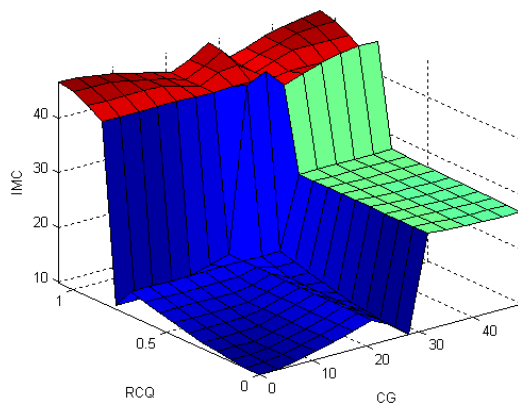
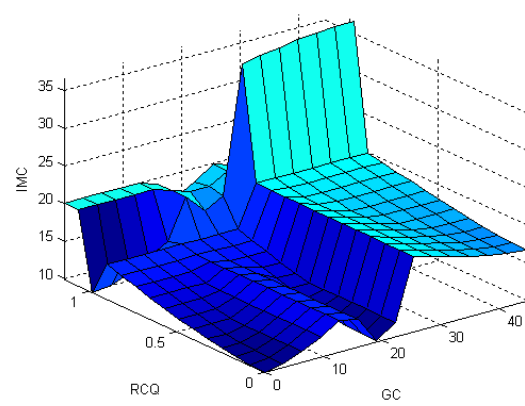


Figura 5: *Gráfico da análise do IMC levando em consideração as variáveis idade e ano de escolaridade.*



(a) Gráfico do IMC em relação ao $RCQ \times GC$ para o gênero feminino.



(b) Gráfico do IMC em relação ao $RCQ \times GC$ para o gênero masculino.

Figura 6: *Superfícies de indicação de obesidade pelo modelo lingüístico fuzzy.*

5.3 Curva ROC

As curvas ROC obtidas para a comparação dos dois modelos propostos (regressão logística e conjuntos fuzzy) considerando as variáveis (1) idade e ano de escolaridade e (2) razão cintura quadril e porcentagem de gordura corporal foram:

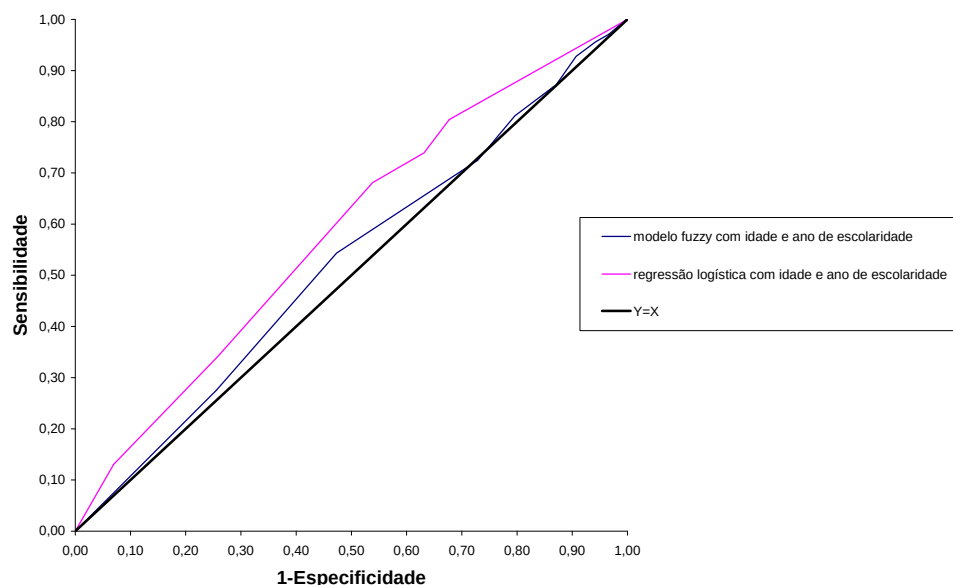


Figura 7: Gráfico da curva ROC comparando o modelo de regressão logística e o modelo fuzzy para as variáveis idade e ano de escolaridade.

Tabela 4: Modelos testados, área abaixo da curva, erro padrão (SE) e significância assintótica (SA) obtidos através do software SPSS, levando em conta as variáveis idade e ano de escolaridade.

Modelos testados	Área	SE	SA
modelo fuzzy	0,5230	0,0300	0,4038
modelo logístico	0,5850	0,0300	0,0022

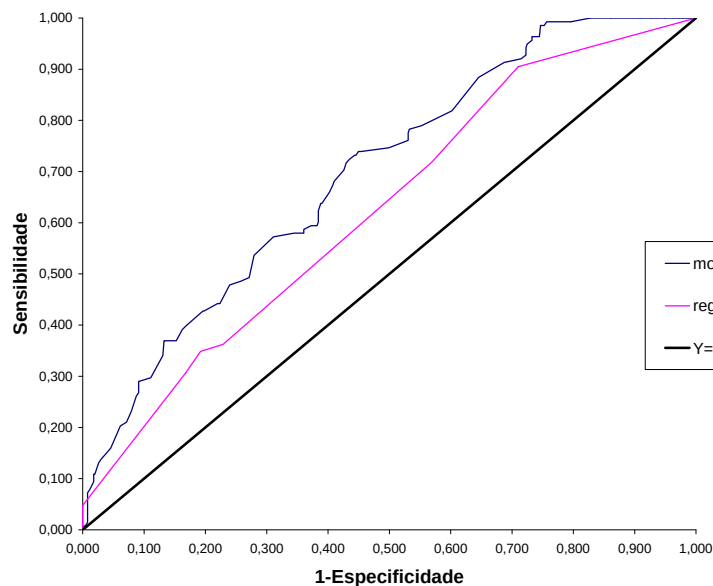


Figura 8: Gráfico da curva ROC comparando o modelo de regressão logística e o modelo fuzzy para as variáveis razão cintura quadril e porcentagem de gordura corporal.

Tabela 5: Modelos testados, área abaixo da curva, erro padrão (SE) e significância assintótica (SA) obtidos através do software SPSS, levando em conta as variáveis RCQ e GC.

Modelos testados	Área	SE	SA
modelo fuzzy	0,7440	0,0210	< 0,0001
modelo logístico	0,6970	0,0240	< 0,0001

Através das tabelas 4 e 5, as quais exibem o cálculo da área abaixo da área das curvas para o modelo logístico e fuzzy, é possível dizer que o modelo de regressão tem um poder de decisão mais confiável do que o modelo fuzzy para as variáveis idade e ano de escolaridade. Por outro lado, o modelo fuzzy aparenta ser superior ao de regressão logística na tomada de decisão para a obesidade quando

as variáveis razão cintura quadril e porcentagem de gordura corporal estão sendo estudadas.

6 DISCUSSÃO

Mais do que um melhor desempenho medido como área sob a curva ROC, o modelo fuzzy demonstrou habilidade para classificar casos em situações onde a distinção não é perceptível ao modelo logístico quando leva-se em consideração as variáveis RCQ e GC. O modelo logístico fica preso aos valores dos dados, ou seja, apresenta um comportamento que não corresponde com a expectativa do especialista, pois não consegue conciliar informações no mesmo modelo das variáveis RCQ e GC para prever a obesidade, e isso ocorre devido a alta associação entre elas, já o modelo fuzzy, permitiu a incorporação dos juízos subjetivos de especialistas no processo de formação da decisão a ser tomada.

No caso dos modelos para idade e ano de escolaridade, obteve-se um melhor desempenho do modelo logístico. Isso ocorre, pois as variáveis idade e ano de escolaridade não são diretamente relacionadas com a obesidade, o que já não é verdade para as variáveis RCQ (32) e GC, que apresentam alto grau de relação com a obesidade.

Conclui-se, no presente estudo, que o modelo fuzzy pareceu ser mais adequado como método para tomada de decisão a respeito do fenômeno abordado quando levadas em consideração as variáveis razão cintura quadril e porcentagem de gordura corporal e, que o modelo logístico se mostrou mais adequado para prever a obesidade quando as variáveis em estudo eram a idade e ano de escolaridade.

Estudos futuros podem ser feitos comparando o modelo de regressão

logística fuzzy com o modelo de regressão logística clássico.

REFERÊNCIAS BIBLIOGRÁFICAS

- [1] World Health Organization. *Physical status: the use and interpretation of anthropometry*. Report of a WHO Expert Committee. WHO Technical Report Series. Geneva; 1995.
- [2] Blumenkrantz, M. *Obesity: the world's metabolic disorder*[online]. Beverly Hills, 1997. Available from URL: <http://www.quantumhcp.com.obesity.htm>
- [3] Popkin, B. M.; Doak, C. M. *The obesity epidemic is a worldwide phenomenon*. Nutrition Reviews, Washington DC, v.56, n.4 (Pt1), p.106-114, 1998.
- [4] Ogden, C. L.; Yanovski S. Z.; Carrol, M. D.; Flegal, K. M. *The epidemiology of obesity*. Gastroenterology, 132: 2087-2102, 2007.
- [5] Kleinbaum, D. G. *Logistic Regression*. Springer-Verlag New York, Inc., página 4-5, 1994.
- [6] Hosmer, D. W.; Lemeshow, S. *Applied Logistic Regression*. New York, John Wiley and Sons Inc., páginas 4-6, 2000.
- [7] Morettin, P. A. *Estatística Básica*. Editora Saraiva, 2002.
- [8] Kleinbaum, D. G. *Logistic Regression*. Springer-Verlag New York, Inc., página 6 , 1994.
- [9] Hastie, T.; Tibshirani, R. *Generalized Additive Models*. Stanford Linear Accelerator Center and Department of Statistics, Stanford University, 1984.
- [10] Pawitan, Y. *In All Likelihood: Statistical Modelling and Inference Using Likelihood* Oxford Science Publications, 2001

- [11] Rao, C. R. *Linear Statistical Inference and Its Application*, Segunda Edição. Wiley, Inc., New York, 1973.
- [12] Souza, C. A. *Teoria de conjuntos fuzzy e regressão logística na tomada de decisão para realização de cintilografia das paratiróides*. Dissertação de mestrado apresentada ao Programa de Pós-Graduação em Saúde Pública da Universidade de São Paulo, 2007, 84p.
- [13] Ortega, N. R. S. *Aplicação de Teoria dos Conjuntos Fuzzy a Problemas da Biomedicina* Tese de Doutorado submetida ao Instituto de Física da Universidade de São Paulo para obtenção do Título de Doutor em Ciências, 2001
- [14] Zadeh, L. A. “Fuzzy Sets”. *Information and Control* **8** (3), 338-353, 1965.
- [15] Chiu, Chui-Yu; Park, Chan S. *Fuzzy cash flow analysis using present worth criterion*. The Engineering Economist, vol 39, 113-137, 1994.
- [16] Drakopoulos, J. A. “Probabilities, possibilities, and fuzzy sets”. *Fuzzy Sets and Systems* **75**, 1-15, 1995.
- [17] Von Altrock, C. *Fuzzy Logic and neurofuzzy applications in business and finance*. New Jersey: Prentice Hall PTR, 1996.
- [18] Massad, E.; Burattini, M. N.; Coutinho, F. A. B.; Struchiner, C. J. “A Fuzzy Model in HIV/AIDS Epidemiology: Predicting the Clinical Course Basing on HIV Viral Load” *Proceedings of Soft Computing SOCO'97*, Nime-France, 58-63, 1997.
- [19] Massad, E.; Burattini, M. N.; Ortega, N. R. S. “Fuzzy Logic and Measles Vaccination: Designing a Control Strategy”. *International Journal of Epidemiology* **28**, 550-557.
- [20] Ferson, S.; Ginzburg, L. R. “Different methods are need to propagate ignorance and variability”. *Reliability Engineering and System Safety* 54, 133-144, 1996.

- [21] Barros, L. C.; Bassanezi, R. C. *Tópicos de lógica fuzzy e biomatemática*. Coleção IMECC, 2006.
- [22] Amendola, M.; Souza, A. L.; Barros, L. C. *Manual do uso da teoria dos conjuntos Fuzzy no MATLAB 6.5*. FEAGRI & IMECC/ UNICAMP
- [23] Reznik, L. *Fuzzy Controllers*. Newnes, Reino Unido, 1997.
- [24] Yen, J.; Langari, R. *Fuzzy Logic: Intelligence, Control, and Information*. Prentice Hall, EUA, 1999.
- [25] Massad, E.; Ortega, N. R. S.; Barros, L. C.; Struchiner, C. J. *Fuzzy Logic in Action: Applications in Epidemiology and Beyond. Studies in Fuzziness and Soft Computing*, Springer, 2008.
- [26] Martinez, E. Z.; Louzada-Neto, F.; Pereira, B. B. *A curva ROC para testes diagnósticos Cadernos Saúde Coletiva*, Rio de Janeiro, 7-31, 2003.
- [27] Braga, A. C. S. *Curvas ROC: Aspectos Funcionais e Aplicações* Dissertação submetida à Universidade do Minho para obtenção do grau de doutor no Ramo de Engenharia de Produção e Sistemas, Área de Métodos Numéricos e Estatísticos, dezembro de 2000.
- [28] Metz, C. E. “Statistical Analysis of ROC Data in Evaluating Diagnostic Performance.” *Multiple Regression Analysis: Applications in the Health Sciences*, number 13, edited by Donald E. Herbert and Raymond H. Myers. 365–384. American Institute of Physics, 1986.
- [29] Hanley, J. A.; McNeil, B. J. “The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve,” *Radiology*, 143 :29–36 (1982).
- [30] Swets, J. A. *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*. New Jersey: LEA, 1996.

- [31] Hanley, J. A.; McNeil, B. J. “A Method of Comparing the Areas under Receiver Operating Characteristic Curves”, *Radiology*, 148 :839–843 (1983).
- [32] Godoy-Matos, A. F.; Oliveira, J. *Sobrepeso e Obesidade: Diagnóstico Sociedade Brasileira de Endocrinologia e Metabologia*, 2004