UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA
DEPARTAMENTO DE ESTATÍSTICA

Regressão Logística utilizando *b-splines*: uma maneira de lidar com relações não lineares

Autor: Nicole Machado Utpott
Orientador: Professor Dr. Álvaro Vigo

Porto Alegre, 03 de Julho de 2015

Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

Regressão Logística utilizando *b-splines*: uma maneira de lidar com relações não lineares

Autor: Nicole Machado Utpott

Trabalho de Conclusão de Curso
apresentado para obtenção
do grau de Bacharel em Estatística.

Banca Examinadora:
Professor Dr. Álvaro Vigo
Professora Dra. Vanessa Bielefeldt Leotti Torman

Porto Alegre, 03 de Julho de 2015

Dedico este trabalho aos meus queridos pais: Gilberto e Stela.

“O futuro pertence àqueles que acreditam na beleza de seus sonhos.”

Eleanor Roosevelt

Agradecimentos

Ao Prof. Álvaro Vigo, pelo constante empenho, pela transmissão de conhecimentos e pela sabedoria presente nos seus conselhos. Devo também agradecer pelo convívio desde o início do curso e pela forma como se manteve presente durante toda a minha graduação, desde minha primeira bolsa de iniciação científica no projeto ELSA, passando pelo apoio à oportunidade de intercâmbio na Holanda e, agora, como meu orientador.

Agradeço também aos excelentes profissionais que tive contato durante a graduação. Agradecimento especial aos professores Bruce Duncan, Cléber Bisognin, Liane Werner, Patrícia Ziegelmann, Stela Castro e Vanessa Torman por terem, de alguma forma, contribuído com o meu crescimento acadêmico e profissional.

À CAPES e ao CNPq, por proporcionarem meu intercâmbio na Holanda através do programa Ciência sem Fronteiras.

Aos amigos e amigas que fiz na graduação e aos que mantive ao longo desses anos. Agradeço também às gurias do ELSA, ao pessoal da Souza Cruz e, agora, aos colegas do Banco Sicredi.

Ao meu amor, Rodrigo, que além de todo o apoio e carinho, me proporciona diariamente momentos de muita felicidade. Agradeço também por me incentivar e acreditar em mim, mesmo quando nem eu acreditei.

Finalmente, agradeço à minha família, cujo suporte tornou possível essa longa jornada. Ao meu pai, Gilberto, à minha mãe, Stela, e ao meu irmão, Gustavo, por terem me incentivado, me apoiado em minhas decisões e por torcerem pelo meu sucesso.

Resumo

Em estudos clínicos onde o desfecho é uma variável dicotômica e o fator de exposição é de natureza quantitativa, uma grande dificuldade reportada pelos pesquisadores é estimar a razão de chances quando a relação entre o preditor e a resposta é não linear no *logito*. Práticas comuns como transformações, utilização de termos polinomiais e categorização das variáveis acarretam uma série de problemas, como: perda de poder, subjetividade da análise e dificuldades na interpretação. Por vezes, essas técnicas não possibilitam a estimação da razão de chances - medida de associação frequentemente utilizada em pesquisas, principalmente na epidemiologia. A abordagem de regressão logística utilizando *b-splines* é uma técnica pouco difundida e que pode ser útil para investigar relações não lineares entre preditores quantitativos com um desfecho binário, pois busca modelar as variáveis sem impor restrições, através de um modelo suavizado. Desta forma, o objetivo principal deste trabalho é revisar a literatura estatística a fim de identificar diferentes abordagens do uso de *splines* e explorar algoritmos implementados para empregar o método. Rotinas computacionais desenvolvidas por *Gregory et al.* (2008) foram utilizadas para ajustar diferentes modelos, variando o grau dos polinômios e a quantidade de pontos de corte. Os modelos ajustados mostraram-se fortemente capazes de identificar a relação funcional de um preditor quantitativo com um desfecho dicotômico, utilizando um banco de dados simulados. Os modelos de regressão logística utilizando *b-splines* podem ser avaliados através da estatística AIC e permitem estimar a razão de chances e o intervalo de confiança para valores pontuais da covariável, a partir de um valor de referência definido pelo usuário.

Palavras-chave: *b-splines*, regressão logística, relação não linear, macro SAS, razão de chances.

Sumário

1 INTRODUÇÃO 8
2 OBJETIVOS 10
3 REVISÃO DA LITERATURA 11
3.1 *Spline* 11
3.2 Motivação 11
3.3 Modelos de regressão por *splines* 14
3.3.1 *Spline* Linear 16
3.3.2 *B-splines* 20
3.3.3 *Spline* cúbico 22
3.3.4 Escolha dos nós 24
3.3.5 Aspectos computacionais 26
3.4 Regressão logística utilizando *splines* 26
4 EXEMPLO DE APLICAÇÃO 29
4.1 Geração do banco de dados 29
4.2 Regressão logística 34
4.3 Regressão logística utilizando *b-splines* 38
4.3.1 Macros 38
4.3.2 Regressão logística utilizando *b-spline* de grau um 41
4.3.3 Regressão logística utilizando *b-spline* de grau dois 44
4.3.4 Regressão logística utilizando *b-spline* de grau três 47
4.3.5 Regressão logística utilizando *b-spline* de grau três e oito nós 49
5 CONSIDERAÇÕES FINAIS 55
6 REFERÊNCIAS BIBLIOGRÁFICAS 58
ANEXO 60
APÊNDICE 65

1 INTRODUÇÃO

Na estatística, medidas de associação são utilizadas para avaliar as chances de um evento ocorrer em um certo grupo de indivíduos. Frequentemente a relação entre os preditores e o desfecho dicotômico é do tipo não linear, principalmente em pesquisas epidemiológicas - por exemplo, a relação entre doenças cardiovasculares e o consumo de álcool (Takahashi *et al.*, 2013). O comportamento dessa associação possui uma peculiaridade que pode ser encontrada em diversas situações. Como pode ser observado na Figura 1 abaixo, a relação entre mortalidade e consumo alcoólico não é exclusivamente crescente e linear.

Figura 1: Exemplo de relação não-linear

Para lidar com o problema de não linearidade, é comum utilizar a categorização dos preditores quantitativos - técnica não recomendada por muitos autores (Bennette *et al.*, 2012; Greenland, 1995; Turner *et al.*, 2010). Uma limitação dessa abordagem é a consequente perda de informação, no entanto, já existem potenciais alternativas na literatura.

A abordagem através de *splines* na regressão logística é um método alternativo à categorização e não exige pressupostos para a utilização. É uma técnica bastante versátil e capaz de captar a relação funcional da variável resposta (ou uma função da resposta, como no modelo logístico) com preditores quantitativos. Atualmente, existem muitas rotinas computacionais disponíveis para a aplicação do método.

Há 30 anos, Silverman (1985) mencionou que o método era pouco conhecido e adotado pelos pesquisadores. Fora do meio estatístico, a situação parece ter mudado. Como o assunto pode se tornar bastante complexo, este trabalho é sugerido como um texto para iniciantes no método e visa introduzir a técnica para ampliar os conhecimentos dos leitores. Os objetivos deste trabalho estão apresentados no Capítulo 2 e, no Capítulo 3, encontra-se uma breve revisão da literatura sobre *splines* e seu uso na regressão logística.

No Capítulo 4 é apresentado um conjunto de rotinas computacionais desenvolvidas em SAS para ajuste do modelo de regressão logística utilizando *b-splines.* As macros foram propostas por Gregory *et al.* (2012) em resposta ao problema de calcular e visualizar a razão de chances e seus intervalos de confiança para preditores contínuos. Ainda no Capítulo 4, é explicado passo-a-passo como utilizar as rotinas, como comparar diferentes modelos - variando o número de nós e os graus do polinômio, e como interpretar pontualmente a razão de chances e seu intervalo de confiança em relação a um valor de referência escolhido arbitrariamente.

Algumas considerações e recomendações são feitas no Capítulo 5, além de conclusões a respeito do método e sugestões para novos trabalhos. Nos Anexos e Apêndices estão o código para gerar os dados simulados e as tabelas com a descrição dos argumentos necessários no uso das macros.

2 OBJETIVOS

O objetivo principal do trabalho foi revisar a literatura estatística e de outras áreas afins para identificar diferentes abordagens do uso de *splines*. Além disso, fazer uma revisão introdutória a respeito de métodos de regressão logística utilizando *splines* e suas derivações.

Dentre os objetivos específicos, este trabalho buscou explorar e apresentar métodos computacionais (macros) em SAS para a aplicação da técnica de regressão logística utilizando *b-splines*, explicando passo-a-passo a utilização das rotinas, bem como a interpretação dos resultados. Além disso, buscou comparar diferentes modelos de *b-spline* e orientar a escolha do melhor ajuste, bem como estimar e interpretar a razão de chances, inclusive graficamente.

3 REVISÃO DA LITERATURA

Este capítulo visa introduzir os métodos de regressão logística por *splines*, porém sem a ambição de fazer uma revisão de literatura exaustiva. Pelo contrário, a ideia é apresentar de forma simples os principais conceitos, cobrindo aspectos desde a motivação para o uso de *splines* e diferentes abordagens, com referências úteis para o aprofundamento no tema.

3.1 *Spline*

O termo *spline* tem sua origem na Inglaterra no final do século XIX, em atividades das indústrias de construção civil, naval e aeronáutica (Wegman *et al.*, 1983). Um *spline* consiste em uma fita ou régua flexível usada para desenhar curvas que passam por pontos pré-determinados, com o intuito de auxiliar na etapa de delineamento de objetos como cascos de navio, peças de avião, etc. No ramo da matemática, a primeira tímida referência teórica ao assunto foi através de um artigo de 1946 escrito pelo matemático romeno Isaac Jacob Schoenberg.

O Glossário Inglês-Português de Estatística (2011), editado em conjunto pela Associação Brasileira de Estatística (ABE) e pela Sociedade Portuguesa de Estatística (SPE), determina *spline* como "*função definida segmentadamente por polinômios (no sentido nato*)". Em outras palavras, *spline* é um conjunto de funções polinomiais, conectadas em determinados pontos de corte (chamados nós), utilizados para ajustar uma curva a um conjunto de dados. No glossário, o termo *regressão por spline* é traduzido como regressão por partes.

3.2 Motivação

A abordagem de regressão através de *splines* é um método alternativo às práticas comuns de modelagem estatística, pois consiste em um modelo bastante flexível e, nos dias de hoje, devido aos avanços da tecnologia, de fácil

implementação computacional. Ainda assim, esse método não é amplamente conhecido e adotado (Silverman, 1985).

Os tópicos discutidos nesta seção servirão para embasar e promover a aplicação de regressão por *splines* e suas vantagens como uma escolha às técnicas radicadas fora do meio estatístico. Pode-se citar essa metodologia como uma alternativa aos testes de linearidade, modelos polinomiais, relações não lineares entre o preditor e a resposta (ou uma função da resposta), entre outros.

Um aspecto importante na modelagem, qualquer que seja o método utilizado, é a avaliação da suposição de linearidade entre o preditor quantitativo e a resposta (ou uma função da resposta, como no modelo logístico). Há diferentes métodos para avaliar o atendimento da suposição de linearidade, como por exemplo, por meio de teste de hipóteses baseados em quartis da distribuição do preditor ou baseado na transformação Box-Tidwell. Uma abordagem empírica, bastante usual, é por meio da inclusão de termos polinomiais, geralmente até o grau três.

Quando a suposição de linearidade não é satisfeita, isto é, se a relação entre a resposta (ou uma função da resposta) e um preditor quantitativo é não linear, pode ser difícil identificar a verdadeira forma funcional. Entre as alternativas ao modelo linear podem ser citadas transformações, modelos não lineares, modelos polinomiais (fracionários ou não) e modelos não paramétricos. Existem muitos casos em que a relação funcional tem um comportamento polinomial, como por exemplo, a associação entre doenças cardiovasculares e o consumo de álcool; ou o risco de derrame *versus* o consumo de café, cujo comportamento se assemelha a uma parábola (*J-shaped curve*) (Takahashi *et al.*, 2013). Para esses casos, a alternativa mais popular é adaptar o ajuste para uma função polinomial de grau superior. A regressão polinomial pode ser vista como uma generalização da regressão linear.

Uma abordagem alternativa de modelagem é a regressão através de *splines*, que é bastante flexível para identificar a forma funcional da relação do preditor com a resposta, podendo ser utilizado também para testar a suposição

de linearidade. Na literatura, as técnicas de modelagem não lineares vêm crescendo muito nas últimas décadas, e a abordagem por *splines* é apontada por muitos autores (Greenland, 1995; Keith *et al.*, 2014; Turner *et al.*, 2010) como uma excelente alternativa.

Na pesquisa clínica e epidemiológica é frequente a necessidade de investigar tendências ou associações do tipo “dose-resposta” com um desfecho dicotômico. O modelo de regressão logística tem sido muito utilizado nestas situações, porém a suposição de linearidade do preditor quantitativo com o *logito* da resposta é frequentemente negligenciada (Greenland, 1995; Gregory *et al.*, 2008; Takahashi *et al.*, 2013).

Modelos com termos polinomiais também são aplicados com frequência para esse tipo de problema, no entanto, a inclusão de termos polinomiais pode, por vezes, ocasionar colinearidade entre as variáveis do modelo.

Outras vezes, na presença de não linearidade, uma abordagem bastante utilizada é a categorização do preditor quantitativo, por meio de percentis (frequentemente tercis, quartis ou quintis) ou de pontos de corte definidos por critérios conceituais ou práticos. O uso deste novo preditor, categórico, acarreta na estimação de associações por meio de uma função do tipo escada, que assume que a associação (ou risco) permanece constante para todos os valores contidos em cada categoria. Isto pode conduzir a resultados incorretos, especialmente se as categorias não representarem grupos biologicamente homogêneos (Greenland, 1995; Bennette *et al.*, 2012).

Outro problema, intrínseco desta abordagem, é a dificuldade de comparação entre estudos, particularmente quando as categorias são escolhidas por um algoritmo mecânico como através de percentis, haja vista que os pontos de corte podem ser muito diferentes entre estudos, impedindo a comparação entre resultados de pesquisas. A suposição de que o risco não varia dentro das categorias pode reduzir o poder do estudo. Quando a distribuição da exposição é bastante assimétrica, possivelmente as categorias das caudas terão amplitudes relativamente maiores, tornando a análise, muitas vezes, implausível (Greenland, 1995; Bennette *et al.*, 2012).

A categorização de preditores quantitativos também pode estar associada com outro potencial problema, que é a necessidade de realizar múltiplos testes para comparações entre as categorias. Com o uso de quartis, por exemplo, frequentemente a categoria mais baixa é usada como referência. Dessa forma, todas as comparações serão em relação àquele grupo de indivíduos, e a pergunta da pesquisa será se o risco do desfecho aumenta conforme as categorias aumentam (ou vice-versa). Também é sabido que a chance de ocorrer um resultado falso-positivo cresce com o aumento do número de testes. Além disso, por vezes, o pesquisador encontra associação significativa para apenas uma das categorias e acaba optando por não incluir tal informação nas suas conclusões. Esses e outros problemas poderiam ser facilmente resolvidos utilizando apenas um teste para toda a amostra (Bennette *et al.*, 2012).

Modelos de regressão por *splines* podem ser particularmente úteis quando existem relações não lineares da resposta com preditores quantitativos. Existem várias abordagens de regressão por *splines*. Este trabalho aborda aspectos do uso de *b-splines* no modelo de regressão logística (desfecho dicotômico), como uma alternativa às categorizações de preditores quantitativos.

3.3 Modelos de regressão por *splines*

Wegman e Wright (1983) publicaram um artigo de revisão sobre o uso de *splines* na estatística, no qual apresentam *spline* como uma técnica de estimação não paramétrica, sendo tratado como uma abordagem de regressão não paramétrica. Nesse contexto, *spline* é definido como uma evolução da inferência clássica paramétrica que preenche o *gap* entre os métodos paramétricos e não paramétricos. Ainda, Greenland (1995) define que qualquer abordagem de regressão via *splines* pode ser considerada uma aproximação à regressão não paramétrica.

Mesmo não sendo considerados como uma forma funcional paramétrica, em muitos casos os *splines* podem ser escritos como combinações lineares de

funções polinomiais, e assim, de certa forma, podem ser vistos como paramétricos (Wegman *et al.*, 1983).

A técnica de regressão via *splines* é puramente um problema de interpolação. Ou seja, o objetivo é interpolar um segmento de reta que passe pelos pontos (x_i, y_i) no plano, onde $i = 1, 2, \dots, n$. Os pontos de corte denominados $\xi_j, j = 0,1, \dots, k$, são escolhidos através de um critério definido pelo pesquisador e devem contemplar todos os possíveis valores de x, cuja malha é definida por $\Delta = \{\xi_1 < \xi_2 < \cdots < \xi_k\}$ e, por razões computacionais, os pontos coincidem com valores assumidos por x (Wegman *et al.*, 1983), onde $\xi_1 > x_{(1)}$ e $\xi_k < x_{(n)}$.

Existem três principais abordagens para o ajuste de curvas utilizando *splines* e esses métodos podem ser diferenciados pelo modo como os resíduos são tratados. O método mais comum utiliza mínimos quadrados penalizados - o problema consiste em minimizar as distâncias dos erros para ajustar o modelo mais aderente aos dados. Uma segunda abordagem, mais aconselhável em certas circunstâncias, consiste em utilizar o método que estima intervalos com 100% de confiança para cada ponto do banco de dados. O terceiro procedimento, abordado neste trabalho, é chamado de regressão por *splines* (*regression splines*), e também faz uso da estimação pelo método dos mínimos quadrados, porém para cada “pedaço” (segmento) da função será ajustado um polinômio com parâmetros diferentes, forçando o encontro das funções adjacentes nos pontos de corte (Wegman *et al.*, 1983).

Para cada caso, em geral, existem infinitos polinômios que poderiam se ajustar aos pontos, sendo necessário apresentar alguns termos e definições importantes com o intuito de limitar as possibilidades de combinações polinomiais: grau da função *spline* (m), quantidade de pontos de corte ou nós (k), posição de cada um dos nós (ξ_j) e o número de coeficientes livres na função spline ($m + k + 1$).

O grau da função *spline* m e o número de nós k geralmente são definidos pelo pesquisador, já a posição dos nós ξ_j pode ser fixa ou livre. No último caso, as posições devem ser estimadas por meio de técnicas

apropriadas, as quais serão apresentadas nesse trabalho. Geralmente a escolha dos nós (posição) e o grau do polinômio são as duas maiores dificuldades reportadas por autores na aplicação da técnica (Wegman *et al.*, 1983). Desta maneira, este trabalho visa apresentar de forma simples e clara os possíveis obstáculos encontrados no uso do método e mostrar ao leitor formas para superá-los.

Silverman (1985) apresentou uma abordagem de regressão não paramétrica para ajuste de curvas, utilizando o método de suavização por *splines* (*smothing splines*), enfatizando que o método raramente é empregado em aplicações práticas. Salienta-se também que qualquer método de regressão tem dois objetivos principais: primeiramente, equipar o pesquisador com informações a respeito da relação entre as variáveis em estudo e, em segundo lugar, fornecer predições para valores que ainda serão observados. Para o primeiro propósito, um método de estimação não paramétrico é o ideal, pois permite que o modelo seja versátil e tenha um bom ajuste para os dados. Assim, o uso de *splines* como agentes suavizadores da curva de regressão encaixa-se perfeitamente, pois é uma técnica bastante flexível e de simples implementação computacional. Surpreendentemente, apesar dos importantes avanços computacionais, 30 anos depois, o método de regressão por *splines* ainda é pouco usado.

3.3.1 *Spline* Linear

A forma mais simples de *spline* é o *spline* linear, também chamado de função linear segmentada (*piecewise linear function*), que é definida como um conjunto de funções lineares com diferentes inclinações em cada intervalo definido pelos nós (Harrell, 2001).

No *spline* linear tem-se um segmento de reta para cada par de nós adjacentes. As conexões nos nós não são suavizadas e apresentam mudanças abruptas na direção dos segmentos de reta. No exemplo da próxima página o autor quis estudar a associação da idade com o salário. Nesse caso, foi

necessário apenas um nó para entender a relação entre essas variáveis, conforme Figura 2, adaptada do material elaborado por Kelly (2014).

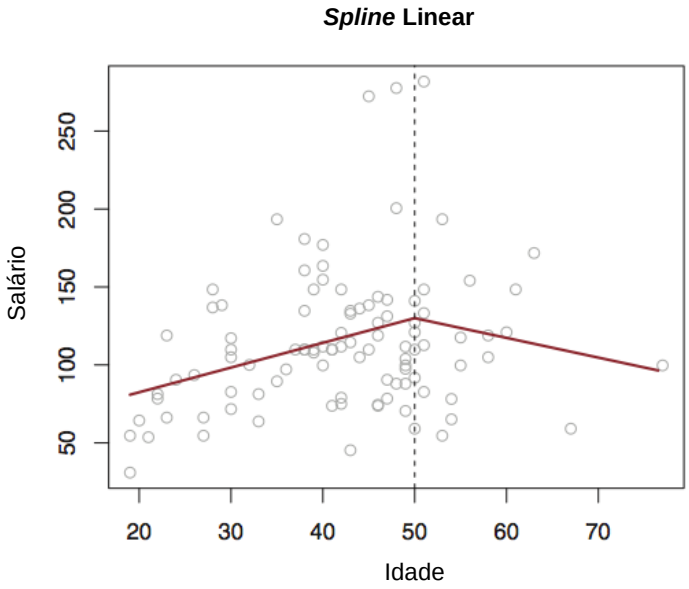

Figura 2: *Spline* linear com um ponto de corte.

Outros tipos de *splines* podem ser usados para obter uma representação suavizada dos dados. A escolha adequada do número de nós possibilita entender o comportamento dos dados e identificar, por exemplo, o tipo de relação funcional com a resposta. O método mais popular para fazer essa suavização é o *spline* cúbico, que possui polinômios de grau três ou inferior (Hastie *et al.*, 1990).

Conforme mencionado anteriormente, a origem do termo *spline* está relacionada a um pedaço de fita ou régua flexível utilizada para desenhar objetos comuns às indústrias de engenharia. Assim, é intuitivo pensar que a função *spline* é composta por um conjunto de retas que se ajustam aos dados. Por exemplo, considerando que o eixo x está dividido em quatro pequenas parcelas, definidas pelos nós internos ξ_1, ξ_2 e ξ_3, o modelo mais simples de *spline* linear possui a seguinte forma:

$$f(x) = \beta_0 + \beta_1 x + \beta_2(x-\xi_1)_+ + \beta_3(x-\xi_2)_+ + \beta_4(x-\xi_3)_+ \qquad (1)$$

em que

$$(u)_+ = \begin{cases} u, u > 0 \\ 0, u \leq 0 \end{cases}$$

onde

$$(u)_+ = \left(x - \xi_j\right)_+.$$

A quantidade de pontos de corte pode variar e depende principalmente da questão da pesquisa e também da quantidade de dados disponíveis para o ajuste da função (Harrell, 2001). Geralmente a abordagem de 4 a 7 nós é adequada.

Assim, à medida que o valor no eixo x cresce, novos termos são adicionados à função. Por exemplo, se $x < \xi_1$, então a equação que irá configurar esse espaço (entre o menor valor de x disponível na amostra e o primeiro nó ξ_1) será $f(x) = \beta_0 + \beta_1 x$, pois os coeficientes β_2, β_3 e β_4 serão multiplicados por zero, nesse caso. Quando x estiver entre o primeiro e o segundo nó, ou seja, $\xi_1 < x < \xi_2$ será adicionado mais um termo à equação: $f(x) = \beta_0 + \beta_1 x + \beta_2(x - \xi_1)$, e ela adotará outro formato. O mesmo ocorre quando $\xi_2 < x < \xi_3$, o termo ξ_3 é acrescentado e a equação será da forma $f(x) = \beta_0 + \beta_1 x + \beta_2(x - \xi_1) + \beta_3(x - \xi_2)$, restando apenas o β_4. E, por fim, para os valores finais de x, que se encontram última na parte, ou seja, $x > \xi_3$, o modelo contemplará todos os termos apresentados anteriormente em (1), $f(x) = \beta_0 + \beta_1 x + \beta_2(x - \xi_1) + \beta_3(x - \xi_2) + \beta_4(x - \xi_3)$.

A Figura 3 abaixo ilustra o modelo de *spline* linear com três nós, descrito na função $f(x)$.

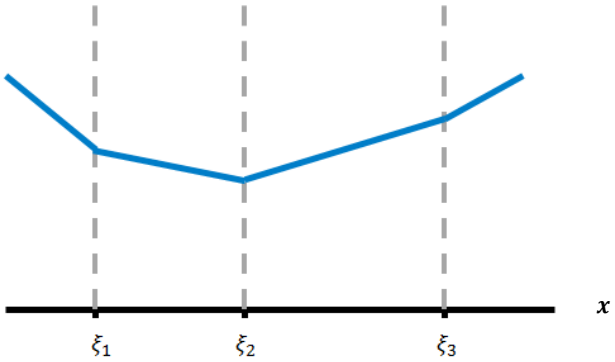

Figura 3: *Spline* linear com três nós.

A suposição de linearidade da relação entre a variável resposta e o preditor pode ser avaliada por meio do seguinte teste de hipóteses, considerando o modelo descrito anteriormente na equação (1):

$$H_0: \beta_2 = \beta_3 = \beta_4 = 0$$
$$H_1: \beta_i \neq \beta_j; para\ pelo\ menos\ um\ i \neq j; i = 2,3,4; j = 2,3,4.$$

Assim, se a hipótese nula for verdadeira, a contribuição dos termos *splines* correspondentes aos coeficientes de regressão β_2, β_3 e β_4 é nula (ou muito pequena), restando somente o intercepto β_0 e o coeficiente de regressão β_1 associado ao termo linear, evidenciando que a linearidade em x está satisfeita.

Naturalmente, o exemplo apresentado pode ser estendido para o caso geral com k nós. O modelo de regressão por *splines* é também atrativo para testar a linearidade de preditores quantitativos. O teste pode ser realizado por meio da estatística do teste da razão de verossimilhanças dos modelos:

$$f(x) = \beta_0 + \beta_1 x \quad (2)$$

e

$$f(x) = \beta_0 + \beta_1 x + \beta_2(x - \xi_1) + \beta_3(x - \xi_2) + \beta_4(x - \xi_3), \quad (3)$$

respectivamente denotadas por $L_0(\beta_0, \beta_1; x)$ *e* $L_1(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4; x)$. Assim, a estatística é da forma:

$$-2 \log \frac{L_0}{L_1} = (-2logL_0) - (-2logL_1) \sim \chi^2_{(3)}.$$

O número de graus de liberdade (três, no exemplo acima) é definido pela diferença entre o número de parâmetros β dos dois modelos.

Embora o *spline* linear seja simples e permita aproximar diversas relações, não tem formato suavizado nos nós e não reflete associações curvilíneas, muitas vezes encontradas em situações práticas, como por exemplo, o comportamento da relação entre o consumo de café (quantitativo) *versus* a ocorrência de infarto no miocárdio (binário) (Takahashi *et al.*, 2013). Para isso, utiliza-se *splines* com graus mais elevados (geralmente quadrático

ou cúbico), cujos polinômios formam curvas com diferentes formatos, dependendo do grau escolhido.

3.3.2 *B-splines*

B-splines é uma abreviação de *splines* básicos (em inglês, *basis* ou *basics splines*). O termo *basics* refere-se à aplicação de uma transformação feita na variável x antes de iniciar a etapa de ajuste do modelo. O *b-spline* é um polinômio definido em partes de grau m definido para uma variável x. Os pontos onde as partes se encontram são chamados nós, conforme definido anteriormente – a diferença aqui é que nos pontos de corte a função é suavizada, ou seja, não ocorre uma mudança abrupta da função entre um nó e outro, como acontece no *spline* linear.

A principal propriedade dos *b-splines* é o fato de que as funções são contínuas nos pontos de encontro e qualquer função de ordem m pode ser expressa como uma combinação de *b-splines*, o caso mais comum é o *b-spline* cúbico (de grau três).

Eilers e Marx (1996) descrevem *b-splines* como partes conectadas de polinômios com as seguintes propriedades, em que m é o grau do polinômio e k é o número de nós internos:

1) A função consiste de $k + 1$ pedaços de polinômio de grau m;
2) Os pedaços se unem nesses k nós, cujas derivadas nos pontos serão existentes;
3) O *b-spline* é positivo no domínio de $k + 2$ nós e nulo no restante;
4) Exceto nas caudas, cada pedaço de polinômio se sobrepõe com $2 \times m$ pedaços de polinômios vizinhos;
5) Os nós podem ser equidistantes ou não.

A curva *b-spline* ajustada para os dados (x_i, y_i) é dada pela seguinte combinação linear:

$$f(x) = \sum_{j=1}^{k+m-1} a_j \, B_j(x) \qquad (4)$$

em que $B_j(x)$ denota o valor do *j-ésimo b-spline* de grau m no ponto x para uma malha de k nós equidistantes. Para o caso de nós livres – não equidistantes – a equação sofre pequenas modificações e pode ser encontrada em De Boor (1978).

Existem certas divergências entre autores com respeito ao posicionamento dos pontos de corte. Alguns autores posicionam o primeiro nó no primeiro valor possível de x, mas a grande maioria define o primeiro nó onde ocorre o primeiro "corte" – nesse caso, os nós são chamados de internos. Por exemplo, na Figura 4, o primeiro nó (ξ_1) está localizado onde começa o primeiro pedaço de curva, no entanto, por vezes, o primeiro nó estará onde na figura é representado o segundo ponto de corte, ξ_2, conforme o exemplo ilustrado pela Figura 3. Essa informação será importante quando for necessário inserir a quantidade de nós nos algoritmos que calculam as transformações splines.

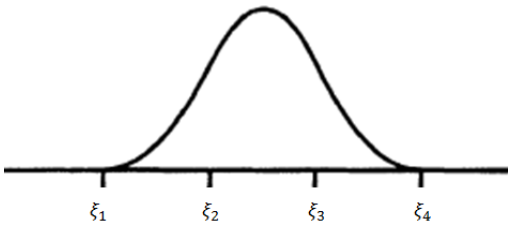

Figura 4: *B-spline* de grau dois.

Em suma, o *b-spline* é uma abordagem geral de regressão por *splines* e outros tipos podem ser vistos como casos especiais, como por exemplo: *splines* cúbicos, *p-splines*, *t-splines* ou *i-splines*.

O termo *p-spline* significa *spline* penalizado, referindo-se ao uso de *b-spline* com coeficientes estimados parcialmente pelos dados e parcialmente por uma função adicional de penalização que tem como objetivo impor a suavização a fim de evitar um superajuste. Existem outras derivações dos *b-splines*, como os *m-splines* (*spline* não negativo), *t-splines* (estima uma superfície), *i-splines* (*spline* monótono), entre outros, porém estão além dos objetivos deste trabalho (Bartels *et al.*, 1987).

Splines cúbicos têm apresentado boas propriedades e alta capacidade para captar comportamentos com formatos curvilíneos (Harrell, 2001). OS autores Hastie e Tibshirani (1990) sugerem que os olhos humanos passam a não compreender a associação devido à complexidade de modelos de *splines* com grau maior do que três. Dessa forma, é consenso entre muitos autores de que a abordagem através do *spline* cúbico é a mais apropriada quando se trata de regressão não paramétrica e verifica-se a necessidade de utilizar um modelo bastante flexível e de fácil ajuste computacional (Eilers *et al.*, 1996; Harrell, 2001).

3.3.3 *Spline* cúbico

Um *spline* cúbico é uma função construída por partes com polinômios de ordem três. É um caso especial dos *b-splines*, pois restringe o grau dos polinômios apenas para a ordem três. O *spline* cúbico pode ser dividido em duas categorias: restrito e não restrito. Será denominado restrito se as caudas (partes de polinômio antes do primeiro nó e após o último nó) forem modeladas através de funções lineares e recebe o título de não restrito caso isso não aconteça.

Supõe-se que o *spline* cúbico não restrito tenha k nós, assim, a função irá requerer estimativas para $k+3$ coeficientes de regressão, além do intercepto (β_0). Além do mais, recomenda-se que não sejam feitas extrapolações para além do primeiro e do último nó. A função de *spline* cúbico não restrito com três nós, ξ_1, ξ_2 e ξ_3, por exemplo, pode ser escrita da seguinte forma:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta^3 x^3 + \beta_4 (x - \xi_1)_+^3 + \beta_5 (x - \xi_2)_+^3 + \beta_6 (x - \xi_3)_+^3. \quad (5)$$

em que

$$(x - \xi_k)_+^3 = \begin{cases} (x - \xi_k)^3, x - \xi_k > 0 \\ 0, \qquad x - \xi_k \leq 0 \end{cases}.$$

Alguns estudos (Stone *et al.*, 1985) mostram que *splines* cúbicos não restritos não apresentam bom comportamento nas caudas, sendo estes, muito suscetíveis a mudanças – principalmente com pequenos tamanhos de amostra.

Sendo assim, por convenção, as funções que vem antes do primeiro e depois do último nó são forçadas a terem formato linear. Uma das vantagens dos *splines* cúbicos restritos (também chamados de *natural splines*) é que, além do intercepto (β_0), é necessário estimar apenas $k+1$ parâmetros, ao contrário dos $k+3$ parâmetros dos *splines* não restritos.

A função de *spline* cúbico restrito com k pontos de corte $\xi_1, \xi_2, \ldots, \xi_k$ pode ser escrita de forma geral como:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 (x - \xi_1)_+^3 + \beta_3 (x - \xi_2)_+^3 + \cdots + \beta_{k+1} (x - \xi_k)_+^3 \qquad (6)$$

em que

$$(x - \xi_k)_+^3 = \begin{cases} (x - \xi_k)^3, & x - \xi_k > 0 \\ 0, & x - \xi_k \leq 0 \end{cases}.$$

Os pedaços de polinômios de *spline* cúbico acabam se tornando uma única curva contínua, pois o encontro das funções nos pontos de corte é forçado através de uma restrição implícita no modelo, onde as derivadas das funções se igualam, com o intuito de atender a restrição de continuidade do modelo.

O gráfico da Figura 5 (Kelly, 2014) na próxima página foi gerado com o mesmo banco de dados que gerou a Figura 2. Essa figura exemplifica bem a relação de continuidade entre as variáveis: mostrando como a função de *spline* cúbico capta de forma mais assertiva a relação entre o salário e a idade ao invés de mostrar apenas dois segmentos de reta (em comparação com o gráfico de *spline* linear apresentado na Figura 2).

A vantagem em utilizar *splines* suavizados em comparação com os lineares é que essa abordagem tem capacidade para captar até pequenas ondulações no comportamento dos dados, dependendo da ordem do polinômio utilizado. A escolha do grau do polinômio deve ser feita com cautela e um dos objetivos deste trabalho é mostrar as diferenças entre os graus utilizados em um banco de dados simulados.

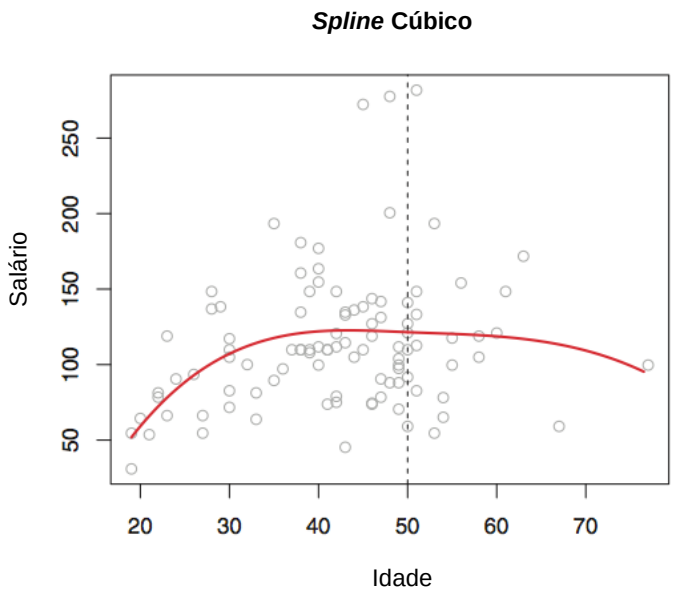

Figura 5: *Spline* cúbico com um ponto de corte.

Além disso, ao utilizar *b-splines*, sejam eles de grau um, dois, três ou mais, ao contrário de modelos polinomiais, é que, ao invés de utilizar polinômios com alto grau para ajuste, utilizam-se vários pequenos polinômios de grau baixo (em geral não ultrapassa três) para cada região de x. Dessa forma, o modelo proposto vai captar comportamentos diferentes para cada parte do conjunto de dados que forem semelhantes entre si.

3.3.4 Escolha dos nós

Tanto a escolha da quantidade dos nós internos quanto a escolha da posição dos mesmos são as principais dificuldades reportadas pelos autores para a especificação do modelo de regressão por *splines*. Em casos pontuais, quando a relação de interesse já foi estudada por outros pesquisadores, é aconselhável utilizar os nós propostos na literatura (Harrell, 2001), por exemplo, quando a variável depende é o IMC (índice de massa corporal) que é uma variável que já possui pré-classificações com embasamento biológico, que vai da magreza acentuada até a obesidade grave. Em geral, na epidemiologia, os nós tem um significado importante: podem representar pontos onde o comportamento da característica muda e a probabilidade de um desfecho

aumenta (Gregory *et al.*, 2008). No entanto, se os nós internos forem parâmetros livres, a função terá mais flexibilidade de ajuste, mas ao custo de maior instabilidade nas estimativas (Keith *et al.*, 2014; Harrell, 2001).

A escolha da posição dos nós ξ_k não é tão decisiva no resultado do modelo, o bom ajuste depende muito mais da quantidade de nós internos k. Popularmente costuma-se utilizar de três a sete nós, por isso, um bom método é utilizar os quantis (tercis, quartis, quintis, etc.). Para o modelo de regressão logística, no entanto, é importante verificar se existe um número de eventos (bem como não eventos) suficiente para cada uma das categorias definidas pelos pontos de corte. Harrell (2001) recomenda que o número de nós seja decidido levando em conta o tamanho da amostra disponível. Para uma amostra menor que 100, o uso de quatro nós internos geralmente produz um ajuste adequado e retorna um modelo balanceado em relação à flexibilidade e à perda de precisão. Já para amostras grandes, o uso de cinco nós é um ponto de partida razoável. A partir de sete nós a classificação passa a perder o significado, tornando a análise subjetiva.

O Critério de Informação de Akaike (AIC), fornecido por diversos pacotes computacionais, pode ser utilizado para fins de comparação de modelos, com o intuito de verificar o melhor ajuste com respeito ao número de nós, suas posições e o grau do polinômio (Harrell, 2001).

Em 1974, o pesquisador Wold participou de um estudo aprofundado que reuniu diversas recomendações “gerais” para a etapa de seleção dos nós, as quais estão listadas abaixo. Essas sugestões foram elaboradas tomando como base os *splines* cúbicos restritos, visto que é o caso mais disseminado entre a comunidade científica. O autor ainda menciona que, para graus maiores que três ($m > 3$), talvez sejam necessárias algumas adaptações nas recomendações a seguir:

1) Os pontos de corte devem coincidir com pontos já existentes no banco de dados;
2) Recomenda-se que, no mínimo, existam quatro ou cinco observações entre cada par de nós;

3) Não mais do que um extremo e um ponto de inflexão deverão ocorrer entre nós adjacentes (limitação associada ao polinômio de grau três);
4) Extremos devem estar centrados em intervalos e pontos de inflexão próximos aos pontos de corte.

3.3.5 Aspectos computacionais

Na literatura, a técnica de regressão via *splines* é extensivamente utilizada na área de ciências físicas e engenharias, mas surpreendentemente incomum na epidemiologia. Para modelos de regressão por *splines*, alguns softwares já dispõem de opções e pacotes para a execução da técnica (Greenland, 1995).

A regressão por *spline* pode ser executada por qualquer programa convencional que já possua a técnica de regressão, simplesmente adicionando ao modelo as transformações da variável de exposição correspondentes a cada parte do *spline*. Já para modelos logísticos via *splines*, a principal dificuldade é encontrar um *script* que calcule as medidas de risco, que frequentemente são do interesse do pesquisador, para poder avaliar os efeitos da associação de uma ou mais variáveis quantitativas em relação a um desfecho dicotômico.

3.4 Regressão logística utilizando *splines*

A utilização de *splines* pode se dar através de diversas técnicas estatísticas, no entanto, este trabalho visa dar enfoque ao uso de *splines* na regressão logística.

Para investigar relações não lineares entre preditores quantitativos e desfechos dicotômicos, a abordagem de regressão logística através de *splines* é atrativa devido à flexibilidade do modelo para identificar a forma da relação funcional entre um ou mais preditores quantitativos e o *logito* da probabilidade do evento (Silverman, 1985). No entanto, a escolha dos pontos de corte ainda é um tópico pouco explorado e que depende da questão do estudo e da população que está sendo investigada. Já existem alguns algoritmos que

auxiliam nessa decisão, entretanto, deve-se tomar o cuidado de não tornar a análise subjetiva, dividindo os possíveis valores de x em grupos que não tem características em comum ou não possuem nenhum significado biológico que motivem o agrupamento.

Existem diferentes abordagens com modelos não paramétricos para avaliar uma relação não linear entre um desfecho dicotômico e um ou mais preditores quantitativos, tais como modelos aditivos generalizados (GAM), LOESS (*locally weighted scatterplot smothing*), etc. Esses modelos, em geral, são usados para estabelecer predições da ocorrência ou não de um evento binário. No entanto, muitas vezes se deseja estimar associações para valores de um preditor quantitativo, com respeito a algum valor de referência, com a ocorrência de um evento dicotômico.

No contexto da previsão, programas usuais de análise estatística de dados, como SAS, R ou STATA são muito versáteis. Para estimar associações, no entanto, cálculos adicionais são necessários, os quais têm sido implementados em rotinas computacionais (macros) complementares. Como exemplo pode-se citar as rotinas desenvolvidas por Gregory *et al.* (2008) e Harrell (2001).

Gregory *et al.* (2008) apresentaram um conjunto de macros para ajustar uma curva de regressão logística não paramétrica através de *b-splines* com um número arbitrário de covariáveis, estimando a razão de chances para um valor de referência específico, bem como os respectivos intervalos de confiança. O modelo de regressão logística usando *b-splines* é definido pela equação (4), substituindo $f(x)$ pelo *logito* da probabilidade de ocorrência do evento $Y = 1|x$.

As macros foram desenvolvidas em resposta ao problema de calcular e visualizar a razão de chances para preditores contínuos na regressão logística, e um ou mais desses preditores podem ser substituídos por expansões *spline* devido à relação não linear. Considerando o preditor quantitativo x, o estimador da razão de chances proposto pelos autores, em relação ao valor de referência x_{ref}, é:

$$\widehat{RC}(x, x_{ref}) = \exp\left(\sum_{i=1}^{n} \widehat{\beta}_i \left[s_i(x) - s_i(x_{ref})\right]\right),$$

em que $\widehat{RC}$ é a estimativa da razão de chances, $\widehat{\beta}_i$ é o coeficiente do i-ésimo *b-spline* estimado pela regressão logística, $s_i(x)$ é o valor do i-ésimo *b-spline* em x e n é o número de graus de liberdade da expansão *spline*. A expansão *spline* é a transformação pela qual a variável passa, cuja quantidade será $k + m + 1$, onde k é a quantidade interna de nós e m o grau da função. Os limites do intervalo de 95% de confiança para a razão de chances são calculados por:

$$\exp\left(log\left(\widehat{RC}(x, x_{ref})\right) \pm 1{,}96 \times \hat{\sigma}^2_{log(\widehat{RC})}\right),$$

em que $\hat{\sigma}^2_{log(\widehat{RC})}$ é a variância do termo $log\left(\widehat{RC}(x, x_{ref})\right)$.

As rotinas foram desenvolvidas para a versão 8.2 ou mais recentes do programa SAS, sendo compostas por três macros: *%regspline*, *%regspline_plot* e *%regspline_subset*. Vale mencionar que a primeira macro requer o uso do módulo SAS/STAT e a segunda do SAS/GRAPH.

A primeira, *%regspline*, computa a expansão *b-spline* e estima os parâmetros do modelo logístico, bem como as estimativas de razão de chances em relação ao valor de referência especificado. Um dos argumentos dessa macro permite que o usuário defina a posição dos nós que vai utilizar ou, ainda, permite definir apenas a quantidade de nós cujas posições serão definidas através de um algoritmo interno da macro que calcula os quantis.

A segunda macro, *%regspline_plot* produz o gráfico das estimativas das razões de chances e dos intervalos de confiança, para uma sequência de valores pré-determinados.

A terceira e última rotina, *%regspline_subset*, produz uma tabela com as razões de chances ajustadas e os respectivos intervalos de confiança para valores pré-determinados pelo usuário da variável de exposição x.

O uso dessas macros será exemplificado no Capítulo 4 utilizando dados simulados. Detalhes dos argumentos dessas macros estão apresentados no Anexo.

4 EXEMPLO DE APLICAÇÃO

Este capítulo descreve detalhes sobre o uso das macros apresentadas por Gregory *et al.* (2008) para exemplificar o ajuste de modelos de regressão logística através de *b-splines*.

4.1 Geração do banco de dados

Para fins de ilustração, um conjunto de dados foi simulado considerando um desfecho dicotômico (y) e dois preditores quantitativos – um deles foi gerado considerando uma relação não linear no *logito* (x_1) e o outro representa um confundidor linear (x_2). O Apêndice contém a rotina computacional desenvolvida para a geração dos dados utilizados.

A Tabela 1 resume aspectos das variáveis geradas.

Tabela 1 - Descrição das variáveis simuladas.

Variável	Tipo	Natureza
$\boldsymbol{y}$	Desfecho	Dicotômico
$\boldsymbol{x_1}$	Preditor	Quantitativo
$\boldsymbol{x_2}$	Confundidor	Quantitativo

Foi utilizado o procedimento IML (*Iterative Matrix Language*) do programa SAS – *Statistical Analysis System* (University Edition, disponível em: http://www.sas.com/en_us/software/university-edition.html) para gerar os dois preditores (x_1 e x_2). Este processo inicia com a geração de uma matriz de dados Z, com $n = 10.000$ linhas e componentes Z_1 e Z_2, a partir da distribuição normal bivariada com vetor de médias $(0{,}0)'$ e matriz de correlação $\begin{bmatrix} 1{,}0 & 0{,}6 \\ 0{,}6 & 1{,}0 \end{bmatrix}$.

Na sequência, a matriz de dados U foi definida por meio da função de distribuição acumulada da distribuição normal nos elementos da matriz Z, de tal forma que as colunas de U (U_1 e U_2) têm distribuição Uniforme em $(0{,}1)$, mas

não são independentes. A inversa da função acumulada de distribuição da Normal padrão foi usada para definir a variável x_1, ou seja, $x_1 = \Phi^{-1}(U_1)$. Similarmente, a inversa da distribuição Gama com parâmetros $\alpha = 1$ e $\beta = 1$ foi usada para gerar a variável x_2. Medidas descritivas das variáveis x_1 e x_2 geradas, bem como a matriz de correlações das mesmas são mostradas nas Tabelas 2 e 3, e as Figuras 6 e 7 mostram os respectivos histogramas.

Tabela 2 - Medidas descritivas dos preditores $\boldsymbol{x_1}$ e $\boldsymbol{x_2}$.

Variável	Mínimo	Média	Desvio Padrão	Máximo
$\boldsymbol{x_1}$	-4,4139	0,0090	1,0009	3,6219
$\boldsymbol{x_2}$	< 0,0001	0,9919	0,9908	9,6581

Tabela 3 - Correlação.

	$\boldsymbol{x_1}$	$\boldsymbol{x_2}$
$\boldsymbol{x_1}$	1,0000	0,5440
$\boldsymbol{x_2}$	0,5440	1,0000

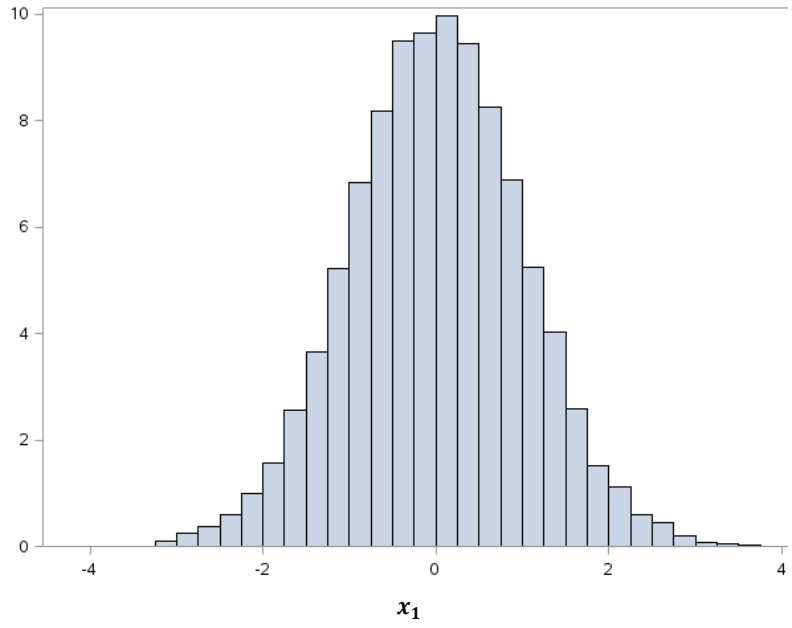

Figura 6: Histograma do preditor x_1 na amostra com n = 10.000.

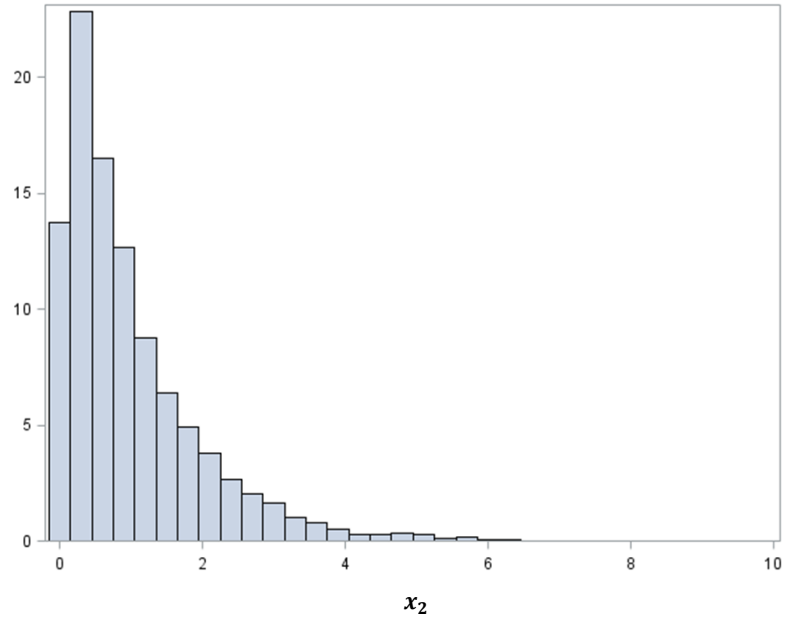

Figura 7: Histograma da variável gerada x_2 na amostra com n = 10.000.

Para gerar a resposta dicotômica Y, foi usada a distribuição Bernoulli com probabilidade de sucesso definida pela equação:

$$p(\boldsymbol{x}) = \frac{exp\{1 - log(3{,}0) \times x_1 + log(0{,}5) \times x_1^2 + log(2{,}0) \times x_1^3 - 0{,}1 \times x_2\}}{0{,}5 + exp\{1 - log(3{,}0) \times x_1 + log(0{,}5) \times x_1^2 + log(2{,}0) \times x_1^3 - 0{,}1 \times x_2\}}$$

em que $\boldsymbol{x} = (x1, x2)'$ e $\varepsilon \sim U(0{,}1)$.

Os coeficientes de regressão $\beta_0 = 1$; $\beta_1 = -log(3{,}0)$; $\beta_2 = log(0{,}5)$; $\beta_3 = log(2{,}0)$ e $\beta_4 = -0{,}1$ foram escolhidos arbitrariamente de forma que o banco gerado fosse plausível e balanceado quanto à variável resposta, ou seja, apresentasse uma quantidade satisfatória de eventos e de não eventos. Isso se faz necessário para que o modelo seja testado em um banco onde exista um balanceamento do desfecho.

Em seguida, para definir a variável resposta Y, para cada linha do banco de dados foi gerado um valor w com distribuição Uniforme (0,1), usado da forma descrita abaixo para atribuir os valores do desfecho:

$$Y = \begin{cases} 1, w < p(x) \\ 0, w \geq p(x) \end{cases}.$$

Dessa forma obteve-se o banco de dados com as variáveis x_1, x_2 e Y. Onde Y é o desfecho e assume valores (0,1), x_1 é um fator de exposição contínuo e x_2 é um confundidor. Essa situação pode ser encontrada em diversas linhas de estudo, tanto na epidemiologia quanto em outras áreas de pesquisa.

A distribuição do desfecho Y ficou bastante balanceada, somando um total de 4.896 observações (48,96%) com $Y = 1$ contra os 5.104 (51,04%) de $Y = 0$. A Figura 8 mostra a relação entre a probabilidade $p(x) = P(Y = 1|x_1, x_2)$ como função do preditor x_1, separadamente para $Y = 0$ e $Y = 1$.

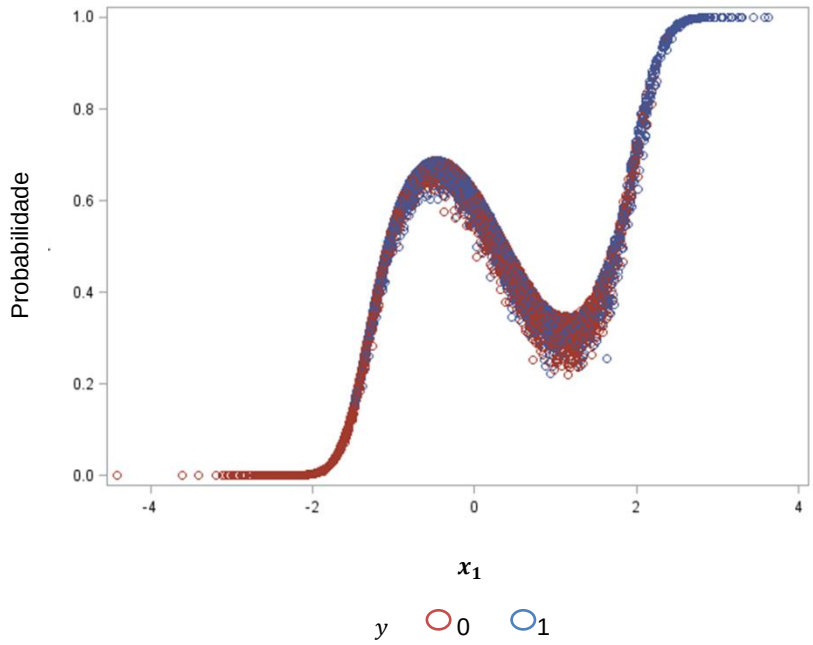

Figura 8: Relação entre a probabilidade $p(x) = P(Y = 1|x_1, x_2)$ como função do preditor x_1.

Na próxima página são encontradas outras figuras que representam a relação da verdadeira $p(x)$ com valores fixos de x_2. Na Figura 9 foi utilizado o primeiro quartil (0,29), na Figura 10, a mediana de x_2 (0,69), e na Figura 11, o terceiro quartil (1,37).

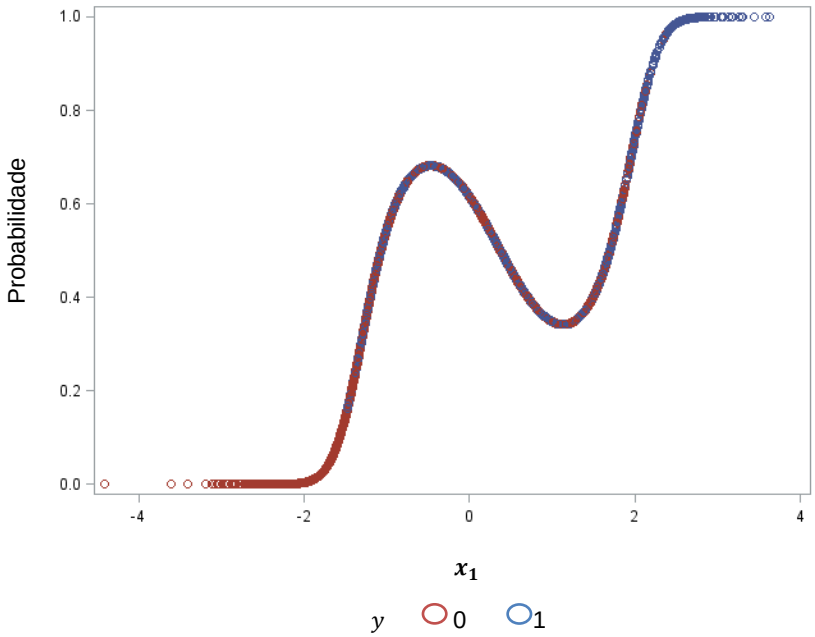

Figura 9: Relação entre a probabilidade $p(x) = P(Y = 1|x_1)$ como função do preditor x_1 fixando $x_2 = 0,29$.

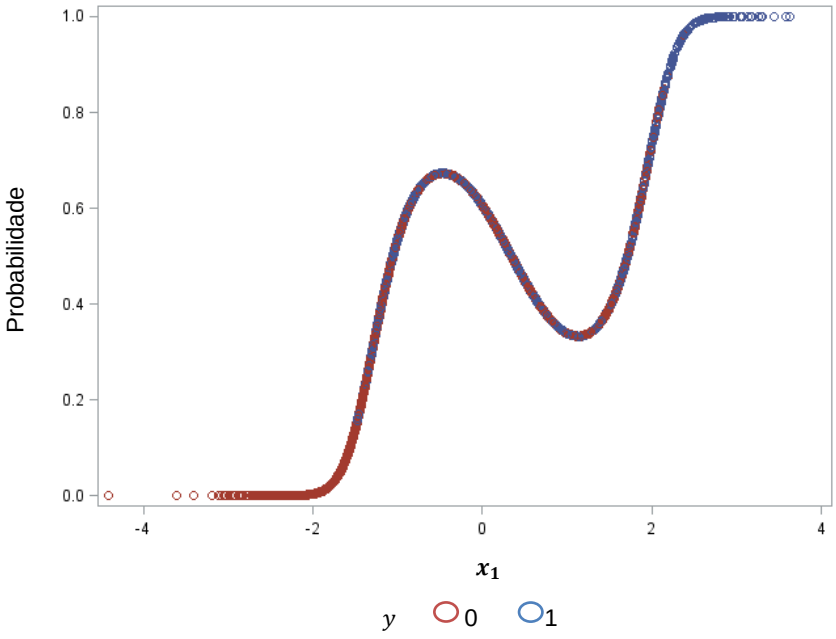

Figura 10: Relação entre a probabilidade $p(x) = P(Y = 1|x_1)$ como função do preditor x_1 fixando $x_2 = 0,69$.

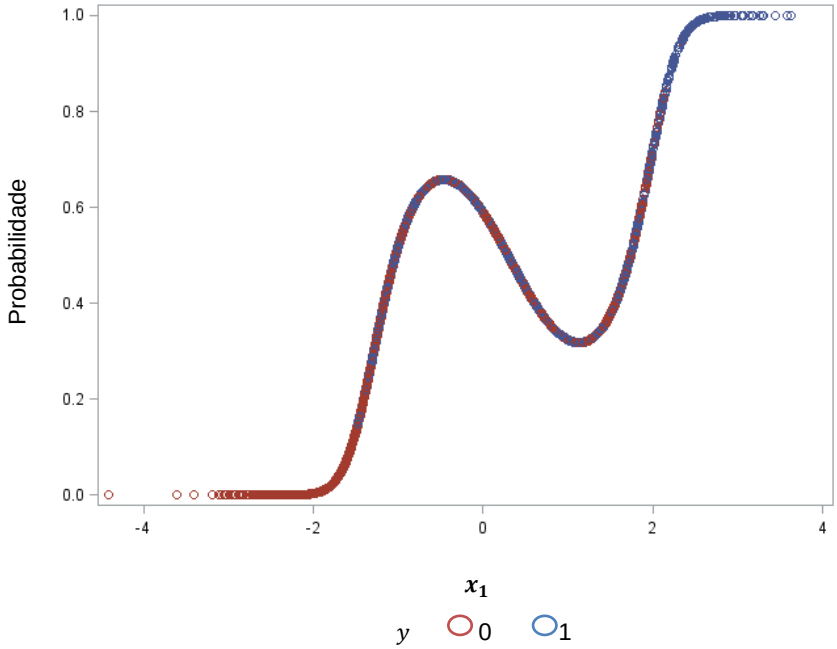

Figura 11: Relação entre a probabilidade $p(x) = P(Y = 1|x_1)$ como função do preditor x_1 fixando $x_2 = 1{,}37$.

4.2 Regressão logística

O primeiro modelo ajustado (Modelo 1) considera a relação linear entre os preditores x_1 e x_2 com o *logito* da probabilidade do evento de $Y = 1$, especificado por:

$$log\frac{P(Y=1|x_1,x_2)}{1-P(Y=1|x_1,x_2)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Na Tabela 4 encontram-se os resultados do ajuste do Modelo 1.

Tabela 4 - Estimativas dos parâmetros (Modelo 1).

Parâmetro	GL	Estimativa	Erro Padrão	Wald Qui-Quadrado	P-valor
Intercepto	1	0,0936	0,0313	8,9626	0,0028
x_1	1	0,2416	0,0242	99,6631	< 0,0001
x_2	1	-0,1390	0,0243	32,6412	< 0,0001

GL: Graus de Liberdade

Na Tabela 5 as correspondentes estimativas de razão de chances. A Tabela 6 mostra estatísticas que podem ser úteis para a comparação de modelos, que serão usadas mais adiante.

Tabela 5 - Estimativas da Razão de Chances (Modelo 1).

Efeito	Estimativa Pontual	Intervalo de Confiança (95%)	
$\boldsymbol{x_1}$	1,273	1,214	1,335
$\boldsymbol{x_2}$	0,870	0,830	0,913

Tabela 6 - Estatísticas de ajuste (Modelo 1).

Critério	Valor
AIC	13860,617
SC	13867,827
-2 log L	13858,617

É importante notar que tanto x_1 quanto x_2 mostraram uma relação linear significativa com o *logito* da probabilidade do evento (p-valor < 0,0001). No entanto, ao contrário de x_2, o preditor x_1 não possui relação linear, como mostrado na Figura 8. Assim, mesmo com um modelo mal especificado, no sentido de que somente o termo linear de x_1 foi considerado, uma relação de linearidade foi identificada pelo Modelo 1, pois de maneira geral, aumentando o valor de x_1 aumenta a probabilidade da ocorrência do evento.

Porém, isto não é verdadeiro para todo o domínio de x_1. A região compreendida entre $x_1 = 0$ e $x_1 = 1$ mostra uma relação não identificada pelo modelo postulado. Consequentemente, a interpretação usual de que para cada aumento de uma unidade em x_1 está associado um aumento de 34,2% na chance do evento é equivocada, especialmente na região supracitada.

A Figura 12 mostra a relação entre x_1 e a probabilidade de ocorrência do evento, ajustada por x_2 (considerando o valor médio $\bar{x}_2 = 0{,}992$) estimada pelo

modelo. Nitidamente o modelo não identifica, nem de perto, a verdadeira relação funcional.

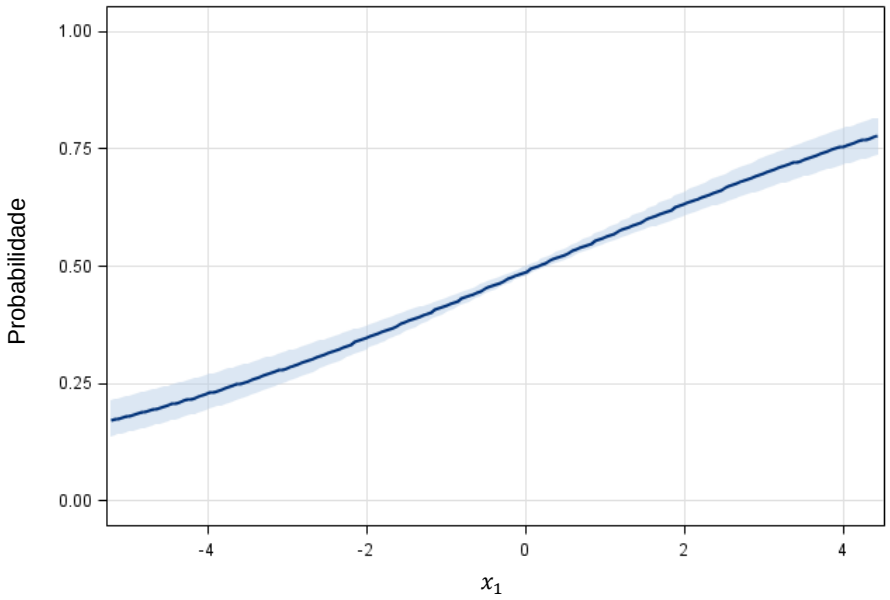

Figura 12: Probabilidades preditas para $y = 1$ com IC 95% para $\bar{x}_2 = 0,992$.

Na sequência, foram incluídos no modelo os termos quadrático e cúbico, para identificar a existência de relação não linear com o *logito*, especificando o modelo (Modelo 2):

$$log\frac{P(Y=1|x_1,x_2)}{1-P(Y=1|x_1,x_2)} = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3 + \beta_4 x_2.$$

A Tabela 7 mostra os resultados do Modelo 2 e as estatísticas usadas para avaliar o ajuste do modelo estão na Tabela 8. Assim, por exemplo, a comparação das estatísticas -2 log L dos Modelos 1 e 2 permitem avaliar a contribuição dos termos x_1^2 e x_1^3 na verossimilhança, ou seja, a contribuição destes termos para explicar a ocorrência do evento, em relação ao Modelo 1.

Assim,

$$(-2logL_1) - (-2logL_2) = 13757,298 - 12383,355 = 1373,943$$

e a hipótese nula $H_0: \beta_2 = \beta_3 = 0$ deve ser rejeitada (p-valor < 0,0001), evidenciando que o Modelo 1 de fato está mal especificado.

Tabela 7 - Estimativas dos parâmetros (Modelo 2).

Parâmetro	GL	Estimativa	Erro Padrão	Wald Qui-Quadrado	P-valor
Intercepto	1	0,4896	0,0372	173,0094	< 0,0001
x_1	1	-1,1162	0,0499	500,2833	< 0,0001
x_1^2	1	-0,6545	0,0314	433,8164	< 0,0001
x_1^3	1	0,6946	0,0259	720,0359	< 0,0001
x_2	1	-0,0978	0,0265	13,6268	< 0,0002

GL: Graus de Liberdade

Tabela 8 - Estatísticas de ajuste (Modelo 2).

Critério	Valor
AIC	12262,142
SC	12298,194
-2 Log L	12252,142

A Figura 13 mostra a relação entre x_1 e a probabilidade de ocorrência do evento, ajustada por x_2 (considerando o valor médio $\bar{x}_2 = 0{,}992$) estimada pelo Modelo 2, captando com clareza a relação não linear.

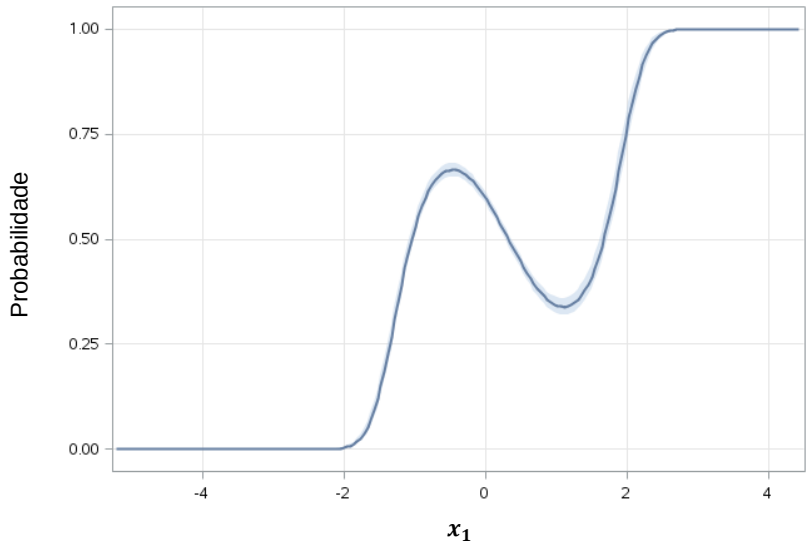

Figura 13: Probabilidades preditas para $y = 1$ para valores de x_1 para $x_2 = 0,992$.

No entanto, uma dificuldade com este modelo é estimar razão de chances da ocorrência do evento para o preditor x_1, que pode ser realizado com relativa flexibilidade utilizando regressão logística com *splines*. A abordagem proposta por Gregory *et al.* (2008), através dos *b-splines* serve perfeitamente pra este propósito.

4.3 Regressão logística utilizando *b-splines*

Esta sessão apresenta uma descrição das macros escritas na linguagem SAS disponibilizadas por Gregory *et al.* (2008), como utilizá-las, e também os resultados dos modelos ajustados. O objetivo principal foi exemplificar o uso das macros em cenários diversos, variando o grau do *spline* e o número de nós. As referidas macros estão disponíveis na página http://www-sciencedirect-com.ez45.periodicos.capes.gov.br/science/article/pii/S0169260708001296#MMCvFirst (acessada em 24/06/15).

4.3.1 Macros

A primeira macro, *%regspline,* é a que, de fato, ajusta o modelo. Para isso, ela depende das seguintes informações: nome do banco onde estão as variáveis, nome do banco de saída, variável dependente, variável *spline*, outras variáveis que serão incluídas no modelo, número de nós internos, grau do polinômio, nome do banco de saída com as estatísticas de ajuste, nome da variável para a qual deve ser estimada a razão de chances e o valor de referência para estimar a razão de chances.

Essa macro retorna dois bancos de dados, o primeiro contém as variáveis originais, as expansões *b-spline*, os valores preditos, os coeficientes da regressão para cada uma das variáveis independentes e as razões de chances comparadas com o valor de referência com seus respectivos intervalos de confiança. O segundo banco de dados retorna as estatísticas usadas para avaliar o ajuste do modelo: o Critério de Informação de Akaike (AIC) e o Critério Bayesiano de Schwarz (SC). A estatística -2 log L pode ser usada para comparar modelos aninhados, dessa forma, não será apresentada

para os próximos modelos. Mais detalhes sobre essas estatísticas estão disponíveis na documentação do procedimento PROC LOGISTIC do programa SAS.

Para ajustar o modelo, primeiramente deve ser usada a macro *%regspline*. A sintaxe do Quadro 1 exemplifica a chamada da macro para o ajuste do modelo *b-spline* com grau um (***degree = 1***) para a variável x_1 (***splinevar = x1***) onde podem ser atribuídas quantas variáveis forem de interesse, usando 4 pontos de corte (***nknots = 4***) definidos pelos quintis da variável x_1 (*default* da macro). Os dados estão no *dataset* BASE1 (***data = base1***), a variável resposta é y1 (***depvar = y1***) e somente o termo linear para a variável x_2 foi especificado (***indepvars = x2***). A instrução ***out = spline_X1*** especifica que a expansão *spline* e seus resultados serão armazenados no *dataset* chamado spline_X1, ***xref = 0*** define o valor de referência para a variável *spline* $x_1 = 0$ como referência para o cálculo das estimativas de razão de chances, ***xref_fuzz = 0,01*** especifica como deve ser o comportamento da macro caso o valor definido em ***xref*** não exista no banco de dados. A instrução ***fit = fit_X1*** especifica o *dataset* no qual serão armazenadas as estatística para avaliação do ajuste do modelo e, finalmente, ***or = or_X1*** define o nome da variável contendo as estimativas de razão de chances e os respectivas estimativas de intervalos de confiança. Uma descrição mais detalhada é apresentada no Anexo A.

```
%regspline(data = base1, out = spline_X1, depvar = Y1, splinevar = X1,
indepvars = X2, nknots = 4, degree = 1, fit = fit_X1, or = or_X1, xref = 0,
xref_fuzz = 0,01);
```

Quadro 1: Chamada da macro *regspline*.

A segunda macro, *%regspline_plot* produz o gráfico da estimativa das razões de chances, utilizando as estimativas armazenadas no *dataset* ***data = spline_X1***, gerado pela rotina *%regspline****.*** É necessário definir a variável spline (***splinevar = X1***) e o seu valor de referência (***xref = 0***), assim como seu *label* (***xref_label = X1***). É preciso definir também o banco onde estão as razões de chances (***riskvar = or_X1***), assim como os seus limites superior (***riskucl =***

or_X1_ucm) e inferior (***risklcl = or_X1_lcm***). Os comandos ***axisheightv = 1.25, axisheightl = 1.25, legendheight = 0,5, linewidth = 1,175)*** especificam opções do gráfico: altura para a fonte usada para os valores do eixo (*default*: 0,75), altura para a fonte usada para os *labels* do eixo (*default:* 0,75), altura para a fonte usada para a legenda (*default:* 0,5) e a largura da linha (*default:* 1,5), respectivamente. No Anexo B podem ser encontrados mais detalhes.

```
%regspline_plot(data = spline_X1, xref = 0, xref_label = X1, splinevar = X1,
riskvar = or_X1, risklcl = or_X1_lcm, riskucl = or_X1_ucm, axisheightv =
1.25, axisheightl = 1.25, legendheight = 0,5, linewidth = 1,175);
```

Quadro 2: Chamada da macro *regspline_plot*.

A última macro, *%regspline_subset*, também depende da rotina *%regspline,* pois utiliza o banco de dados (***data = spline_X1***) que contém as informações necessárias para gerar uma tabela com as razões de chances (***splinevar = X1***) ajustadas e os respectivos intervalos de confiança, onde o limite inferior é definido por ***or_lcm = or_X1_lcm*** e o superior por ***or_ucm = or_X1_ucm***. Essas estimativas consideram os valores da variável de exposição x definidos pelo usuário no argumento **subset = -3 -2 -1 0 1 2 3**. Também é necessário definir o nome do banco de dados que armazena todas essas informações, ***out = rr_list***. O Anexo C contém explicações mais detalhadas.

```
%regspline_subset(data = spline_X1, out = rr_list, subset = -3 -2 -1 0 1 2 3,
splinevar = X1,  or = or_X1, or_lcm = or_X1_lcm, or_ucm = or_X1_ucm);
```

Quadro 3: Chamada da macro *regspline_subset*.

Os resultados do ajuste do modelo gerado pela macro *%regspline* são mostrados parcialmente na Tabela 9, na seção 4.3.2. O gráfico gerado pelo comando %regspline_plot está apresentado na Figura 15, assim como os resultados da macro *%regspline_subet* estão na Tabela 11.

Nos Anexos encontram-se breves explicações sobre a função de cada um dos argumentos que devem ser definidos para a funcionalidade das rotinas.

É recomendado que as três macros sejam rodadas na ordem em que foram apresentadas *%regspline*, *%regspline_plot* e *%regspline_subset*, pois existe uma relação de dependência entre elas.

Um aspecto importante é observar que a macro *%regspline* modela o evento definido pelo menor valor da variável resposta especificada pela instrução que define a variável dependente (***depvar =***). Assim, no exemplo em discussão, foi necessário recodificar a variável resposta y em uma variável y_1, conforme condições abaixo, pois, nesse caso, o interesse é modelar o evento ($y = 1$).

$$y_1 = \begin{cases} 2, y = 0 \\ 1, y = 1 \end{cases}.$$

Para ilustrar o uso das macros, modelos com *b-splines* de diferentes graus e quantidades de pontos de corte serão estudados. Em todos os modelos estimados o valor de referência para a estimação das razões de chances para a variável x_1 será zero.

4.3.2 Regressão logística utilizando *b-spline* de grau um

Em um modelo *b-spline* de grau um as conexões nos pontos de corte não são suavizadas. Para estudar o comportamento do spline de grau um, foram utilizados quatro pontos de corte, ou seja, foram utilizados os quintis da distribuição (Modelo 3). As estimativas dos parâmetros são mostradas na Tabela 9 e as estatísticas de ajuste do modelo estão na Tabela 10. Esses índices serão utilizados posteriormente para comparar o desempenho dos modelos ajustados ao longo deste trabalho.

A Figura 14 mostra a relação funcional entre a probabilidade estimada pelo Modelo 3 e o preditor x_1, que se parece com a Figura 8 (Modelo 2), mas não está suavizada.

Tabela 9 - Estimativas de parâmetros para o modelo *b-spline* com grau um e quatro nós (Modelo 3).

Parâmetro[1]	GL	Estimativa	Erro Padrão	Wald Qui-Quadrado	P-valor
Intercepto	1	3,3521	0,2520	176,9998	< 0,0001
x_1_0	1	-16,5558	0,6712	608,4577	< 0,0001
x_1_1	1	-2,4032	0,2547	88,9952	< 0,0001
x_1_2	1	-2,7240	0,2540	115,0471	< 0,0001
x_1_3	1	-3,1358	0,2417	168,3908	< 0,0001
x_1_4	1	-4,1754	0,2761	228,7692	< 0,0001
x_2	1	3,3521	0,2520	176,9998	< 0,0001

GL: Graus de Liberdade

[1]***Nota:*** *os termos x_1_0, x_1_1, x_1_2, x_1_3 e x_1_4 são os termos da expansão spline.*

Tabela 10 - Estatísticas de ajuste (Modelo 3).

Critério	Valor
AIC	12397,355
SC	12447,827

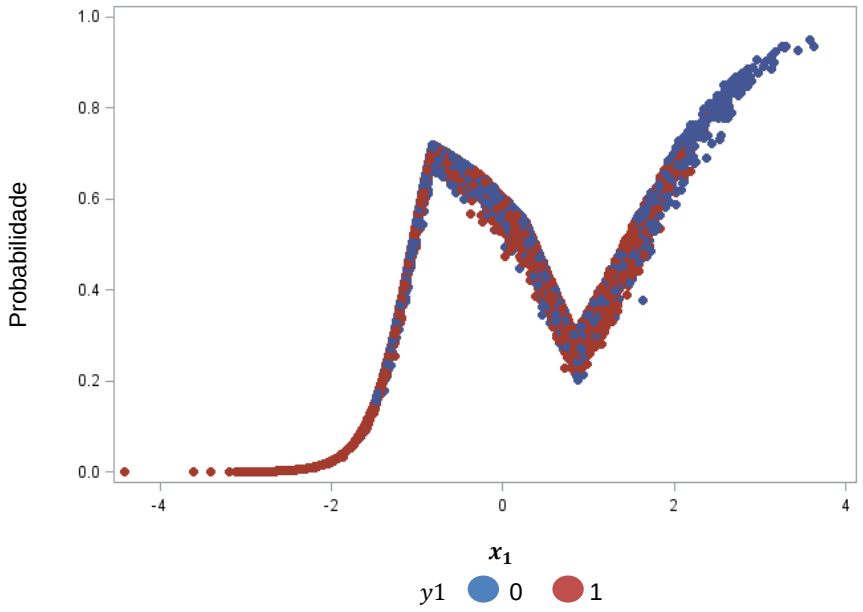

Figura 14: Relação entre a probabilidade estimada como função do preditor x_1 (Modelo 3).

A Tabela 11 apresenta as estimativas de razão de chances (IC 95%) obtidas pelo Modelo 3, para valores selecionados da variável x_1 (entre -3 e 2),

em relação ao valor $\boldsymbol{x_1 = 0}$ (referência), mostradas também na Figura 15. É importante notar que a curva tem arestas nos pontos de corte, pouco comum em situaçoes reais. As informações presentes na Tabela 11 são armazenadas na tabela *rr_list* e, por *default*, não são mostradas na saída do programa. É necessário abrir a tabela para encontrar as estimativas e os limites do intervalo de confiança.

Tabela 11 - Estimativas da Razão de Chances para valores pré-definidos de $\boldsymbol{x_1}$ (Modelo 3).

$\boldsymbol{x_1}$	RC	Intervalo de Confiança (95%)	
2	1,721	1,584	1,870
1	0,341	0,323	0,360
0	1,000	Referência	
-1	0,722	0,685	0,762
-2	0,015	0,013	0,017
-3	< 0,001	< 0,001	< 0,001

Estimativas suavizadas podem ser obtidas por meio de um modelo *b-spline* de grau maior do que um. A reta vermelha representa a estimativa pontual para a razão de chances e as retas azul e verde, os limites superior e inferior para o IC 95%. Essa relação pode ser observada no gráfico da Figura 15.

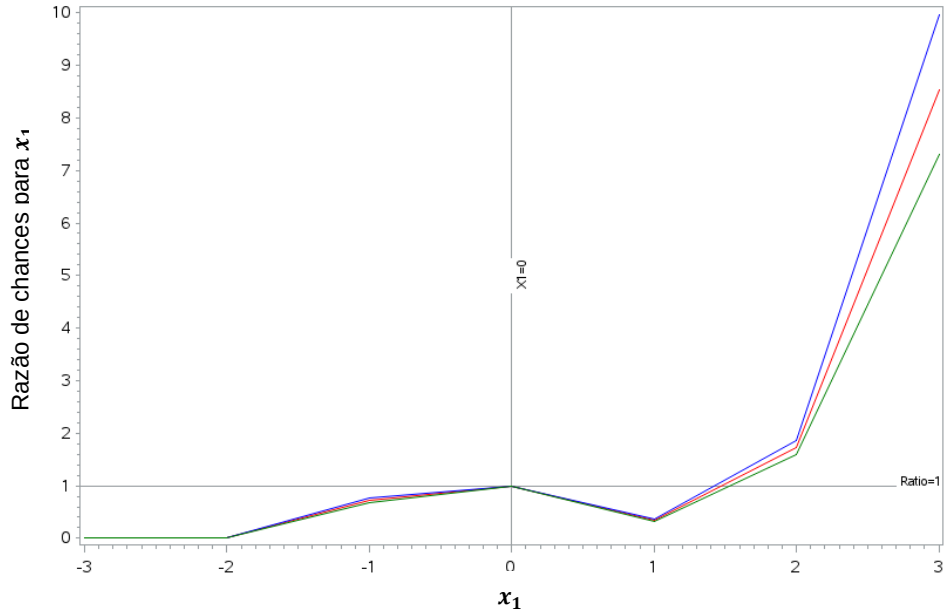

Figura 15: Estimativas da razão de chances (IC 95%) para x_1 (Modelo 3).

O modelo com grau um também foi ajustado considerando $k = 3$ nós, produzindo resultados bastante semelhantes, não serão apresentados neste trabalho.

4.3.3 Regressão logística utilizando *b-spline* de grau dois

Para o Modelo 4 foi definido um polinômio quadrático, de grau dois, a fim de verificar se a forma funcional seria melhor captada por este modelo. O número de nós utilizado também foi quatro.

Na Tabela 12 estão os resultados do ajuste e as estatísticas de diagnóstico para avaliação do modelo estão na Tabela 13. A Figura 16 apresenta o gráfico gerado pela macro *%regspline_plot*, com as estimativas de razão de chances (IC 95%) para os valores -3, -2, -1, 0, 1 e 2. O valor 3 foi excluído pois sua estimativa é muito extrema e, se incluído no gráfico, dificulta a visualização para os outros valores. A Tabela 14 apresenta as estimativas para as razões de chances.

Tabela 12 - Estimativas de parâmetros para o modelo *b-spline* com grau dois e quatro nós (Modelo 4).

Parâmetro	GL	Estimativa	Erro Padrão	Wald Qui-Quadrado	P-valor
Intercepto	1	14,7897	1,1645	161,3120	< 0,0001
x_1_0	1	-66,4689	6,0944	118,9534	< 0,0001
x_1_1	1	-16,1094	1,2956	154,6142	< 0,0001
x_1_2	1	-13,8005	1,1585	141,9016	< 0,0001
x_1_3	1	-14,4630	1,1793	150,4044	< 0,0001
x_1_4	1	-14,8450	1,1293	172,7900	< 0,0001
x_1_5	1	-17,2110	1,3862	154,1648	< 0,0001
x_2	1	-0,0960	0,0264	13,1965	< 0,0001

GL: Graus de Liberdade

Tabela 13 - Estatísticas de ajuste (Modelo 4).

Critério	Valor
AIC	12256,747
SC	12314,430

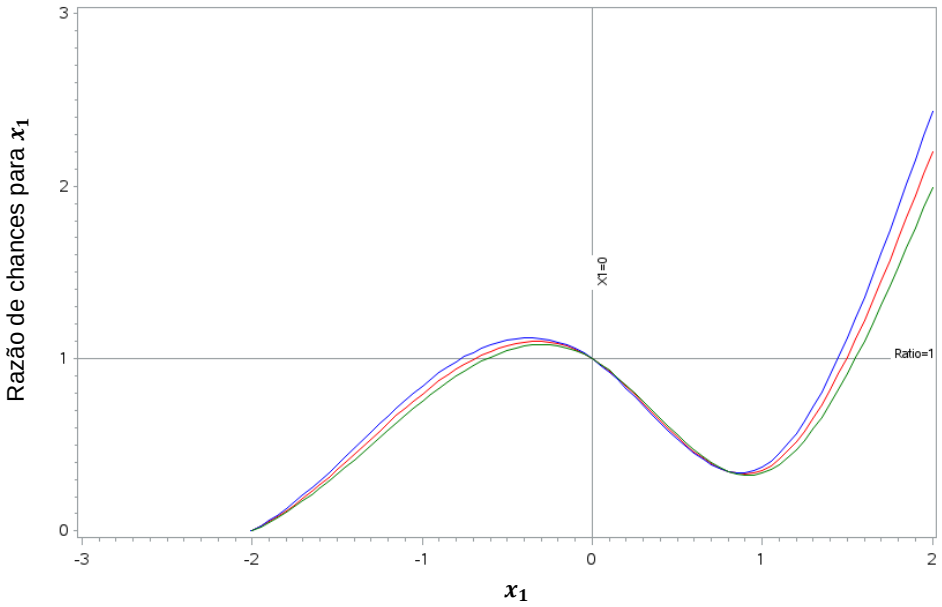

Figura 16: Estimativas da razão de chances (IC 95%) para x_1 (Modelo 4).

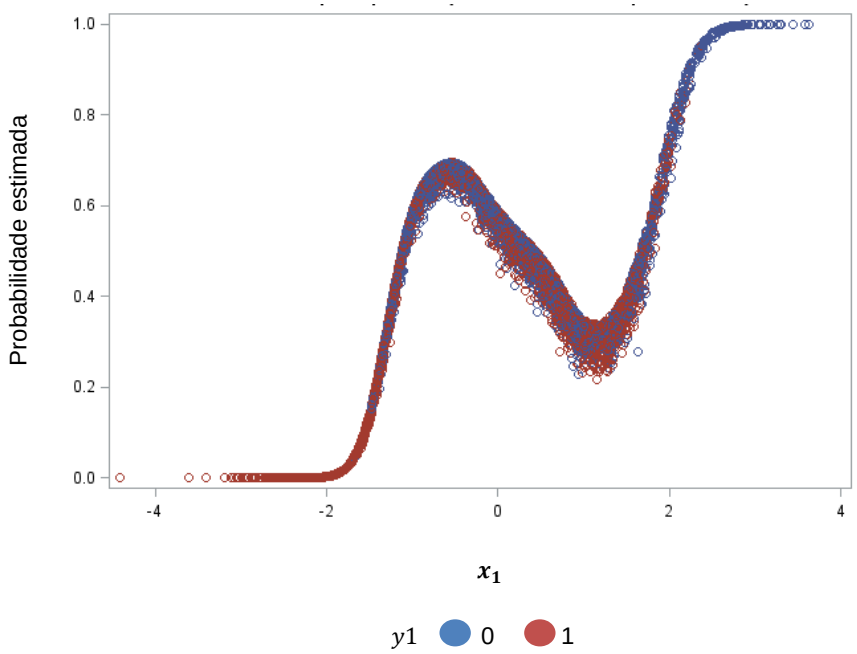

Figura 17: Probabilidade estimada como função de x_1 (Modelo 4)

Através das Figuras 16 e 17 observa-se que o modelo quadrático é muito mais flexível que o modelo linear com *spline* e foi capaz de captar corretamente a forma funcional da relação de x_1 com a variável resposta $y1$.

Tabela 14 - Estimativas da Razão de Chances para valores pré-definidos de x_1 (Modelo 4).

x_1	RC	Intervalo de Confiança (95%)	
2	2,201	1,991	2,433
1	0,353	0, 334	0,373
0	1,00	Referência	
-1	0,795	0,751	0, 842
-2	0,003	0,002	0,004
-3	< 0,001	< 0,001	< 0,001

Também foi ajustado um modelo *b-spline* de grau dois utilizando três nós, no entanto, os resultados não serão apresentados, pois, assim como no modelo de ordem um, foram bastante semelhantes.

4.3.4 Regressão logística utilizando *b-spline* de grau três

O *b-spline* de ordem três se assemelha ao *spline* cúbico não restrito. Essa é a abordagem preferida dos autores, pois é um modelo que combina flexibilidade e boas propriedades de estimação, além de ser o mais explorado na literatura e, consequentemente, o que tem mais referências e exemplos aplicados. A quantidade de nós utilizada no ajuste desse modelo (Modelo 5) foi de quatro pontos de corte.

Na Tabela 15 são apresentadas as estatísticas de diagnóstico para avaliar o ajuste do Modelo 5.

Na Tabela 16 são apresentadas as estimativas para os coeficientes do modelo. Conforme já mencionado na revisão bibliográfica, um modelo de grau três $(m = 3)$ e quatro nós $(k = 4)$ resulta em $k + m + 1 = 8$ coeficientes, um para cada expansão *spline*. No entanto, um dos termos da expansão *spline* é escrito como uma função dos outros e, portanto, não é estimado. Por exemplo, na Tabela 15, o termo x_1_7 pode ser escrito como: $x_{1_7} = intercepto - (x_1\_0 + x_1\_1 + x_1\_2 + x_1\_3 + \ldots x_1\_6)$.

Tabela 15 - Estatísticas de ajuste (Modelo 5).

Critério	Valor
AIC	12254,760
SC	12326,027

Tabela 16 - Estimativas de parâmetros para o Modelo *b-spline* com grau três e quatro nós (Modelo 5).

Parâmetro	GL	Estimativa	Erro Padrão	Wald Qui-Quadrado	P-valor
Intercepto	1	23,8029	4,3378	30,1103	< 0,0001

x_1_0	1	-89,9670	41,0246	4,8093	0,0283
x_1_1	1	-40,2958	5,0090	64,7169	< 0,0001
x_1_2	1	-22,0806	4,4570	24,5434	< 0,0001
x_1_3	1	-23,1317	4,3133	28,7609	< 0,0001
x_1_4	1	-23,6505	4,3790	29,1690	< 0,0001
x_1_5	1	-25,2334	4,1362	37,2183	< 0,0001
x_1_6	1	-23,6730	5,2732	20,1538	< 0,0001
x_2	1	-0,0974	0,0264	13,5910	0,0002

GL: Graus de Liberdade

A Figura 18 e a Tabela 17 apresentam as estimativas das razões de chance (IC 95%) para os valores pré-definidos da variável x_1.

É possível observar que a curva ajustada na Figura 16 (Modelo 4) é muito semelhante à curva ajustada para o gráfico presente na Figura 18 (Modelo 5). Isso pode ser comprovado também comparando as estimativas dos coeficientes nas Tabelas 14 e 17. Além disso, o Critério de Informação de Akaike (AIC) teve um decréscimo insignificante, ou seja, os Modelos 4 e 5 apresentados alcançaram resultados extremamente parecidos e são praticamente equivalentes do ponto de vista estatístico.

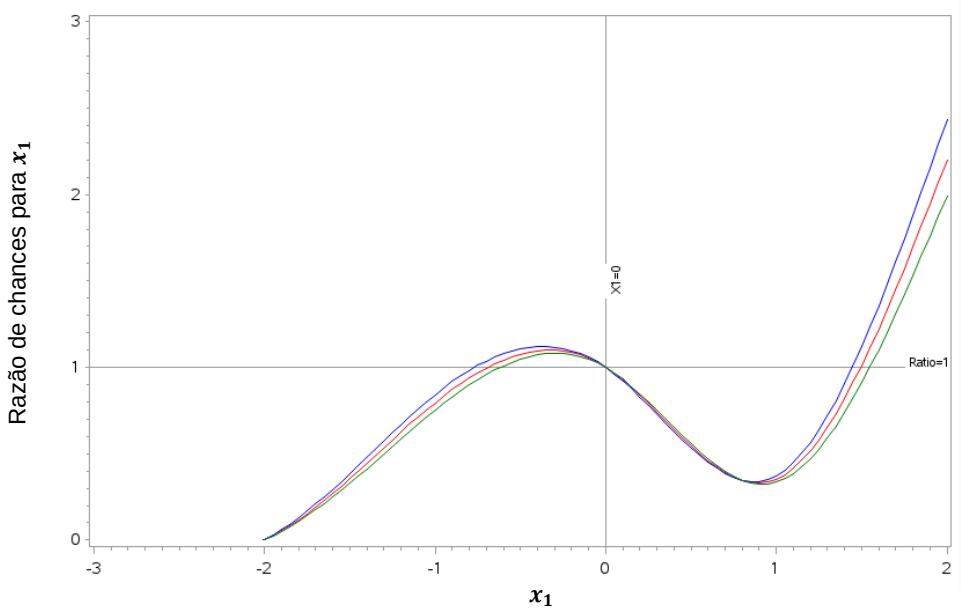

Figura 18: Estimativas da razão de chances (IC 95%) para x_1 (Modelo 5).

Tabela 17 - Estimativas da Razão de Chances para valores pré-definidos de $\boldsymbol{x_1}$ (Modelo 5).

$\boldsymbol{x_1}$	RC	Intervalo de Confiança (95%)	
2	2,084	1,901	2,287
1	0,366	0,347	0,385
0	1,00	Referência	
-1	0,781	0,742	0,823
-2	0,002	0,001	0,003
-3	< 0,001	< 0,001	< 0,001

A partir das informações da Tabela 17, pode-se dizer que aos indivíduos com valor $x_1 = 2$, está associado a um aumento médio de 108,4% na chance de ocorrer o evento $y1$, comparados aos indivíduos com $x_1 = 0$. Porém, essa interpretação é exclusivamente para a variação pontual de 0 para 2. Similarmente, na comparação entre $x_1 = 1$ versus $x_1 = 0$, o modelo sugere que está associada uma redução média de 73,3% na chance de ocorrência do evento. Essa relação pode ser claramente observada no gráfico apresentado pela Figura 18, onde a curva que passa em $x_1 = 0$ tem um pequeno decréscimo em relação ao valor $x_1 = 1$, mas em seguida aumenta rapidamente quando $x_1 = 2$.

Foi ajustado também um modelo *b-spline* de ordem três utilizando três nós. O resultado foi muito semelhante ao Modelo 5 e não será apresentado neste trabalho.

4.3.5 Regressão logística utilizando *b-spline* de grau três e oito nós

Os Modelos 3, 4 e 5, de ordem um, dois e três, respectivamente, foram testados com três e quatro nós e os resultados foram extremamente semelhantes. Para verificar se um verdadeiro aumento de nós é impactante, optou-se por ajustar um modelo de *b-spline* cúbico com oito nós internos (Modelo 6).

Na Tabela 18 são apresentadas as estimativas para os coeficientes do Modelo 6 e na Tabela 19 as estatísticas de diagnóstico para avaliação do modelo.

Tabela 18 - Estimativas de parâmetros para o modelo *b-spline* com grau três e oito nós (Modelo 6).

Parâmetro	GL	Estimativa	Erro Padrão	Wald Qui-Quadrado	P-valor
Intercepto	1	15,1581	7,1507	4,4936	0,0340
x_1_0	1	-146,7000	156,7000	0,8768	0,3941
x_1_1	1	-33,1196	10,7958	9,4116	0,0022
x_1_2	1	-17,6209	7,2291	5,9413	0,0148
x_1_3	1	-14,2209	7,1491	3,9569	0,0467
x_1_4	1	-14,4337	7,1561	4,0683	0,0437
x_1_5	1	-14,4836	7,1432	4,1112	0,0426
x_1_6	1	-15,0755	7,1662	4,4255	0,0354
x_1_7	1	-15,0803	7,1226	4,4828	0,0342
x_1_8	1	-15,5808	7,2071	4,6737	0,0306
x_1_9	1	-16,6633	6,8439	5,9281	0,0149
x_1_10	1	-9,9654	8,6159	1,3378	0,2474
x_2	1	-0,0961	0,0265	13,2079	0,0003

GL: Graus de Liberdade

Tabela 8 - Estatísticas de *goodness of fit* (Modelo 6).

Critério	Valor
AIC	12263,946
SC	12357,681

Pela Tabela 19 verifica-se que, com o aumento no número de nós, a estatística de diagnóstico AIC também aumentou algumas unidades, quando comparado ao Modelo 5. Mesmo que o modelo tenha mais nós, essa estatística penaliza o acréscimo de variáveis no modelo - pois com a

transformação da variável x_1, foram criadas novas expansões *spline*, cuja quantidade depende diretamente no número de nós k e do grau do polinômio m.

A Figura 19 e a Tabela 20 apresentam as estimativas das razões de chance (IC 95%) para os valores pré-determinados da variável x_1.

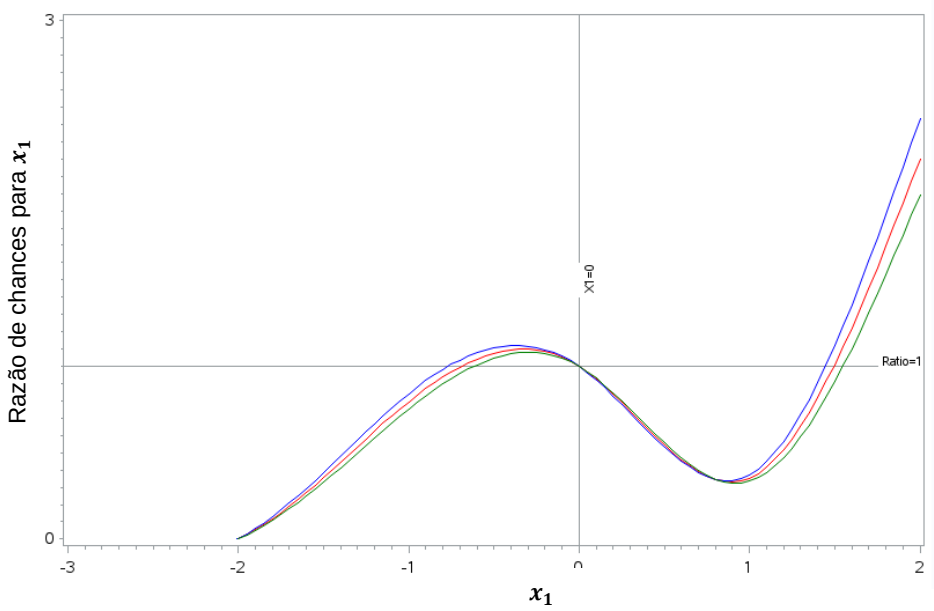

Figura 19: Estimativas da razão de chances (IC 95%) para x_1 (Modelo 6).

Tabela 20 - Estimativas da Razão de Chances para valores pré-definidos de x_1 (Modelo 6).

x_1	RC	Intervalo de Confiança (95%)	
2	2,089	1,919	2,274
1	0,366	0,350	0,383
0	1,00	Referência	
-1	0,734	0,700	0,770
-2	< 0,001	< 0,001	< 0,001
-3	< 0,001	< 0,001	< 0,001

4.3.5 Regressão logística utilizando *b-spline* de grau oito

Conforme já foi verificado, aumentar o número de nós (passando de três ou quatro para oito) não impactou diretamente nos resultados do modelo. O objetivo agora é aumentar significativamente o grau do polinômio para avaliar se essa mudança modifica os resultados do ajuste. Dessa forma, o Modelo 7 possui quatro pontos de corte e é de ordem oito.

Na Tabela 21 estão as estimativas para os coeficientes do Modelo 7 e na Tabela 22 as estatísticas de diagnóstico do modelo.

Tabela 21 - Estimativas de parâmetros para o modelo *b-spline* com grau oito e quatro nós (Modelo 7).

Parâmetro	GL	Estimativa	Erro Padrão	Wald Qui-Quadrado	P-valor
Intercepto	1	14,2944	600,7000	0,0006	0,9810
x_1_0	1	-798,1000	2761,3000	0,0835	0,7726
x_1_1	1	496,0000	1236,3000	0,1610	0,6883
x_1_2	1	-305,6000	582,1000	0,2756	0,5996
x_1_3	1	70,8079	632,9000	0,0125	0,9109
x_1_4	1	-61,8816	589,6000	0,0110	0,9164
x_1_5	1	34,1807	612,9000	0,0031	0,9555
x_1_6	1	-53,4696	590,2000	0,0082	0,9278
x_1_7	1	15,8864	610,1000	0,0007	0,9792
x_1_8	1	-36,9148	593,7000	0,0039	0,9504
x_1_9	1	23,5197	618,5000	0,0014	0,9697
x_1_10	1	-65,4767	553,6000	0,0140	0,9059
x_1_11	1	77,5416	747,6000	0,0108	0,9174
x_2	1	-0,0962	0,0264	13,2483	0,0003

GL: Graus de Liberdade

Tabela 22 - Estatísticas de ajuste (Modelo 7).

Critério	Valor
AIC	12264,877
SC	12365,821

A Figura 20 e a Tabela 23 apresentam as estimativas das razões de chance (IC 95%) para os valores pré-determinados da variável x_1 do Modelo 7.

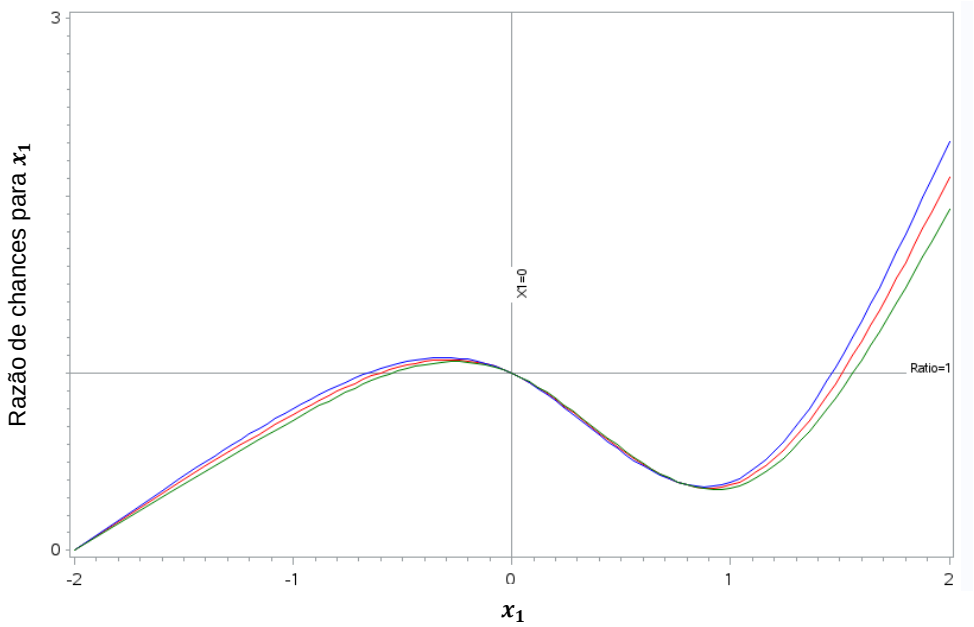

Figura 20: Estimativas da razão de chances (IC 95%) para x_1 (Modelo 7).

Tabela 23 - Estimativas da Razão de Chances para valores pré-definidos de x_1 (Modelo 7).

x_1	RC	Intervalo de Confiança (95%)	
2	2,104	1,921	2,396
1	0,369	0,354	0,385
0	1,00	Referência	
-1	0,766	0,735	0,799
-2	< 0,001	< 0,001	0,001

A partir dos resultados apresentados na Tabela 23 e na Figura 20, é possível identificar semelhanças entre o Modelo 7 e os outros ajustados anteriormente (Modelos 4, 5 e 6).

A Tabela 24 faz um comparativo de todos os modelos ajustados, inclusive os que não foram apresentados neste trabalho (com três nós). Uma das ferramentas para avaliação do ajuste (*goodness of fit*) é o Critério de

Informação de Akaike (AIC), e, dentre os modelos com transformação *spline*, o melhor foi aquele com quatro nós e de ordem três (Modelo 5). Entretanto, é importante salientar que as diferenças entre as medidas de AIC entre alguns modelos foram pequenas, como acontece nos Modelos 4 e 5, tanto com três quanto com quatro nós.

O Modelo 2, com termos polinomiais, também apresentou um bom desempenho e seria adequado para estudar a associação das variáveis x_1 e x_2 com o desfecho y. No entanto, conforme citado anteriormente, a dificuldade com este modelo é estimar razão de chances da ocorrência do evento para o preditor x_1. Nesse contexto as rotinas propostas por Gregory *et al.* (2008) utilizando *b-splines* serviram perfeitamente.

Tabela 24 – Estatísticas AIC para os modelos ajustados.

Modelo	*b-spline*	Nº de nós ($\boldsymbol{k}$)	Grau ($\boldsymbol{m}$)	AIC
1	-	-	1	13860,617
2	-	-	3	12262,142
-	✓	3	1	12496,781
3	✓	4	1	12397,355
-	✓	3	2	12263,405
4	✓	4	2	12256,747
-	✓	3	3	12258,620
5	✓	4	3	12254,760
6	✓	8	3	12263,946
-	✓	3	8	12257,044
7	✓	4	8	12264,877

A macro permite modelar simultaneamente mais do que uma variável utilizando *spline*, que pode ser especificado na chamada da macro conforme no Quadro 4. No entanto, no caso estudado neste trabalho, não faz sentido a aplicação em função de que x_2 tem relação linear no logito.

```
%regspline(data = base1, out = spline_X1, depvar = Y1, splinevar = X1 X2,
indepvars = X2, nknots = 4, degree = 1, fit = fit_X1, or = or_X1, xref = 0,
xref_fuzz = 0,01);
```

Quadro 4: Chamada da macro com 2 variáveis *spline*.

5 CONSIDERAÇÕES FINAIS

Relações não lineares entre preditores quantitativos e desfechos dicotômicos são encontradas com frequência, principalmente em estudos de associação e de dose-resposta na área da pesquisa clínica. Uma prática comum para contornar esse problema é o uso da categorização dos preditores quantitativos - o que acarreta uma série de problemas nas suposições do Modelo e não tem sido bem recomendado por diversos autores (Bennette *et al.*, 2012; Greenland, 1995; Turner *et al.*, 2010).

Nesse contexto, este trabalho apresenta um método alternativo para o ajuste do Modelo de regressão logística, quando se tem o interesse de estimar as associações (razão de chances). O uso de *splines* permite ajustar um modelo versátil e que se aproxima à relação entre os preditores e a variável resposta (ou uma função da resposta, como no modelo logístico), sem imposição de restrições.

Um modelo com termos polinomiais, como o Modelo 2, apresentado neste trabalho, conseguiu captar corretamente a relação não linear característica da variável x_1 com a variável resposta y. No entanto, não é factível estimar a razão de chances para diferentes valores da variável x_1. Além disso, a multicolinearidade também pode ser apontada como um possível problema a ser investigado na inclusão de termos polinomiais, visto que os termos do modelo x_1^2 e x_1^3 são parcialmente correlacionadas com a variável original x_1.

Dessa forma, a abordagem através de *splines* é particularmente útil quando se deseja estimar associações. Os diferentes modelos apresentados de regressão logística utilizando *b-splines* (Modelos 3 ao 7) foram assertivos e mostraram-se capazes em identificar a relação da forma funcional da variável x_1 com o desfecho, inclusive o Modelo 3, cuja ordem definida para o polinômio foi igual a um.

Duas dificuldades reportadas pelos autores (Wegman *et al.*, 1983; Harrell, 2001) são a escolha da quantidade de nós e suas respectivas posições. Com o banco de dados gerado e através dos modelos ajustados,

observa-se que não há uma regra geral a ser seguida, até por que os modelos apresentaram estatísticas AIC bastante próximas. Uma possibilidade é limitar os testes utilizando de três a sete nós internos, segundo sugerido por Harrell (2001), e posicioná-los conforme os quantis da variável com transformação *spline*. A macro desenvolvida por Gregory *et al.* (2008) já trata os nós dessa forma, caso o usuário não defina a posição dos mesmos.

Os modelos de regressão logística utilizando *b-splines* permitem estimar a razão de chances para valores pontuais, cujas interpretações são exclusivas de cada valor, em relação a um valor referência, definido pelo usuário. O Modelo 5, por exemplo, que obteve o menor AIC dentre os modelos com transformação *spline*, permite interpretações como: para indivíduos com valor $x_1 = 2$ está associado um aumento (médio) de 108,4% na chance de ocorrer o evento em estudo ($y = 1$), comparados aos indivíduos com $x_1 = 0$, ao passo que para indivíduos com $x_1 = 0$ está associada uma redução (média) de 73,3% na chance de ocorrência do desfecho em comparação aos indivíduos com valor $x_1 = 1$. Estas estimativas estão ajustadas pela variável x_2. Além disso, o modelo permite fazer predições para valores que estejam dentro do *range* observado para a variável *spline* x_1.

Em estudos epidemiológicos, por exemplo, é muito comum encontrar relações do com uma forma de J (*j-shaped curve*) como ilustrada pela associação entre a pressão sanguínea e o risco de doença arterial coronariana (Takahashi *et al.*, 2013). O modelo de regressão logística utilizando *b-splines* proposto neste trabalho parece ser ideal para ajustar esse tipo de associação e estimar a razão de chances com IC 95%.

Como continuidade a este trabalho, sugere-se uma exploração mais profunda do método, cobrindo aspectos em relação à análise de resíduos do Modelo e também de diagnósticos, os quais foram pouco explorados. Além disso, sabe-se da existência de outras rotinas computacionais capazes de estimar a razão de chances para o método de regressão logística utilizando *splines*, uma delas, inclusive, desenvolvida por F. Harrell através de *splines* cúbicos restritos. Dessa forma, recomenda-se testar outras macros, inclusive em outros pacotes, para ampliar o tema.

Outros métodos não paramétricos alternativos estão disponíveis para estudar esse tipo de relação como, por exemplo, modelos aditivos generalizados, regressão local (LOESS ou LOWESS - *locally weighted scatterplot smoothing*), entre outros. Detalhes e aplicações dessas técnicas estão disponíveis na documentação dos procedimentos PROC GAM e PROC LOESS do programa SAS, por exemplo.

A abordagem com *b-splines* explorada neste trabalho tem grande potencial para estudar a associação entre preditores quantitativos e um desfecho dicotômico (ou uma função da resposta) quando se tem interesse em estudar a razão de chances. Sem dúvida, a técnica merece mais atenção. Contudo, conforme apontado por *Gregory et al.* (2008), a análise utilizando *splines* é trabalhosa e as interpretações podem ser bastante complexas, dessa forma, é importante comparar o *trade-off* entre um modelo complicado e o resultado do ajuste, de forma a avaliar se o modelo via *splines* fornece um ajuste significativamente melhor que outras alternativas citadas.

6 REFERÊNCIAS BIBLIOGRÁFICAS

Bartels RH, Beatty JC, Barsky BA. An Introduction to Splines for Use in Computer Graphics and Geometric Modelling. *Morgan Kaufmann, Berkeley,* 1987.

Bennette C, Vickers A. Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. *Medical Research Methodology. 2012; 12(21) :1-5.*

Bindinganavle K. An optimal approach to geometric trimming of B-splines surfaces. URL: http://scholar.lib.vt.edu/theses/available/etd-04192001-172731/unrestricted/chapter_4.pdf, acesso em 10/06/15. (2000)

Eilers PHC, Marx BD. Flexible Smoothing with B-splines and Penalties. *Statistical Science, 11, 2, 89-102.* (1996)

Greenland S. Dose-Response and Trend Analysis in Epidemiology: Alternatives to Categorical Analysis. *Epidemiology, 6, 4, 356-365.* (1995)

Gregory M, Ulmer H, Pfeiffer KP, Lang S, Strasak, AM. A set of SAS macros for calculating and displaying adjusted odds ratio (with confidence intervals) for continuous covariates in logistic B-spline regression models. *Computer Methods and Programs in Biomedicine, 92, 109-114.* (2008)

Harrell, FE Jr. Regression Modeling Strategies. NY: *Springer-Verlag,*2001.

Hastie TJ, Tibshirani, RJ. Generalized Additive Models. *Chapman & Hall, New York.* (1990)

Keith SW, Alisson DB. A free-knot spline modeling framework for piecewise linear logistic regression in complex samples with body mass index and mortality as an example. *National Institute of Health, 16.* (2014)

Kelly R. Extending Linear Models: Non-Linearity. URL: https://rpubs.com/ryankelly/GAMs, acesso em 17/06/15. (2014)

SAS/STAT® 13.2 User's Guide. SAS Institute Inc. Cary, NC. URL: http://support.sas.com/documentation/cdl/en/statug/67523/HTML/default/viewer.

htm#statug_chap0000018_statug_chap0000018_sect001.htm, acesso em 24/06/15. (2013)

Silverman, BW. Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting. *Journal of the Royal Statistical Society, 47, 1, 1-52.* (1985)

Stone SJ, Koo, CY. Additive splines in statistics. *Proceedings of the Statistical Computing Section ASA, 45-48.* (1985)

Takahashi K, Nakao H, Hattori S. Cubic Spline Regression of J-shaped Dose-Response Curves with Likelihood-based Assignments of Grouped Exposure Levels. *Biometrics & Biostatistics, 4(5):1-6.* (2013)

Turner EL, Dobson JE, Pocock SJ. Categorization of continuous risk factors in epidemiological publications: a survey of current practice. *Epidemiology Perspectives & Innovations, 7:9.* (2010)

Wegman EJ, Wright IW. Splines in Statistics. *Journal of the American Statistical Association, 78, 382, 351-365.* (1983)

Wold, S. Spline functions in data analysis. *Technometrics, 16, 1-11.* (1974)

ANEXOS

Anexo A – Argumentos da macro SAS *regspline*

Anexo B – Argumentos da macro SAS *regspline_plot*

Anexo C – Argumentos da macro SAS *regspline_subset*

ANEXO A - Argumentos da macro SAS *%regspline*.

Parâmetro	Descrição
DATA	Nome do banco de dados de entrada (*input*). Deve conter a variável resposta e as variáveis independentes.
OUT	Nome do banco de dados de saída (*output*). Além das variáveis originais, conterá as expansões *spline* os coeficientes de regressão, as razões de chances e os respectivos intervalos de confiança (95%).
OUTTEST	Nome do banco de dados com os coeficientes.
DEPVAR	Variável dependente do Modelo.
SPLINEVAR	Uma ou mais variáveis independentes que serão substituídas por expansões *splines*.
INDEPVARS	Variáveis restantes que serão incluídas no Modelo.
KNOTS	Lista os nós que serão usados para cada variável *spline*. Essa variável deverá ser preenchida caso o pesquisador já saiba a posição dos nós.
NKNOTS	Especifica o número de nós internos para cada expansão *spline*. Essa variável deverá ser preenchida caso o pesquisador não saiba a posição dos nós.
DEGREE	Especifica o grau do polinômio utilizado para gerar cada expansão spline (o *default* é grau três).
XREF	Especifica o valor de referência para cada variável *spline* no cálculo da razão de chances.
XREF_FUZZ	Especifica como deverá ser o comportamento da macro caso o valor definido em XREF não esteja presente no banco de dados. Se XREF_FUZZ=0, a macro para automaticamente. Especificando um valor diferente de zero, a macro busca um valor entre XREF e XREF_FUZZ no banco de dados e o utiliza como referência para o cálculo da razão de chances.
DEBUG	Opções de debug.
LINK	Variável original (*default*), log, etc.
MODEL_OPTS	Opções do Modelo.
FIT	Nome do banco de dados para as estatísticas de ajuste.
OR	Nome da variável para a qual será calculada a razão de chances.

LOGOR	Fornece o $log(RC)$, o default é fornecer apenas a RC.

Anexo B - Argumentos da macro SAS *%regspline_plot.*

Parâmetro	Descrição
DATA	Nome do banco de dados que contém os dados para fazer o gráfico (banco de saída da macro %regspline).
SPLINEVAR	Nome da variável *spline* (será colocada no eixo x).
RISKVAR	Nome da variável para a qual será calculada a razão de chances.
RISKLCL	Nome da variável que contém o menor valor para o limite de RISKVAR.
RISKULC	Nome da variável que contém o maior valor para o limite de RISKVAR.
XREF	Valor para a linha vertical de referência.
XREF_LABEL	Label para a linha vertical de referência.
LINEWIDTH	Largura da linha (*default* é 1,5).
AXISHEIGHTV	Altura para a fonte usada para os valores do eixo (*default* é 0,75).
AXISHEIGHTL	Altura para a fonte usada para os labels do eixo (*default* é 0,75).
LEGENDHEIGHT	Altura para a fonte usada para a legenda (*default* é 0,5).

Anexo C - Argumentos da macro SAS *%regspline_subset*.

Parâmetro	Descrição
DATA	Nome do banco de dados de entrada (*input*).
OUT	Nome do bando de dados de saída (*output*).
SUBSET	Especifica uma lista de valores pré-especificados da variável spline, para os quais se deseja obter a razão de chances.
OR	Nome da variável que contém a razão de chances.
OR_LCM	Nome da variável que contém o limite inferior (95%) para a razão de chances.
OR_UCM	Nome da variável que contém o limite superior (95%) para a razão de chances.

APÊNDICES

Apêndice A – Sintaxe para gerar banco de dados

Apêndice B – Modelagem

Apêndice A – Sintaxe para gerar banco de dados

```
options ps=58 ls=80 nocenter nodate nonumber formchar='|----|+|---+=|-/\<>*';
* Gera distribuicao dos preditores correlacionados;
proc iml;
        call randseed(54);
        * Especifica a matriz de correlacoes da distribuicao normal multivariada
          com medias 0 e variancias 1;
        sigma = {1.0 0.6,
                 0.6 1.0};
        * Gera 10000 obervacoes da normal multivariada;
    Z = randnormal(1e4, {0,0}, sigma);
        * Se F a f.d. da v.a. X, entao U=F(X)~U(0,1)
          a funcao 'cdf' aplica a f.d. da v.a.;
        U  = cdf("normal", Z); * As colunas de U sao v.a. U(0,1), mas nao sao
ind.;
        * Se U~U(0,1), entao X=invF(U) ~ F;
        X1 = quantile("Normal", U[,1],0,1);     /* X1 ~ Normal */
        X2 = quantile("Gamma", U[,2],1.0);      /* X2 ~ Gama   */
        X = X1||X2;
        /* Se Z ~ MVN(0,Sigma), corr(X) geralmente é proxima de Sigma,
           em que X=(X1,X2,...,Xm) and X_i = F_i^{-1}(Phi(Z_i)) */
        rhoZ = corr(Z);
        rhoX = corr(X);
        print rhoZ rhoX;
        varNames = ("X1":"X2");
        create PREDITORES from X[c=varNames];
        append from X;
        close PREDITORES;
run;
```

```
quit;
* Verificando distribuicoes;
proc means data=PREDITORES min mean std var max;
        var X1 X2;
run;
proc corr data=PREDITORES;
        var X1 X2;
run;
proc sgplot data=PREDITORES;
   histogram X1;
run;
proc sgplot data=PREDITORES;
   histogram X2;
run;
*********************************************************
  Gera relacao funcional com a resposta Y e Y
*********************************************************;
data MODELO;
        set PREDITORES;
        intercepto = 0.5;
        beta11 = - log(3.0);
        beta12 = + log(0.5);
        beta13 = + log(2.0);
  XBETA = intercepto + Beta11*X1 + Beta12*(X1*X1) + Beta13*(X1*X1*X1)-
0.1*X2;
  P = logistic(XBETA);
  format P f8.4;
  U = uniform(56);
  Y = (U < P);
run;
```

```
options ls=80;
proc freq data=MODELO;
        table Y;
run;
proc sgplot data=MODELO;
        scatter x=X1 y=P / group=Y markerattrs=(symbol=dot);
run;
proc means data=MODELO q1 median q3;
        var X2;
   output out=atX2values;
run;
data GRAFICO1;
        set PREDITORES;
        intercepto = 0.5;
        beta11 = - log(3.0);
        beta12 = + log(0.5);
        beta13 = + log(2.0);
  XBETA_X2A = intercepto + Beta11*X1 + Beta12*(X1*X1) +
Beta13*(X1*X1*X1)- 0.1*0.29;
  P_X2A = logistic(XBETA_X2A);
  format P_X2A f8.4;
  U = uniform(56);
  Y_X2A = (U < P_X2A);
  XBETA_X2B = intercepto + Beta11*X1 + Beta12*(X1*X1) +
Beta13*(X1*X1*X1)- 0.1*0.69;
  P_X2B = logistic(XBETA_X2B);
  format P_X2B f8.4;
  Y_X2B = (U < P_X2B);
  XBETA_X2C = intercepto + Beta11*X1 + Beta12*(X1*X1) +
Beta13*(X1*X1*X1)- 0.1*1.37;
  P_X2C = logistic(XBETA_X2C);
```

```
  format P_X2C f8.4;
  Y_X2C = (U < P_X2C);
run;
proc sgplot data=GRAFICO1;
        scatter x=X1 y=P_X2A / group=Y_X2A markerattrs=(symbol=dot);
run;
proc sgplot data=GRAFICO1;
        scatter x=X1 y=P_X2B / group=Y_X2B markerattrs=(symbol=dot);
run;
proc sgplot data=GRAFICO1;
        scatter x=X1 y=P_X2C / group=Y_X2C markerattrs=(symbol=dot);
run;
proc sgplot data=MODELO;
        where Y=1;
        scatter x=X1 y=P / group=Y markerattrs=(symbol=dot);
run;
proc logistic data=MODELO plots(only)=effect(clband);
        model Y(event='1') = X1 X1*X1 X1*X1*X1 X2/ rl;
run;
proc logistic data=MODELO plots(only)=effect(clband);
        model Y(event='1') = X1 X2/ rl;
run;
```

Apêndice B – Modelagem

Obs.: Antes de rodar a sintaxe, certificar-se de que as macros *sasname*, *regspline*, *regspline_plot* e *regspline_subset* já foram carregadas.

```
data BASE1;
        set MODELO;
        Y1 = (Y = 0) + 1;
run;
proc freq data=BASE1;
        table Y1*Y / nopercent nocol norow;
run;
proc means data=BASE1 min mean std max;
        var Y Y1 X1 X2;
run;
* Modelos com Grau = 1;
title1 "Modelo B-Spline para X1 (nknots=4, Grau do polinomio=1)";
%regspline(data=BASE1, out=Spline_X1, depvar=Y1, splinevar=X1, indepvars=X2 ,
        NKNOTS=4, degree=1, fit=fit_X1, or=or_X1, xref=0, xref_fuzz=0.001, _debug_=0);
%regspline_plot(data=Spline_X1, color=y, xref=0, xref_label=X1, splinevar=X1, riskvar=or_X1,
            risklcl=or_X1_lcm, riskucl=or_X1_ucm, axisheightv=1.25, axisheightl=1.25, legendheight=0.5, linewidth=1.175);
data Spline_X1;
        set Spline_X1;
        XBETAhat = log(p_Y1/(1-p_Y1));
run;
options ls=120;
proc means data=Spline_X1 min mean median max;
        var p_Y1 XBETAhat;
run;
```

```
proc sgplot data=Spline_X1;
        scatter x=X1 y=XBETAhat / group=Y1 markerattrs=(symbol=DOT);
run;
proc sgplot data=Spline_X1;
        scatter x=X1 y=p_Y1 / group=Y1 markerattrs=(symbol=CircleFilled);
run;
%regspline_subset(data=Spline_X1, out=rr_list, subset= -2.5 -1 0 1 2 3, splinevar=X1,
            or=or_X1, or_lcm=or_X1_lcm, or_ucm=or_X1_ucm);
%regspline_plot(data=rr_list, color=y, xref=0, xref_label=X1, splinevar=X1, riskvar=or_X1,
            risklcl=or_X1_lcm, riskucl=or_X1_ucm, axisheightv=1.25, axisheightl=1.25, legendheight=0.5, linewidth=1.175);
* Modelos com Grau = 2;
title1 "Modelo B-Spline para X1 (nknots=3, Grau do polinomio=2)";
%regspline(data=BASE1, out=Spline_X1, depvar=Y1, splinevar=X1, indepvars=X2 ,
        NKNOTS=3, degree=2, fit=fit_X1, or=or_X1, xref=0, xref_fuzz=0.001, _debug_=1);
data Spline_X1;
        set Spline_X1;
        XBETAhat = log(p_Y1/(1-p_Y1));
run;
options ls=120;
proc means data=Spline_X1 min mean median max;
        var p_Y1 XBETAhat;
run;
proc sgplot data=Spline_X1;
        scatter x=X1 y=p_Y1 / group=Y1 markerattrs=(symbol=dot);
run;
proc sgplot data=Spline_X1;
```

```
        scatter x=X1 y=XBETAhat / group=Y1 markerattrs=(symbol=star);

run;

%regspline_plot(data=Spline_X1, color=y, xref=0, xref_label=X1, splinevar=X1,
riskvar=or_X1,

        risklcl=or_X1_lcm, riskucl=or_X1_ucm, axisheightv=1.25,
axisheightl=1.25, legendheight=0.5, linewidth=1.175);

%regspline_plot_splined(data=Spline_X1, color=y, xref=0, xref_label=X1,
splinevar=X1, riskvar=or_X1,

        risklcl=or_X1_lcm, riskucl=or_X1_ucm, axisheightv=1.25,
axisheightl=1.25, legendheight=0.5, linewidth=1.175);

%regspline_subset(data=Spline_X1, out=rr_list, subset=-3 -2 -1 0 1 2,
splinevar=X1,

        or=or_X1, or_lcm=or_X1_lcm, or_ucm=or_X1_ucm);

%regspline_plot_splined(data=rr_list, color=y, xref=0, xref_label=X1,
splinevar=X1, riskvar=or_X1,

        risklcl=or_X1_lcm, riskucl=or_X1_ucm, axisheightv=1.25,
axisheightl=1.25, legendheight=0.5, linewidth=1.175);

* Modelos com Grau = 3;

title1 "Modelo B-Spline para X1 (nknots=3, Grau do polinomio=3)";

%regspline(data=BASE1, out=Spline_X1, depvar=Y1, splinevar=X1,
indepvars=X2 ,

        NKNOTS=3, degree=3, fit=fit_X1, or=or_X1, xref=0, xref_fuzz=0.001,
_debug_=1);

data Spline_X1;

        set Spline_X1;

        XBETAhat = log(p_Y1/(1-p_Y1));

run;

options ls=120;

proc means data=Spline_X1 min mean median max;

        var p_Y1 XBETAhat;

run;

proc sgplot data=Spline_X1;

        scatter x=X1 y=XBETAhat / group=Y1 markerattrs=(symbol=CircleFilled);
```

```
run;

%regspline_plot(data=Spline_X1, color=y, xref=0, xref_label=X1, splinevar=X1,
riskvar=or_X1,

        risklcl=or_X1_lcm, riskucl=or_X1_ucm, axisheightv=1.25,
axisheightl=1.25, legendheight=0.5, linewidth=1.175);

%regspline_plot_splined(data=Spline_X1, color=y, xref=0, xref_label=X1,
splinevar=X1, riskvar=or_X1,

        risklcl=or_X1_lcm, riskucl=or_X1_ucm, axisheightv=1.25,
axisheightl=1.25, legendheight=0.5, linewidth=1.175);

%regspline_subset(data=Spline_X1, out=rr_list, subset=-3 -2 -1 0 1 2,
splinevar=X1,

        or=or_X1, or_lcm=or_X1_lcm, or_ucm=or_X1_ucm);

%regspline_plot_splined(data=rr_list, color=y, xref=0, xref_label=X1,
splinevar=X1, riskvar=or_X1,

        risklcl=or_X1_lcm, riskucl=or_X1_ucm, axisheightv=1.25,
axisheightl=1.25, legendheight=0.5, linewidth=1.175);

* Modelos com Grau = 3, 8 nos;

title1 "Modelo B-Spline para X1 (nknots=8, Grau do polinomio=3)";

%regspline(data=BASE1, out=Spline_X1, depvar=Y1, splinevar=X1,
indepvars=X2 ,

        NKNOTS=8, degree=3, fit=fit_X1, or=or_X1, xref=0, xref_fuzz=0.001,
_debug_=1);

data Spline_X1;

        set Spline_X1;

        XBETAhat = log(p_Y1/(1-p_Y1));

run;

options ls=120;

proc means data=Spline_X1 min mean median max;

        var p_Y1 XBETAhat;

run;

proc sgplot data=Spline_X1;

        scatter x=X1 y=XBETAhat / group=Y1 markerattrs=(symbol=CircleFilled);

run;
```

```
%regspline_plot(data=Spline_X1, color=y, xref=0, xref_label=X1, splinevar=X1,
riskvar=or_X1,

        risklcl=or_X1_lcm, riskucl=or_X1_ucm, axisheightv=1.25,
axisheightl=1.25, legendheight=0.5, linewidth=1.175);

%regspline_plot_splined(data=Spline_X1, color=y, xref=0, xref_label=X1,
splinevar=X1, riskvar=or_X1,

        risklcl=or_X1_lcm, riskucl=or_X1_ucm, axisheightv=1.25,
axisheightl=1.25, legendheight=0.5, linewidth=1.175);

%regspline_subset(data=Spline_X1, out=rr_list, subset=-3 -2 -1 0 1 2,
splinevar=X1,

        or=or_X1, or_lcm=or_X1_lcm, or_ucm=or_X1_ucm);

%regspline_plot_splined(data=rr_list, color=y, xref=0, xref_label=X1,
splinevar=X1, riskvar=or_X1,

        risklcl=or_X1_lcm, riskucl=or_X1_ucm, axisheightv=1.25,
axisheightl=1.25, legendheight=0.5, linewidth=1.175);

* Modelos com Grau = 8, 4 nos;

title1 "Modelo B-Spline para X1 (nknots=4, Grau do polinomio=8)";

%regspline(data=BASE1, out=Spline_X1, depvar=Y1, splinevar=X1,
indepvars=X2 ,

      NKNOTS=4, degree=8, fit=fit_X1, or=or_X1, xref=0, xref_fuzz=0.001,
_debug_=1);

data Spline_X1;

      set Spline_X1;

      XBETAhat = log(p_Y1/(1-p_Y1));

run;

options ls=120;

proc means data=Spline_X1 min mean median max;

      var p_Y1 XBETAhat;

run;

proc sgplot data=Spline_X1;

      scatter x=X1 y=XBETAhat / group=Y1 markerattrs=(symbol=CircleFilled);

run;

%regspline_plot(data=Spline_X1, color=y, xref=0, xref_label=X1, splinevar=X1,
riskvar=or_X1,
```

```
        risklcl=or_X1_lcm, riskucl=or_X1_ucm, axisheightv=1.25,
axisheightl=1.25, legendheight=0.5, linewidth=1.175);

%regspline_plot_splined(data=Spline_X1, color=y, xref=0, xref_label=X1,
splinevar=X1, riskvar=or_X1,

        risklcl=or_X1_lcm, riskucl=or_X1_ucm, axisheightv=1.25,
axisheightl=1.25, legendheight=0.5, linewidth=1.175);

%regspline_subset(data=Spline_X1, out=rr_list, subset=-2 -1 0 1 2,
splinevar=X1,

        or=or_X1, or_lcm=or_X1_lcm, or_ucm=or_X1_ucm);

%regspline_plot_splined(data=rr_list, color=y, xref=0, xref_label=X1,
splinevar=X1, riskvar=or_X1,

        risklcl=or_X1_lcm, riskucl=or_X1_ucm, axisheightv=1.25,
axisheightl=1.25, legendheight=0.5, linewidth=1.175);
```