MODELO DE REGRESSÃO LOGÍSTICA E PROBITO.

FORTALEZA

2022

UNIVERSIDADE FEDERAL DO CEARÁ

DEPARTAMENTO DE ESTATÍSTICA E MATEMÁTICA APLICADA - DEMA

PROFESSOR:

Prof. Dr.Juvêncio Santos Nobre.

ALUNOS:

Thalis Rebouças de Oliveira.

Claudia Lima da Silva.

Sumário

0.1 Introdução

Os ser humano sempre procurou uma maneira de entender o mundo com o advento da criação da matemática conseguimos calcular e modelar por meio de funções simples ou mais complexa alguns fenômenos, com o desenvolvimento da probabilidade conseguimos entender os fenômenos por meio de funções de probabilidades. A probabilidade foi um dos princípios para tentar modelar, análise e saber a chance de um fenômeno ocorrer.Atualmente com os avanços da estatística e as ajudas do computadores,conseguimos chegar em um nível muito bom em regressões lineares, seja a regressão simples ou múltipla, mas tudo tem suas qualidades e defeitos e os defeitos da regressões são que precisamos de condições especificas para que ocorra regressão para modelar os fenômenos aleatórios. na Década de 70 e 80 foi a época que teve o crescimento do desenvolvimento dos Modelos de Regressão Lineares Generalizados (MLG's) sendo Nelder e Wedderburn (1972) foram os primeiros a estudar e escrever sobre o assunto, de maneira a entender e modelar mesmo quando a suposição de normalidade ou mesmo utilizado algumas transformações, como a do Box e Cox (1964), não fosse atendidas. Com isso vamos explicar os modelos generalizados Probito e o modelo Logístico.

0.2 Modelos Lineares Generalizados (MLG)

0.2.1 O que são Modelos Lineares Generalizados

A explicação geral descrita no livro do Gauss e Clarice(2013) é da seguinte forma:

> Esses modelos envolvem uma variável resposta uni-variada, variáveis explanatórias e uma amostra aleatória de n observações independentes, sendo que:
>
> i) a variável resposta, componente aleatório do modelo, tem uma distribuição pertencente à família de distribuições exponenciais que engloba as distribuições normal,gama e normal inversa para dados contínuos; binomial para proporções; Poisson e binomial negativa para contagens;
>
> ii) as variáveis explanatórias entram na forma de uma estrutura linear, constituindo o componente sistemático do modelo;
>
> iii) a ligação entre os componentes aleatório e sistemático é feita por meio de uma função adequada como, por exemplo, logarítmica para os modelos log-lineares,denominada função de ligação.

Com isso a variável resposta seguindo, por exemplo uma Poisson e que a relação funcional entre a média de Y e o preditor linear é dada por $log(\mu) = n$, com isso a relação funcional é adequada, uma vez que para quaisquer valores dos parâmetros do preditor linear é um valor positivo para μ. Outro exemplo comum é tentar ver a variável resposta por meio de proporção e a distribuição mais conhecida é a Binomial e uma relação funcional é o $\log(\frac{\mu}{1-\mu})$, neste caso a proporção de sucesso esperada fica entre $0 \leq \mu \leq 1$.

Deste modo um dos pontos dos MLG's é a família uniparamétrica exponencial, que muitas famílias de probabilidades fazem parte, como a normal, binomial, gamma, entre outras. Essa família é caracterizada por uma função de densidade ou probabilidade na forma de:

$$f(x;\theta) = h(x)exp[n(\theta)t(x) - b(\theta)]$$

Neste caso a função descrita acima tem valores no subconjuntos dos reais e não são única. O suporte da família exponencial são para valores reais maiores que zero e não pode depender de parâmetros e temos que a estatística $t(x)$ é suficiente para θ, pelo teorema de Neyman-Fisher.

0.2.1.1 Componente aleatório

Sabemos que os modelos lineares generalizados tem um componente aleatório e esse componente é um conjunto de variáveis aleatórias independentes provenientes de uma distribuição exponencial uniparamétrica na forma canônica.Nelder e Wedderburn (1972) ao proporem essa modelagem, conseguiram incorporar distribuições biparaméticas no componente aleatório do modelo.

$$f(y;\theta,\phi) = exp\{\phi^{-1}[y\theta - b(\theta)] + c(y,\phi)\}$$

Deste modo com a o ϕ é conhecido, temos uma família exponencial linear, que é idêntica à família exponencial na forma canônica, além disso, $b(.)$ e $c(.)$ são conhecidos também. Temos que:

$$E(Y) = \mu = b'(\theta) \; e \; Var(Y) = \phi b''(\theta).$$

O resultado acima está provado no livro de Gauss e Clarisse na seção (1.4).

0.2.1.2 Componente sistemático

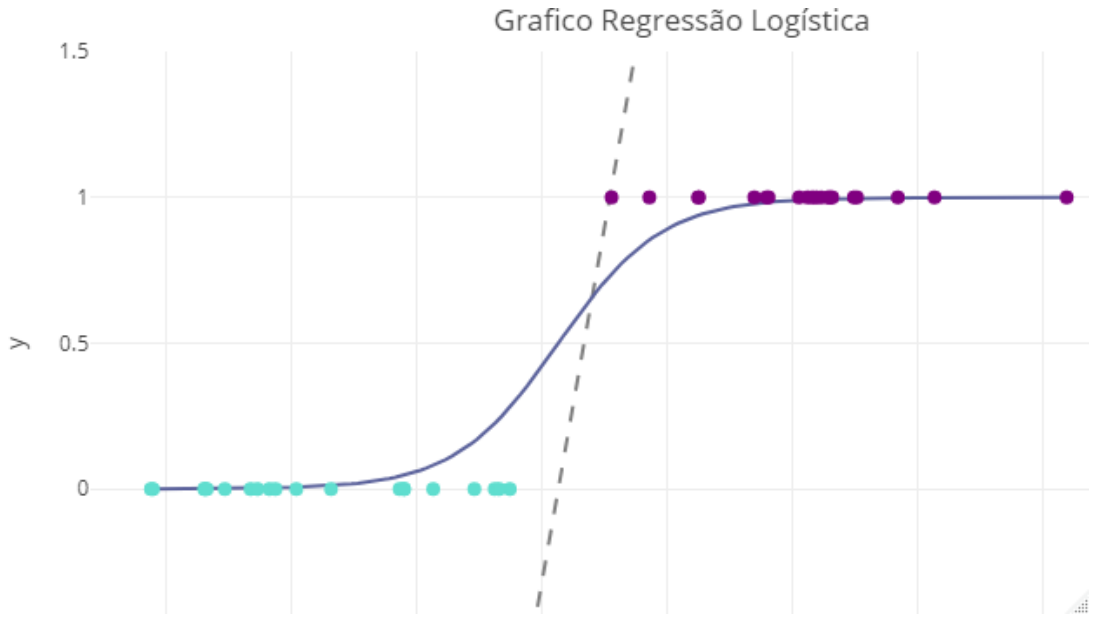
Temos no livro de Gauss e Clarisse a seguinte explicação desse componente:

> O componente sistemático é estabelecido durante o planejamento (fundamental para a obtenção de conclusões confiáveis) do experimento, resultando em modelos de regressão (linear simples, múltipla, etc.), de análise de variância (delineamentos inteiramente causalizados, causalizados em blocos, quadrados latinos com estrutura de tratamentos fatorial, parcelas subdivididas, etc.) e de análise de covariância. O componente aleatório é especificado assim que são definidas as medidas a serem realizadas, que podem ser contínuas ou discretas, exigindo o ajuste de diferentes distribuições.

Com isso temos que esse componente represente ao preditor linear, que é o conjunto das combinações lineares dos parâmetros.

$$Yi = \beta_0 + \beta_1 x_{i1} + \ldots + B_p x_{ip}$$

Figura 1 – Gráfico do modelo de regressão logístico



Fonte:(HAIR et al.,2009).

E neste caso a **função ligação** é uma função real, monótona e diferenciável que associa o componente aleatório, mais especificamente a média da sua distribuição, com a função de linearizar a relação entre os componentes aleatório e sistemático.

$$g(\mu_i) = Y_i = \beta_0 + \beta_1 x_{i1} + \ldots + B_p x_{ip} s$$

Com isso a função de ligação é o $g(.)$.

0.3 Modelos Lineares Generalizados para Variável Categórica

0.3.1 Logística

O termo ”regressão” como um conceito estatístico foi utilizado primeiramente pelo pesquisador britânico Francis Galton (1822-1911) em estudos sobre hereditariedade. A técnica da regressão logística foi descoberta no século XIX em estudos sobre o crescimento das populações e as reações químicas no curso de autocatálise. Em 1845 Pierre-François Verhulst (1804-1849) publicou um artigo na revista Proceedings, no qual define a curva de crescimento populacional por meio de uma função denominada por ele de “logística” (SOUZA, 2006).

Os modelos de regressão são utilizados em análises estatísticas nas quais se busca descrever as relações entre a variável resposta (Y) e a variável explicativa (X). Quando se tem apenas uma variável independente, pode-se estabelecer uma regressão linear simples.

$$Y_i = \beta_0 + \beta_1 xi + e_i, i = 1, \ldots, n$$

onde, Y_i é a variável resposta, xi a variável explanatória, β_0 e β_1 os parâmetros de regressão e e_i o erro do modelo. Para os casos em que se têm duas ou mais variáveis explanatórias pode-se modelar os dados por meio da regressão linear múltipla, a qual se constitui uma generalização do modelo anterior:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + ... + \beta_h x_{hi} + ei, i = 1, \ldots, n$$

A regressão logística se diferencia dos modelos de regressão linear porque no seu caso a variável dependente é qualitativa e binária. Na regressão logística a variável resposta assume apenas valores 0 e 1, sendo geralmente "1" a ocorrência do evento de interesse e "0" a sua ausência, ou em outros termos, "1" corresponde ao sucesso e "0" ao fracasso.Portanto, o valor da previsão de Y sempre estará no intervalo $0 \geq Y \leq 1$. Conforme pontua Hair et al (2009), devido a natureza binária da variável dependente,as suposições da regressão linear e múltipla são violadas. Neste sentido, os resíduos seguem distribuição binomial ao invés da normal e a variância não é constante, apresentando heterocedasticidade, além disso, as transformações não são suficientes para corrigir essas violações. Sendo assim, a regressão logística é um método que se ocupa particularmente com esses problemas.

0.3.2 Probito

Este modelo foi utilizado no estudo do Bliss em 1935, na área da saúde para calcular a curva de dosagem-mortalidade, neste caso o estudo fazia a inferência que a dosagem era medida a partir da mortalidade observada na suposição de que a suscetibilidade é distribuída normalmente, tais dosagens inferidas, em termos de unidades que eram chamadas de probits.

Neste caso precisamos de um ligação probito, que se dá por:

$$\phi^{-1}(\mu) = n$$

em que $\phi(.)$ é a função de distribuição da normal padrão.

Com isso, considerando um a variável tem uma V que possui um distribuição normal de média μ pertencente aos reais e $\sigma^2 > 0$:

$$F_v(v, \mu, \sigma^2) = \sqrt{2\pi\sigma^2}^{-1} \; exp\left[-\frac{(v-\mu)^2}{2\sigma^2}\right]$$

Deste modo, temos uma função não-linear em um conjunto linear de parâmetros, em que é linearizado:

$$probit(\pi_i) = \phi^{-1}(\pi_i) = \beta_1 + \beta_2 d_i$$

Com,

$$\beta_1 = -\frac{\mu}{\sigma} \; e \; \beta_2 = \frac{1}{\sigma d_i}$$

logo ,

$$\pi_i = \phi(\beta_1 + \beta_2 d_i)$$

Este resultado é obtido pela normalização da variável V, mais detalhes está no livro da Gauss e Clarice(2013) na página 31.

De maneira conceitual vamos ter a seguinte forma de regressão.

$$Y = X^T \ \beta + \mu$$

O parâmetro β é estimado por máxima verosimilhança, todas as provas necessárias deste modelo são disponibilizadas no Wikipédia no artigo de *Probit model*. A variância é dada por:

$$Var(Y_i|n_i) = \frac{\mu_i(1-\mu_i)}{n_i}$$

No R temos o comando **glm(Y ~ x, family=binomial(link="logit"))**, neste comando precisamos da variável resposta Y_i seguidas das variáveis explicativas x_i.

0.4 Técnicas de diagnóstico para Modelos Lineares Generalizados

A maioria dos diagnósticos para modelos lineares se estendem de forma relativamente diretamente aos GLMs.Essas extensões normalmente tiram vantagem do cálculo dos estimadores de máxima verossimilhança e máxima quase-verossimilhança para GLMs obtidos por mínimos quadrados ponderados iterados.O trabalho sobre a extensão de diagnósticos de mínimos quadrados lineares para GLMs foi feito por Pregibon (1981), Landwehr, Pregibon e Shoemaker (1984), Wang (1985, 1987) e Williams (1987).O ajuste final por mínimos quadrados ponderados lineariza o modelo e fornece uma aproximação quadrática para o logaritmo da verossimilhança. Os diagnósticos aproximados são baseados diretamente na solução WLS ou são derivados de estatísticas facilmente calculadas a partir desta solução.

Modelos lineares ajustados por mínimos quadrados fazem suposições fortes e às vezes os dados não seguem essas suposições. Quando essas condições não são atendidas, as estimativas de mínimos quadrados podem se comportar de maneira inadequada e podem até mostrar os dados de maneira completamente incorreta e incoerente com a realidade. Os diagnósticos de regressão podem revelar esses problemas e, muitas vezes, apontar o caminho para as soluções deste problemas e podemos até elencar algumas possíveis geradores desse problema.

Conforme Souza (2006) é importante que se faça uma análise dos resíduos e diagnósticos do modelo ajustado a fim de detectar possíveis problemas, como por exemplo:

1. Presença de observações discrepantes.
2. Inadequação das pressuposições para os erros aleatórios ou para as médias.
3. Colinearidade entre as colunas da matriz do modelo.
4. Forma funcional do modelo inadequada.
5. Presença de observações influentes.

0.4.1 Matriz Chapéu ou Diagonal da matriz H

Os elementos de h_i, da matriz H, para um modelo linear generalizado podem ser obtidos diretamente da iteração final do procedimento de mínimos quadrados ponderados iterados para ajustar o modelo, e têm a interpretação usual - exceto que, ao contrário de um modelo linear, os valores h_i em um modelos linear generalizado dependem da variável de resposta Y, bem como na configuração dos X. Com isso podemos verificar pontos extremos no espaço designado. Com isso os pontos exercem um papel importante no ajuste final dos parâmetros de um modelo estatístico, sua eliminação pode ocasionar mudanças importantes em uma análise estatística, por exemplo.

Essa matriz é dada por:

$$H = W^{\frac{1}{2}} X (X^T W X)^{-1} X^T W^{\frac{1}{2}}$$

No qual a matriz W é o peso da iteração final do processo de estimação.Tal matriz sugere a utilização dos elementos da diagonal principal de H para detectar a presença de pontos de alavanca no modelo. Neste caso a medida mais comum de alavancagem é o h_i ou *hat-values*.

O nome *hat-values* é obtido da relação entre o vetor de respostas observado e os valores ajustados. O vetor de valores ajustados é dado por $\hat{y} = X\hat{\beta} = Hy$, onde H, definida acima e chamada de matriz *hat*, projeta y, os valores observados da variável resposta Y, no subespaço estendido pelas colunas da matriz do modelo X. Como $H = H^T H$, os valores *hat* h_i são simplesmente as entradas diagonais da matriz chapéu.

Os h_i são limitados entre 0 e 1; em modelos com um intercepto, eles são limitados entre 1/n e 1 e sua soma $\sum_i h_i$ é sempre igual ao número de coeficientes no modelo, incluindo o intercepto.

Situações nas quais há alguns h_i muito grandes podem ser problemáticas: em particular, a normalidade de grandes amostras de algumas combinações lineares dos regressores tende a falhar e as observações de alta alavancagem podem exercer influência indevida sobre os resultados.

na *library* **car** temos o comando **influenceIndexPlot(ajuste)** que nos permitem ver o gráficos de índice de resíduos estudentizados, os p-valores Bonferroni correspondentes para o teste de outlier, os hat-values e as distâncias de Cook.

0.4.2 Resíduos

0.4.2.1 Ordinários

Os resíduos ordinários são simplesmente as diferenças entre a resposta observada e seu valor esperado estimado: $e_i = y_i - \hat{\mu}_i$, onde :

$$\hat{\mu} = g^{-1}(\hat{n}_i) = g^{-1}(\hat{\alpha} + \hat{\beta}_1 X_{il} + \ldots + \hat{\beta}_K X_{ik}$$

Na regressão por mínimos quadrados ponderados, a soma residual dos quadrados é igual a $\sum_i e_i^2$. Caso o modelo de regressão inclua o intercepto, então $\sum_i e_i = 0$. Os resíduos ordinários não estão correlacionados com os valores ajustados ou mesmo qualquer combinação linear dos regressores e, portanto, os padrões nos gráficos de resíduos ordinários versus combinações lineares dos regressores podem ocorrer apenas se uma ou mais suposições do modelo são inadequadas. Se o modelo de regressão estiver correto, então os resíduos ordinários são variáveis aleatórias com média 0 e com variância dada por:

$$Var(e_i) = \phi(1 - h_i)$$

A quantidade h_i é chamada de alavancagem ou hat-value. Em modelos lineares com preditores fixos, h_i é um valor não aleatório restrito a estar entre 0 e 1, dependendo da localização dos preditores para uma observação específica em relação às outras observações. Em um modelo com intercepto, o hat-value mínimo é $\frac{1}{n}$. Valores grandes de h_i correspondem a observações com valores X_i relativamente incomuns, enquanto um pequeno valor h_i corresponde a observações próximas ao centro do espaço do regressor. Resíduos comuns para observações com grande h_i têm variâncias menores. Para corrigir a variância não constante dos resíduos ordinários, podemos dividi-los por uma estimativa de seu desvio padrão. Considerando que $\hat{\phi}$ represente a estimativa de ϕ, os resíduos padronizados são:

$$e_i sd = \frac{e_i}{\hat{\phi}\sqrt{1 - h_i}}$$

0.4.2.2 Pearson

O resíduo de Pearson ajuda a classificar observações que podem ser consideradas outliers. O resíduo ordinário, definido como a diferença entre os valores observados e os valores preditos é dado por:

$$e_i = y_i - \hat{\pi}_i$$

Por não ser útil para detectar outliers, é necessário transformar esse resíduo a fim de eliminar o efeito de medição da variável resposta e preditora. Os resíduos de Pearson

fazem parte da estatística qui-quadrado de Pearson, e a indicação de um bom ajuste para o modelo ocorre quando o valores resultantes são pequenos.

$$rp_i = \frac{y_i - \hat{\pi}_i}{\hat{\pi}\sqrt{1-\hat{\pi}_i}}$$

Embora os $e_i sd$ tenham variância constante, eles não são mais não correlacionados com os valores ajustados ou combinações lineares dos regressores, portanto, usar resíduos padronizados em gráficos não é uma melhoria óbvia.

0.4.2.3 Deviance

São componentes da Deviance, utilizados para detectar os erros no ajuste do modelo. Tais resíduos medem se existem discrepâncias entre o modelo saturado e o modelo restrito em relação às observações y_i. O resultado da *deviance* é baseado no logaritmo da verossimilhança, definido por:

$$Rd_i = \begin{cases} -\sqrt{-2ln(1\hat{\pi}_i)}, \text{ se } y_i = 0 \\ + \text{ ou - } \sqrt{2\left[y_i \ln \frac{y_s}{\pi_i} + (-y_i)\ln \frac{1-y_i}{1-\pi_i}\right]}, \text{ se } 0 < y_i < 1 \\ \sqrt{-2\ln \hat{\pi}_i}, \text{se } y_i = 1 \end{cases}$$

Os resíduos no podem ser encontrados utilizando a função genérica R *residuals* e podem calcular-se vários tipos de resíduos. O padrão para um modelo linear é retornar os resíduos ordinários, mesmo se houver pesos. Definir o argumento type = "pearson", retorna os resíduos de Pearson, que produzem resíduos corretamente ponderados se houverem pesos e resíduos ordinários se não houverem pesos. Resíduos de Pearson são o padrão quando os resíduos são usados com um GLM. As funções rstandard e rstudent retornam os resíduos padronizados e estudentizados, respectivamente. A função hatvalues retorna os hat-values.

0.4.3 Outros

Temos também o **gráfico do modelo marginal** que é uma variação no gráfico de resíduos básico é o gráfico do modelo marginal, proposto por Cook and Weisberg (1997).

Além disso, temos o **teste de Wald** que avalia se cada coeficiente é significativamente diferente de zero.Neste sentido, o teste de Wald avalia se a relação uma determinada variável independente com a variável dependente é estatisticamente significativa.Há casos em que o teste de Wald costuma não rejeitar a hipótese nula quando esta deveria ser rejeitada (MESQUITA, 2014). Sendo assim, recomenda-se que o teste da razão de verossimilhança seja utilizado nos casos em que houver dúvidas acerca da eficiência do teste de Wald.

Temos também a **curva ROC** mede a capacidade de predição do modelo, sendo produzida bi-dimensionalmente através das predições de sensibilidade e especificidade. A sensibilidade indica a proporção de verdadeiros positivos e a especificidade a proporção de

verdadeiros negativos. A área abaixo da curva ROC, denominada AUC (Area Under the ROC Curve) compara os classificadores da curva em um único valor, indicando a probabilidade do modelo realizar predições corretas (FAWCETT, 2006). O valor apresentado pela AUC é sempre entre 0 e 1, e segundo Hosmer e Lemershow (2013) deve ser considerado aceitável acima de 0,7.

Critério de Informação de Akaike que é critério de informação de Akaike penaliza os modelos com mais variáveis, apresentando valores menores para modelos mais parcimoniosos, entre outros.

0.5 Aplicações Práticas

Este é um exemplo do Professor José Rodrigo de Moraes de estatística da UFF.

O professor começa falando sobre o modelo de regressão logística binária, caso particular que é a Probito:

> Um dos modelos lineares generalizados mais utilizados na área de saúde é o modelo de regressão logística binária, onde a variável resposta do modelo tem distribuição de Bernoulli (ou Binomial) e a função de ligação é a função logística. Na área de saúde, o referido modelo poderia ser adotado, por exemplo, para estimar a probabilidade do paciente: aderir ao tratamento medicamentoso (adesão=1; não adesão=0); reportar um estado de saúde não bom (não bom=1; bom=0); ter uma determinada doença crônica (ter DC=1; não ter DC=0).

0.5.0.1 base de dados

Neste caso os dados utilizados foram de um estudo sobre autoavaliação geral de saúde (1=não boa, 0=boa) com um valor de n igual a 30 indivíduos com idade variando de 20 a 95 anos. O objetivo do estudo é estudar a relação entre a autoavaliação de saúde (Y) e as seguintes variáveis explicativas: idade(em anos) e renda familiar per capita (1=Mais de 3 s.m, 0= Até 3 s.m=base). Então começa utilizando abrindo a base de dados :

```
idade=c(21,20,25,26,22,35,36,40,42,46,59,
50,60,72,85,59,29,45,39,45,20,25,36,58,95,52,80,85,62,72)
renda=c(1,1,1,1,0,0,1,1,1,1,1,0,1,1,1,1,1,0,0,0,0,0,1,0,0,0,0,0,1)
saúde=c(0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
```

0.5.0.2 Modelos de Regressão Probito e Logística

Neste caso atualizamos o caso de estudo incluindo o modelo probit para fazer a comparação entre os modelos. Vamos utilizar a função **glm** do r, a seguir:

```
modelo1=glm(saude~idade+renda,family=binomial(link="logit"));modelo1
modelo2=glm(saude~idade+renda,family=binomial(link="probit"));modelo2
```

Agora analisando os modelos com o código **summary** temos:

```
glm(formula = saude ~ idade + renda, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9396  -0.3251   0.1493   0.5154   2.1727

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.93790    1.74439  -1.684  0.09214 .
idade        0.13296    0.05123   2.595  0.00945 **
renda       -3.17898    1.45863  -2.179  0.02930 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 38.191  on 29  degrees of freedom
Residual deviance: 18.711  on 27  degrees of freedom
AIC: 24.711
Number of Fisher Scoring iterations: 6



glm(formula = saude ~ idade + renda, family = binomial(link = "probit"))

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.90361  -0.32822   0.09788   0.50326   2.11414

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.71528    0.97535  -1.759   0.0786 .
idade        0.07703    0.02669   2.886   0.0039 **
renda       -1.76117    0.76583  -2.300   0.0215 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 38.191  on 29  degrees of freedom
Residual deviance: 18.590  on 27  degrees of freedom
AIC: 24.59
Number of Fisher Scoring iterations: 8
```

Vemos que temos valores bem diferentes para todas as estimativas, porém todos os teste mostram que todas as variáveis são significativas em ambos os ajustes, percebemos também que o modelo *Probito* teve um melhor resultados no resíduos e para isso vamos usar a função **AIC()** do R BASE para ver qual modelo escolher:

```
 > AIC(modelo1,modelo2)
        df      AIC
modelo1  3 24.71078
modelo2  3 24.59033
```

Deste resultados, selecionamos a função de ligação complementar log-log (menor valor de AIC), neste caso o modelo 2 vamos continuar as análises.

Utilizamos o **pacote rms : Regression Modeling Strategies** para calcular o R^2 de Nagelkerke, para isso devemos utilizar a função lrm da seguinte forma:

```
 > lrm(modelo1)$stats[10]
       R2
0.6633273
> lrm(modelo2)$stats[10]
       R2
0.6633273
```

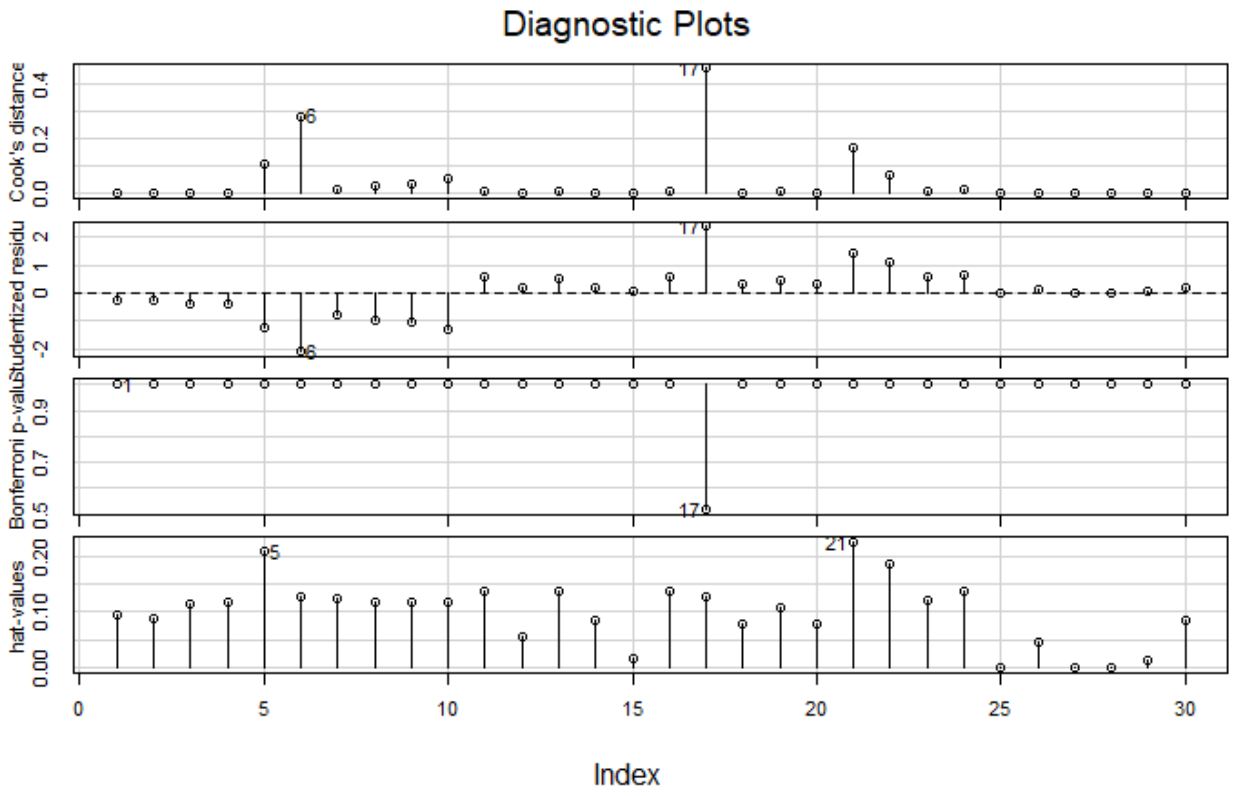
Olhando para R^2 obtemos um valor relativamente médio,porém iguais nos dois modelos, será coincidência? Obvio que não.

0.5.0.3 Diagnóstico

Vamos ver o que pode estar influenciando o modelo, para isso vamos usar a função **influenceIndexPlot(modelo2)** da biblioteca **car**:

```
influenceIndexPlot(modelo2)
```

Figura 2 – Gráfico do influência do modelo 2

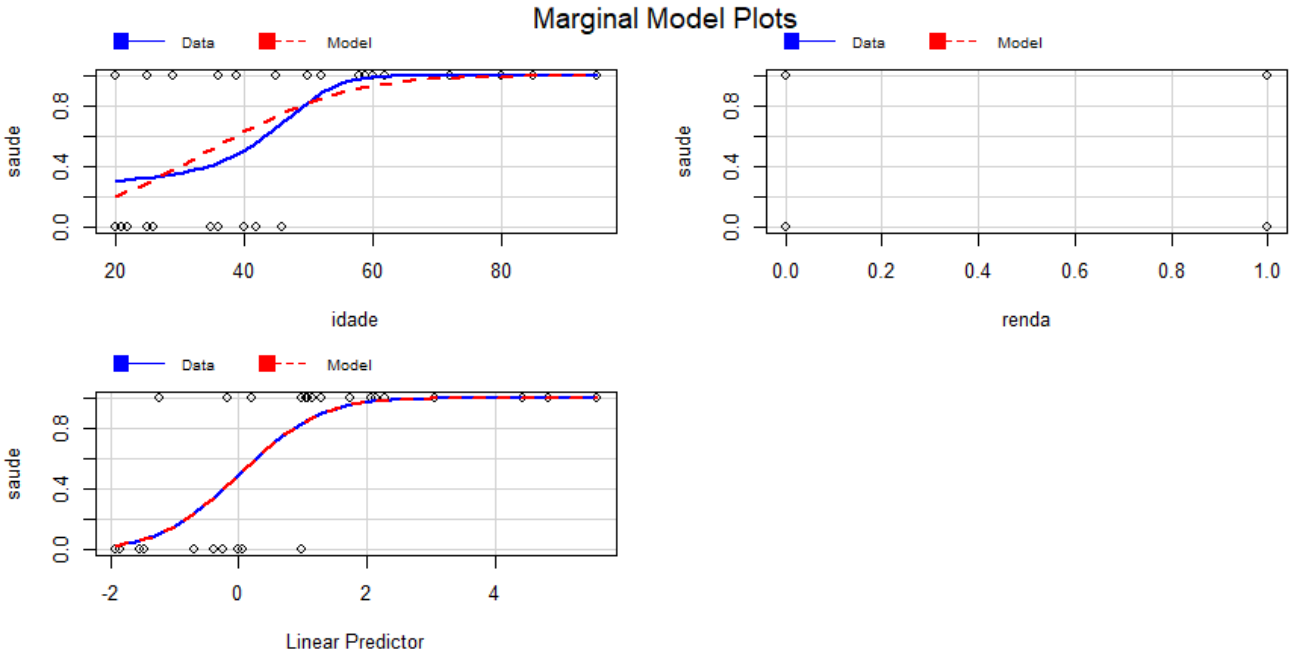


Na figura acima, que inclui gráficos de índice de resíduos estudentizados, os p-valores Bonferroni correspondentes para o teste de outlier, os hat-values e as distâncias de Cook (discutida em Medidas de influência) para a regressão de intertravamentos. Observe que nas três primeiras observações se destacam na distância de Cook , no hat-value e os resíduos estudentizados o sujeitos 6 e 17 se destacam, pois ambos são um oposto do outro, pois o individuo 6 tem uma saúde boa e não tem uma doença crônica e o sujeito 17 tem uma saúde ruim e tem uma doença crônica.

Olhando agora para ver se temos normalidade nos resíduos vamos usar a função genérica R residuals e podem calcular-se vários tipos de resíduos. O padrão para um modelo linear é retornar os resíduos ordinários, mesmo se houver pesos. Definir o argumento type = "pearson", retorna os resíduos de Pearson, que produzem resíduos corretamente ponderados se houverem pesos e resíduos ordinários se não houverem pesos. Resíduos de Pearson são o padrão quando

os resíduos são usados com um GLM. As funções rstandard e rstudent retornam os resíduos padronizados e estudentizados, respectivamente. A função hatvalues retorna os hat-values.

Figura 3 – Gráfico de resíduos do modelo 2

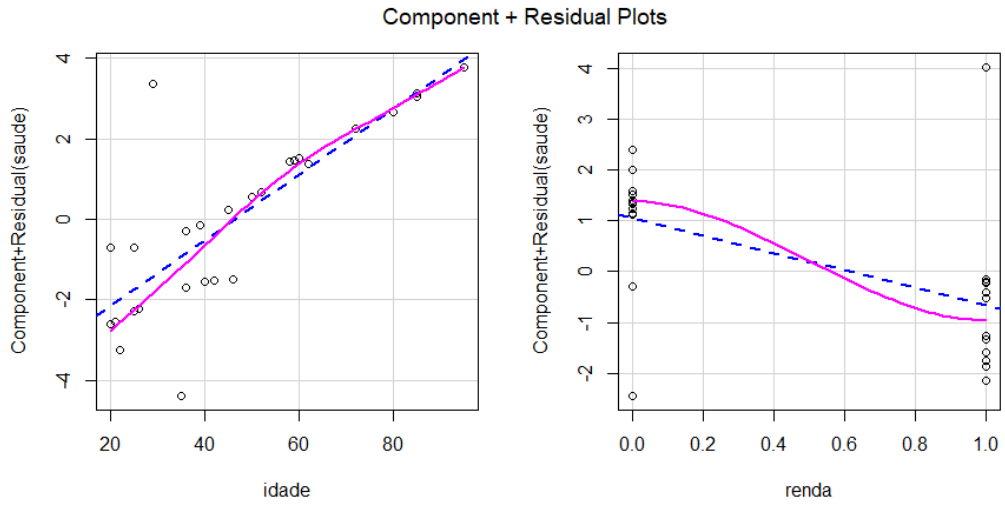


Gráficos de resíduos em relação aos valores ajustados e em relação a cada um dos preditores, por sua vez, são os gráficos de diagnóstico mais básicos. Se um modelo linear for especificado corretamente, os resíduos de Pearson são independentes dos valores ajustados e dos preditores, e esses gráficos devem ser gráficos nulos, sem características sistemáticas - no sentido de que a distribuição condicional dos resíduos, no eixo vertical do gráfico, não deve ser alterada com os valores ajustados ou com um preditor, no eixo horizontal. A presença de características sistemáticas geralmente implica uma falha de uma ou mais suposições do modelo. De interesse nesses gráficos são as tendências não lineares, as tendências de variação no gráfico e os pontos isolados.

Fazendo o Gráfico do modelo marginal que é uma variação no gráfico de resíduos básico é o gráfico do modelo marginal, proposto por Cook and Weisberg (1997):

```
library(car)
marginalModelPlots(modelo2)
```

Figura 4 – Gráfico de resíduos do modelo 2



Os gráficos da resposta versus preditores individuais exibem a distribuição condicional da resposta dado cada preditor, ignorando os outros preditores; estes são gráficos marginais no sentido de que mostram a relação marginal entre a resposta e cada preditor contínuo. O gráfico em relação aos valores ajustados é um pouco diferente, pois exibe a distribuição condicional da resposta de acordo com o ajuste do modelo. Agora imagine um segundo gráfico que substitui o eixo vertical com os valores ajustados do modelo. Se o modelo for apropriado para os dados, então, sob condições bastante suaves, o ajuste suave para este segundo gráfico também deve estimar a esperança condicional da resposta dado o preditor no eixo horizontal. A segunda suavização também é desenhada no gráfico do modelo marginal, como uma linha tracejada. Se o modelo se ajusta bem aos dados, então as duas suavizações devem corresponder em cada um dos gráficos do modelo marginal; se algum par de alisamentos não corresponder, então temos evidências de que o modelo não se ajusta bem aos dados. Uma característica interessante dos gráficos do modelo marginal é que, embora o modelo que ajustamos aos dados especifique relações parciais lineares entre idade e renda, ele é capaz de reproduzir relações marginais não lineares para esse preditor. Na verdade, o modelo, conforme representado pelas linhas tracejadas, faz um trabalho bastante bom em combinar as relações marginais representadas pelas linhas sólidas, embora as falhas sistemáticas descobertas nos gráficos de resíduos sejam visíveis aqui.

0.5.0.4 Conclusões da análise

Fazendo agora um intervalo de confiança de 95% onfiança para os parâmetros do modelo, com base na estatística de Wald ,temos :

```
> OR1=exp(modelo2$coefficients)
> ICbeta1=confint.default(modelo2,level=0.95)
> ICOR1=exp(ICbeta1)
> round((cbind(OR1, ICOR1)),3)
              OR1 2.5 % 97.5 %
(Intercept) 0.180 0.027  1.217
idade       1.080 1.025  1.138
renda       0.172 0.038  0.771
```

Pela medidas de associação (razões de chance) pode-se demonstrar matematicamente que a razão de chance é o exponencial da estimativa pontual e fazendo os intervalos de confiança para as razões de chance (odds ratio – OR), fixando o nível de confiança de 95%.Através do comando executados acima os resultados de interesse são condensados, de modo a facilitar a interpretação das medidas de razão de chance e a análise sobre a significância da associação entre cada variável explicativa e a chance do indivíduo reportar um estado de saúde não bom. Tanto a idade quanto a renda familiar per capita estão significativamente relacionadas com a chance de autoavaliação de saúde não boa (OBS: Note que o p-valor é menor que o nível de significância de 5% e o IC para OR não inclui a unidade). A chance do indivíduo reportar um estado de saúde não bom aumenta em 8% ao aumentar em 1 ano a idade. Indivíduos com mais de 3 salários mínimos tem uma chance de reportar um estado de saúde não bom 82,8% menor do que os indivíduos que ganham no máximo 3 salários mínimos.

0.6 Considerações Finais e Conclusão

A regressão logística é usada no aprendizado de máquina (ML) para ajudar a criar previsões precisas. É semelhante a regressão linear, exceto que, em vez de um resultado gráfico, a variável de destino é binária; o valor é 1 ou 0.Existem dois tipos de mensuráveis, as variáveis características explicativas (item sendo medido) e a variável de resposta variável binária alvo, que é o resultado.Os modelos logit e probit resolvem ambos os problemas: os valores (representando probabilidades) estarão sempre entre (0,1) e o efeito parcial mudará dependendo dos parâmetros.

Referências

McCullagh, P. and Nelder, J.A. (1989). Generalized Linear Models, Second Edition. London: Chapman and Hall.

Clarice,G.B. Demétrio.; Gauss,Moutinho Cordeiro. Modelos Lineares Generalizados e Extensões - 1 Edição, ESALQ/USP, 2013.

Gilberto,A. Paula. MODELOS DE REGRESSÃO com apoio computacional - 2 Edição, ESALQ/USP, 2013.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Oxford English Dictionary, 3rd ed. s.v. probit (article dated June 2007): Bliss, C. I. (1934). "The Method of Probits". Science. 79 (2037): 38–39. Bibcode:1934Sci....79...38B. doi:10.1126/science.79.2037.38. PMID 17813446. These arbitrary probability units have been called 'probits'.

Probit model,https://en.wikipedia.org/wiki/Probit_model;acesso 18/11/2022.

SOUZA, É. C. d. Análise de influência local no modelo de regressão logística. Tese (Doutorado) — Universidade de São Paulo, 2006. Citado 3 vezes nas páginas 12, 18 e 20.