UNIVERSIDADE DE ÉVORA
ESCOLA DE CIÊNCIAS E TECNOLOGIA

Escola de Ciências e Tecnologia da Universidade de Évora

Mestrado em Modelação Estatística e Análise de Dados

Ano Letivo 2017/2018

2º Semestre

U.C.: Análise Categórica de Dados

Regressão Logística e Multinomial

Docente:

Dulce Pereira

Discente:

Ana Sapata n.º39504

1.

a) Para se saber se existe associação entre duas variáveis usualmente utiliza-se o teste do Qui-Quadrado para a independência. Neste caso, as hipóteses para este teste são:

$$\begin{cases} H_0: O\ incumprimento\ é\ independente\ da\ idade\ categorizada \\ H_1: O\ incumprimento\ não\ é\ independente\ da\ idade\ categorizada \end{cases}$$

No entanto para a realização deste teste é preciso que se verifiquem dois pressupostos:

1. Não mais de 20% com $E_{ij} < 5$;
2. Todas as classes com $E_{ij} > 1$.

Para as variáveis idadeCat e Incumprimento obteve-se a Tabela1 com os valores esperados

		Incumprimento	
		0(Não)	1(Sim)
idadeCat	0(<35)	195.146	67.854
	1(>=35)	175.854	61.146

Tabela 1 Valores Esperados obtidos pelo Qui-Quadrado

Pelo que se verificam os pressupostos para a realização do teste do Qui-Quadrado para a independência. Através da realização deste teste obtiveram-se os seguintes resultados

$$\chi^2 = 5.6832;\ p - value = 0.01713$$

Utilizando por omissão um nível de significância de 5% tem-se que $p - value = 0.01713 < 0.05 = \alpha$, pelo que se rejeita H_0. Sendo assim, para um nível de significância de 5% não existe evidencia estatística para afirmar que as variáveis incumprimento e a idade categorizada são independentes, pelo que, estas variáveis são dependentes.

Como as variáveis são dependentes de seguida ter-se-á de analisar qual o seu nível de associação, como ambas as variáveis são nominais, podem usar-se medidas baseadas no Qui-Quadrado ou baseadas na redução proporcional do erro de previsão, para medir esta associação.

Foram então utilizadas as medidas baseadas no Qui-Quadrado, ou seja, o Coeficiente de Contingência, o Coeficiente V de Cramer e o Coeficiente Phi. Os resultados obtidos foram os seguintes(Tabela2):

Coeficiente de Contingência	0.111
Coeficiente V de Cramer	0.111
Coeficiente Phi	0.111

Tabela 2Medidas de Associação

Através da Tabela2 conclui-se que existe uma associação fraca entre o incumprimento e a idade categorizada.

Para o Risco Relativo (RR) foi obtido o valor de 0.8771742. Ou seja, o risco de pessoas com 35 anos de idade ou mais sofrerem de incumprimento é 0.87% vezes superior ao risco das pessoas com idade inferior a 35 anos.

O valor obtido para o Rácio das Chances (OR) foi de 0.596210106. O que significa que as pessoas com idade igual ou superior a 35 anos têm 0.60% mais possibilidade de estarem em incumprimento com o banco.

b) Para se realizar o teste de Breslow-Day foi necessário passar a variável educação de 5 categorias para 4, tendo-se juntado a categoria 4 e 5.

Realizou-se o teste de Mantel-Haenszel cujas suas hipóteses são

$$\begin{cases} H_0: O\ incumprimento\ \text{é}\ independente\ da\ idade\ categorizada\ nos\ diferentes\ niveis\ de\ escolariadade \Leftrightarrow\ OR = 1 \\ H_1: O\ incumprimento\ \text{não é}\ independente\ da\ idade\ categorizada\ nos\ diferentes\ niveis\ de\ escolaridade \Leftrightarrow OR \neq 1 \end{cases}$$

Para tal teste obtiveram-se os seguintes resultados

$$\chi^2 = 6.222; p - value = 0.01262$$

Como se tem que $p - value = 0.01262 < 0.05 = \alpha$, rejeita-se H_0 para um nível de significância de 5%. Portanto, com um nível de significância de 5% não existe evidencia estatística para afirmar que o incumprimento é independente da idade categorizada nos diferentes níveis de escolaridade.

Foi também realizado o teste de Breslow-Day, com as seguintes hipóteses

$$\begin{cases} H_0: Os\ OR\ entre\ o\ incumprimento\ e\ o\ facto\ de\ a\ pessoa\ ter\ idade\ superior\ ou\ igual\ a\ 35 \\ anos/ou\ não\ são\ identicos\ \ nos\ diferenets\ niveis\ \ de\ escolaridade \\ \Leftrightarrow OR\ homogeneos\ nos\ diferentes\ niveis\ de\ escolaridade \\ \\ H_0: Os\ OR\ entre\ o\ incumprimento\ e\ o\ facto\ de\ a\ pessoa\ ter\ idade\ superior\ ou\ igual\ a\ 35 \\ anos/ou\ não, não\ são\ identicos\ nos\ diferentes\ niveis\ de\ escolaridade \\ \Leftrightarrow OR\ não\ homogeneos\ nos\ diferentes\ niveis\ de\ escolaridade \end{cases}$$

Para este teste foram obtidos os seguintes resultados:

$$\chi^2 = 3.3389; p - value = 0.3423$$

Como se tem que $p - value = 0.3423 > 0.05 = \alpha$ então para 5% de significância não se rejeita H_0. Conclui-se então que para um nível de significância de 5% existe evidencia significativa para afirma que os OR são homogéneos nos diferentes níveis de escolaridade, pelo que não será necessário estratificar pelo nível de escolaridade.

c) Como a variável idadeCat é qualitativa ordinal, mas a variável dívida é uma variável quantitativa para se medir a correlação entre estas duas variáveis utilizou-se uma medida de correlação que é utilizada para variáveis ordinais ou superiores. Sendo assim para a medição da correlação foi utilizado o Tau de Kendall tendo-se obtido um valor de 0.00298, pelo que se considera que as variáveis são independentes não havendo correlação entre estas.

Regressão Logística

d) Foram realizados vários modelos univariados de modo a saber quais as variáveis que são significativas para a elaboração de uma boa previsão de risco de incumprimento.

- Incumprimento ~ idade

$$logit(\pi) = -0.05631 - 0.02918\ idade$$

Como $\beta_{idade} = -0.02918$, tem-se que $OR = e^{-0.02918} = 0.9712393$. Logo $(OR - 1) * 100\% = -2.876071$, o que significa que as chances de a pessoa estar em incumprimento com o banco diminuem 2.88% com o aumentar 1 ano na idade.

O intervalo de confiança (IC) para β_{idade} é $]0.9459198, 0.9972365[$, como o valor 1 não se encontra no intervalo significa que esta variável é significativa para o modelo.

Realizou-se ainda um teste para a significância da variável através da razão de verosimilhanças, com as seguintes hipóteses

$$\begin{cases} H_0: \beta_{idade} = 0 \\ H_1: \beta_{idade} \neq 0 \end{cases}$$

Para o qual se obteve um $p - value = 0.02814 < 0.05 = \alpha$, rejeitando-se assim H_0 com 5% de significância. Portanto, não existe evidencia estatística, para um nível de significância de 5%, de que $\beta_{idade} = 0$, ou seja, a variável é idade significativa.

- Incumprimento ~ educação

$$logit(\pi) = -1.5925 + 0.3019\ educação$$

Como $\beta_{educação} = 0.3019$, tem-se que $OR = e^{0.3019} = 1.352418$. Logo $(OR - 1) * 100\% = 35.2418\%$, o que significa que as chances de a pessoa estar em incumprimento com o banco aumentam 35.24% com o aumentar 1 nível de escolaridade.

O intervalo de confiança (IC) para $\beta_{educação}$ é]1.101749,1.660120[, como o valor 1 não se encontra no intervalo significa que esta variável é significativa para o modelo.

Realizou-se ainda um teste para a significância da variável através da razão de verosimilhanças, com as seguintes hipóteses

$$\begin{cases} H_0: \beta_{educação} = 0 \\ H_1: \beta_{educação} \neq 0 \end{cases}$$

Para o qual se obteve um $p - value = 0.0039 < 0.05 = \alpha$, rejeitando-se assim H_0 com 5% de significância. Portanto, não existe evidencia estatística, para um nível de significância de 5%, de que $\beta_{educação} = 0$, ou seja, a variável educação é significativa.

- Incumprimento ~ t_emprego

$$logit(\pi) = -0.25370 - 0.11138 t_{emprego}$$

Como $\beta_{t\_emprego} = -0.11138$, tem-se que $OR = e^{-0.11138} = 0.8946013$. Logo $(OR - 1) * 100\% = -10.53987$, o que significa que as chances de a pessoa estar em incumprimento com o banco diminuem -10.53987% com o aumentar 1 valor em t_emprego.

O intervalo de confiança (IC) para $\beta_{t\_emprego}$ é]0.8602589,0.9303147[, como o valor 1 não se encontra no intervalo significa que esta variável é significativa para o modelo.

Realizou-se ainda um teste para a significância da variável através da razão de verosimilhanças, com as seguintes hipóteses

$$\begin{cases} H_0: \beta_{t\_emprego} = 0 \\ H_1: \beta_{t\_emprego} \neq 0 \end{cases}$$

Para o qual se obteve um $p - value = 9.486e^{-10} < 0.05 = \alpha$, rejeitando-se assim H_0 com 5% de significância. Portanto, não existe evidencia estatística, para um nível de significância de 5%, de que $\beta_{t\_emprego} = 0$, ou seja, a variável t_emprego é significativa.

- Incumprimento ~ t_endereço

$$logit(\pi) = -0.69386 - 0.04705 t_{endereço}$$

Como $\beta_{t_{endereço}} = -0.04705$, tem-se que $OR = e^{-0.04705} = 0.9540374$. Logo $(OR - 1) * 100\% = -4.59626$, o que significa que as chances de a pessoa estar em incumprimento com o banco diminuem -4.59% com o aumentar 1 valor em t_endereço.

O intervalo de confiança (IC) para $\beta_{t\_endereço}$ é]0.9229821,0.9861375[, como o valor 1 não se encontra no intervalo significa que esta variável é significativa para o modelo.

Realizou-se ainda um teste para a significância da variável através da razão de verosimilhanças, com as seguintes hipóteses

$$\begin{cases} H_0: \beta_{t\_endereço} = 0 \\ H_1: \beta_{t\_endereço} \neq 0 \end{cases}$$

Para o qual se obteve um $p - value = 0.003799 < 0.05 = \alpha$, rejeitando-se assim H_0 com 5% de significância. Portanto, não existe evidencia estatística, para um nível de significância de 5%, de que $\beta_{t\_endereço} = 0$, ou seja, a variável t_endereço é significativa.

- Incumprimento ~ rendimento

$$logit(\pi) = -1.0213515 - 0.0007766 rendimento$$

Como $\beta_{rendimento} = -0.0007766$, tem-se que $OR = e^{-0.0007766} = 0.9992237$. Logo $(OR - 1) * 100\% = -0.07763138$, o que significa que as chances de a pessoa estar em incumprimento com o banco diminuem 0.08% com o aumentar 1 valor em rendimento.

O intervalo de confiança (IC) para $\beta_{rendimento}$ é]0.9936486,1.0048300[, como o valor 1 pertence ao intervalo significa que esta variável não é significativa para o modelo.

Realizou-se ainda um teste para a significância da variável através da razão de verosimilhanças, com as seguintes hipóteses

$$\begin{cases} H_0: \beta_{rendimento} = 0 \\ H_1: \beta_{rendimento} \neq 0 \end{cases}$$

Para o qual se obteve um $p - value = 0.7832 > 0.05 = \alpha$, não se rejeitando assim H_0 com 5% de significância. Portanto, existe evidencia estatística, para um nível de significância de 5%, de que $\beta_{rendimento} = 0$, ou seja, a variável rendimento não é significativa.

- Incumprimento ~ dívida

$$logit(\pi) = -2.50414 + 0.12776 dívida$$

Como $\beta_{dívida} = 0.12776$, tem-se que $OR = e^{0.12776} = 1.136275$. Logo $(OR - 1) * 100\% = 13.62746$, o que significa que as chances de a pessoa estar em incumprimento com o banco aumentam 13.63% com o aumentar 1 valor na dívida.

O intervalo de confiança (IC) para $\beta_{dívida}$ é]1.099764,1.173998[, como o valor 1 não se encontra no intervalo significa que esta variável é significativa para o modelo.

Realizou-se ainda um teste para a significância da variável através da razão de verosimilhanças, com as seguintes hipóteses

$$\begin{cases} H_0: \beta_{dívida} = 0 \\ H_1: \beta_{dívida} \neq 0 \end{cases}$$

Para o qual se obteve um $p - value < 2.2e^{-16} < 0.05 = \alpha$, rejeitando-se assim H_0 com 5% de significância. Portanto, não existe evidencia estatística, para um nível de significância de 5%, de que $\beta_{dívida} = 0$, ou seja, a variável dívida é significativa.

- Incumprimento ~ dívida_cc

$$logit(\pi) = -1.54737 + 0.29432 dívida\_cc$$

Como $\beta_{dívida\_cc} = 0.29432$, tem-se que $OR = e^{0.29432} = 1.34222$. Logo $(OR - 1) * 100\% = 34.22198$, o que significa que as chances de a pessoa estar em incumprimento com o banco aumentam 34.22% com o aumentar 1 valor na dívida_cc.

O intervalo de confiança (IC) para $\beta_{dívida\_cc}$ é]1.196263,1.505984[, como o valor 1 não se encontra no intervalo significa que esta variável é significativa para o modelo.

Realizou-se ainda um teste para a significância da variável através da razão de verosimilhanças, com as seguintes hipóteses

$$\begin{cases} H_0: \beta_{dívida\_cc} = 0 \\ H_1: \beta_{dívida\_cc} \neq 0 \end{cases}$$

Para o qual se obteve um $p - value = 3.222e^{-09} < 0.05 = \alpha$, rejeitando-se assim H_0 com 5% de significância. Portanto, não existe evidencia estatística, para um nível de significância de 5%, de que $\beta_{dívida\_cc} = 0$, ou seja, a variável dívida_cc é significativa.

• Incumprimento ~ outras_dív

$$logit(\pi) = -1.44216 + 0.11872 outras\_div$$

Como $\beta_{outras\_div} = +0.11872$, tem-se que $OR = e^{0.11872} = 1.126054$. Logo $(OR - 1) * 100\% = 12.6054$, o que significa que as chances de a pessoa estar em incumprimento com o banco aumentam 12.61% com o aumentar 1 valor em outras_dív.

O intervalo de confiança (IC) para β_{outras_div} é]1.061831,1.194162[, como o valor 1 não se encontra no intervalo significa que esta variável é significativa para o modelo.

Realizou-se ainda um teste para a significância da variável através da razão de verosimilhanças, com as seguintes hipóteses

$$\begin{cases} H_0: \beta_{outras\_div} = 0 \\ H_1: \beta_{outras\_div} \neq 0 \end{cases}$$

Para o qual se obteve um $p - value = 4.384e^{-05} < 0.05 = \alpha$, rejeitando-se assim H_0 com 5% de significância. Portanto, não existe evidencia estatística, para um nível de significância de 5%, de que $\beta_{outras\_div} = 0$, ou seja, a variável outras_dív é significativa.

• Incumprimento ~ idadeCat

$$logit(\pi) = -0.8275 - 0.5172 idadeCat$$

Como $\beta_{idadeCat} = -0.5172$, tem-se que $OR = e^{-0.5712} = 0.5962101$. Logo $(OR - 1) * 100\% = -40.37899$, o que significa que as chances de a pessoa estar em incumprimento com o banco diminuem 40.38% com o aumentar 1 ano na idadeCat.

O intervalo de confiança (IC) para $\beta_{idadeCat}$ é $]0.3958005, 0.8980952[$, como o valor 1 não se encontra no intervalo significa que esta variável é significativa para o modelo.

Realizou-se ainda um teste para a significância da variável através da razão de verosimilhanças, com as seguintes hipóteses

$$\begin{cases} H_0: \beta_{idadeCat} = 0 \\ H_1: \beta_{idadeCat} \neq 0 \end{cases}$$

Para o qual se obteve um $p - value = 0.01251 < 0.05 = \alpha$, rejeitando-se assim H_0 com 5% de significância. Portanto, não existe evidencia estatística, para um nível de significância de 5%, de que $\beta_{idadeCat} = 0$, ou seja, a variável idadeCat é significativa.

- Incumprimento ~ educaçãoCat

$$logit(\pi) = -1.6225 + 0.3201 educaçãoCat$$

Como $\beta_{educaçãoCat} = 0.3201$, tem-se que $OR = e^{0.3201} = 1.377239$. Logo $(OR - 1) * 100\% = 37.72388$, o que significa que as chances de a pessoa estar em incumprimento com o banco aumentam 37.72% com o aumentar 1 valor em educaçãoCat.

O intervalo de confiança (IC) para $\beta_{educaçãoCat}$ é $]1.115042, 1.701089[$, como o valor 1 não se encontra no intervalo significa que esta variável é significativa para o modelo.

Realizou-se ainda um teste para a significância da variável através da razão de verosimilhanças, com as seguintes hipóteses

$$\begin{cases} H_0: \beta_{educaçãoCat} = 0 \\ H_1: \beta_{educaçãoCat} \neq 0 \end{cases}$$

Para o qual se obteve um $p - value = 0.003247 < 0.05 = \alpha$, rejeitando-se assim H_0 com 5% de significância. Portanto, não existe evidencia estatística, para um nível de significância de 5%, de que $\beta_{educaçãoCat} = 0$, ou seja, a variável educaçãoCat é significativa.

Portanto as variáveis significativas para se elaborar uma boa previsão de risco de incumprimento são a idade, educação, t_emprego, t_endereço, dívida, dívida_cc, outras_dív, idadeCat e educaçãoCat.

e) Depois de efetuada a análise univariada do modelo efetuou-se uma análise multivariada chegando-se ao seguinte modelo final

$$logit(\pi) = -0.86791 - 0.22954\ t\_emprego - 0.07435\ t\_endereço + 0.08242\ dívida + 0.57561\ dívida\_cc$$

f) Começou por se usar um modelo constituído pelas variáveis consideradas significas em 1.d), retirando-se de seguida uma a uma as que tinham maior $p - value$ e atualizando a cada retirada também o modelo com as já retiradas anteriormente para se verificar que realmente não eram necessárias no modelo. Após se ter chegado à conclusão que o modelo seria da forma $logit(\pi) = \beta_o + \beta_1 t\_emprego + \beta_2\ t\_endereço + \beta_3 dívida + \beta_4 dívida\ \_cc$, testou-se ainda as interações entre estas variáveis no modelo, dando sempre que não eram significativas.

Através do modelo obtido definido em 1.e) podem retirar-se as seguintes conclusões:

- t_emprego

  $$OR(t\_emprego = 1, t\_emprego = 0) = e^{-0.22954}$$

  Logo tem-se que $(OR - 1) * 100 = -20{,}51$, o que significa que quando se passa de uma pessoa que está a trabalhar à menos de 1 ano para uma que está a trabalhar à 1ano as chances de a pessoa sofrer incumprimento diminuem 20,51%. Portanto, quanto mais tempo a pessoa estiver no emprego as chances de cumprir incumprimento vão diminuir mais.

- t_endereço

  $$OR(t\_endereço = 1, t\_endereço = 0) = e^{-0.07435}$$

  Logo tem-se que $(OR - 1) * 100 = -7{,}17$, o que significa que quando se passa de uma pessoa que está à menos de 1 ano na mesma casa para uma que está à 1ano na mesma casa as chances de a pessoa sofrer incumprimento diminuem 7,17%. Portanto, quanto mais tempo a pessoa estiver na mesma casa as chances de cumprir incumprimento vão diminuir mais.

- dívida

  $$OR(dívida = 1, dívida = 0) = e^{0.08242}$$

  Logo tem-se que $(OR - 1) * 100 = 8{,}59$, pelo que quando se passa de uma pessoa que tem dívidas inferiores a 1 milhar para uma pessoa que tem dívidas na ordem de 1 milhar as chances de incumprimento dessa pessoa aumentam 8,59%. Sendo assim quando maior é a dívida da pessoa maior é a chance desta vir a estar em incumprimento.

- dívida_cc

$$OR(dívida\_cc = 1, dívida\_cc = 0) = e^{0.57561}$$

Logo tem-se que $(OR - 1) * 100 = 77{,}82$, pelo que quando se passa de uma pessoa em que as dívidas do cartão de crédito são inferiores a 1 milhar para uma pessoa em que as dívidas do cartão de crédito são na ordem de 1 milhar as chances de incumprimento dessa pessoa aumentam 77,82%. Sendo assim quando maior é a dívida do cartão de crédito da pessoa maior é a chance desta vir a estar em incumprimento.

g) Para a adequabilidade do modelo foram obtidos os seguintes valores

Medida	Valor
R^2 *McFadden*	0.290796
R^2 *Cox and Snell*	0.282558
R^2 *Nagelkerke*	0.415043

Tabela 3 Valores da adequabilidade do modelo

Segundo os valores da Tabela 3 o modelo não é muito adequado aos dados que temos, isto poderá dever-se ao facto de existirem outliers como se verá e seguida.

Para o teste de Hosmer e Lemeshow têm-se as seguintes hipóteses

$$\begin{cases} H_0: O\ modelo\ ajusta - se\ aos\ dados \\ H_1: O\ modelo\ não\ se\ ajusta\ aos\ dados \end{cases}$$

Obtendo-se então um p-value=0.409 que é superior a um nível de significância de 5%, pelo que não se rejeita H_0. Portanto, com 5% de significância, existe evidencia estatística para afirmar que o modelo se ajusta aos dados. Embora o ajustamento não seja o melhor, como se viu pelos valores dos R^2 anterior.

Na análise de resíduos foram obtidos os seguintes gráficos (Fig.1-2), através dos quais se consegue visualizar a existência de alguns outliers. Como é o caso das observações 36, 53, 152, 281, 301, 344, 374, 445, 467 e 479.

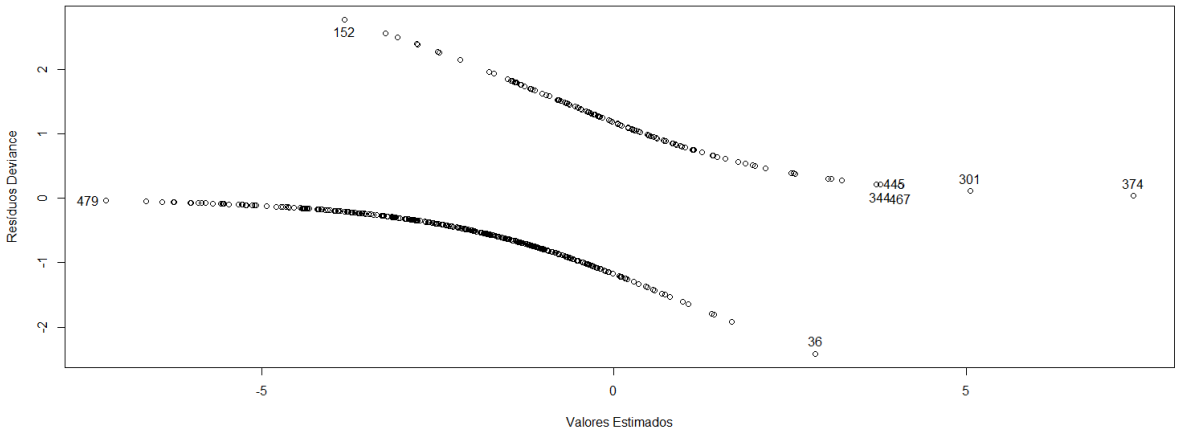

Fig. 1 Resíduos Deviance

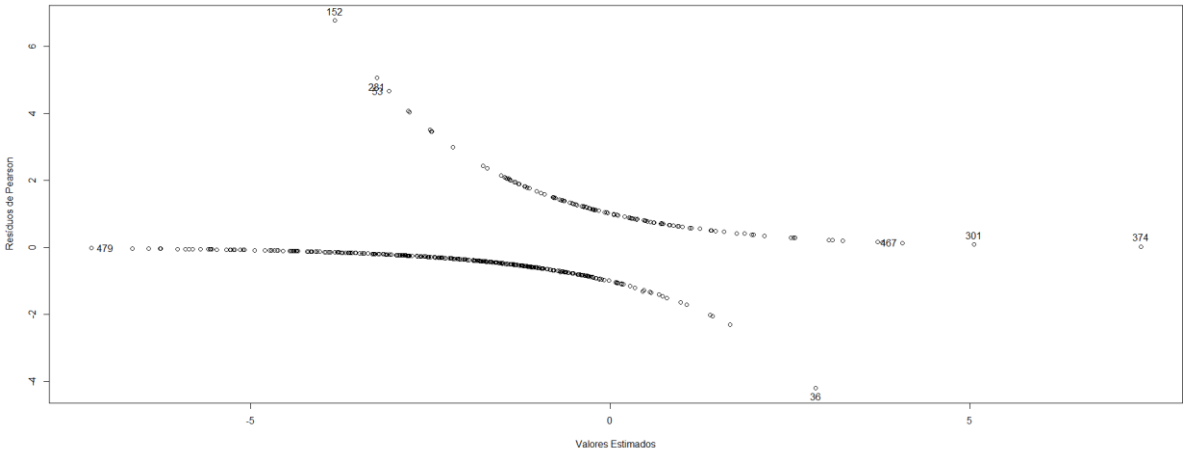

Fig. 2 Resíduos Standardizados

Os gráficos das Fig.3-5 representam a distância de Cook para os resíduos do modelo, observando-se alguns outliers como é o caso das observações 36, 152 e 281, já constatadas anteriormente.

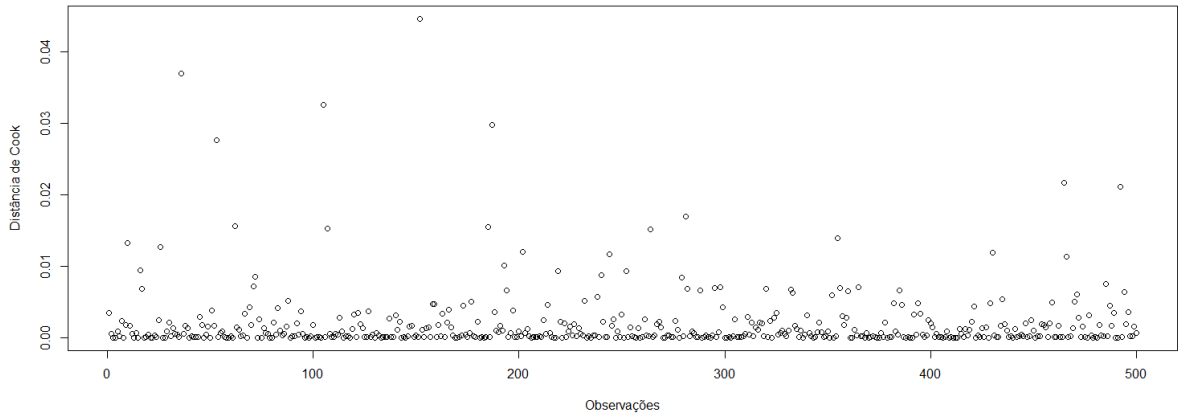

Fig. 3 Distância de Cook

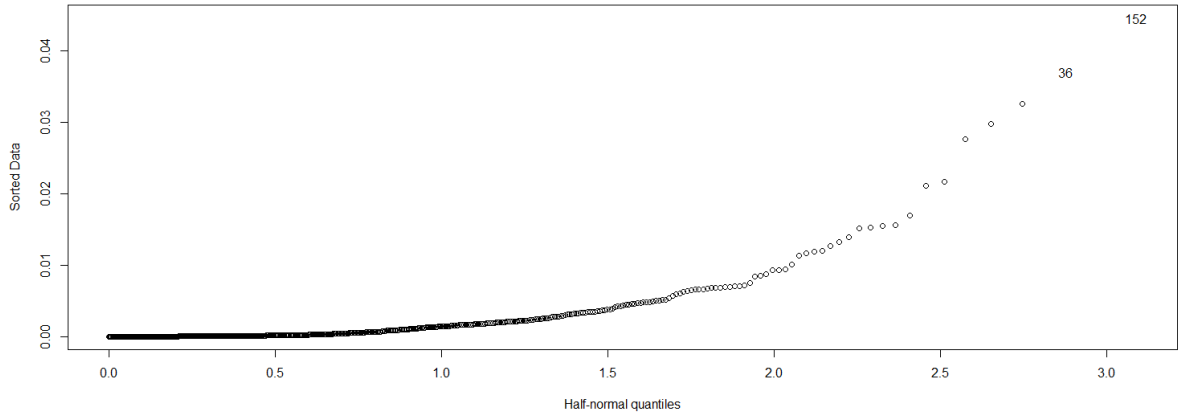

Fig. 4 HalfNormal para distância de Cook

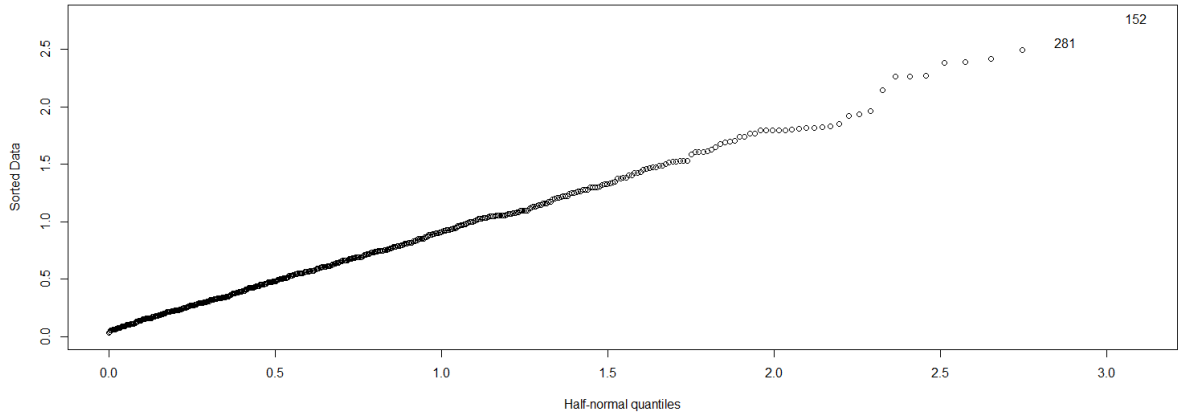

Fig. 5 HalfNormal para os resíduos

h) A curva de ROC adquirida para o modelo foi a representada pela Fig.6, podendo-se ver que o ponto de corte iria ter uma sensibilidade de 90.7%, ou seja, neste ponto o modelo classifica corretamente 90.7% das pessoas que estão em incumprimento. O valor da especificidade seria de 64.4%, o que significa que o modelo classifica corretamente 64.4% das pessoas que não estão em incumprimento. Neste ponto a percentagem de falsos positivos seria de 4.8%, ou seja, 4.8% das pessoas não estão em incumprimento, mas o modelo prevê que sim. Já a percentagem de falsos negativos seria de 53.0%, ou seja, 53.0% das pessoas estão em incumprimento, mas o modelo prevê que não estão.

A área abaixo da curva ROC é de 0.848, o que é um valor bastante bom, indicando que o modelo está adequado.

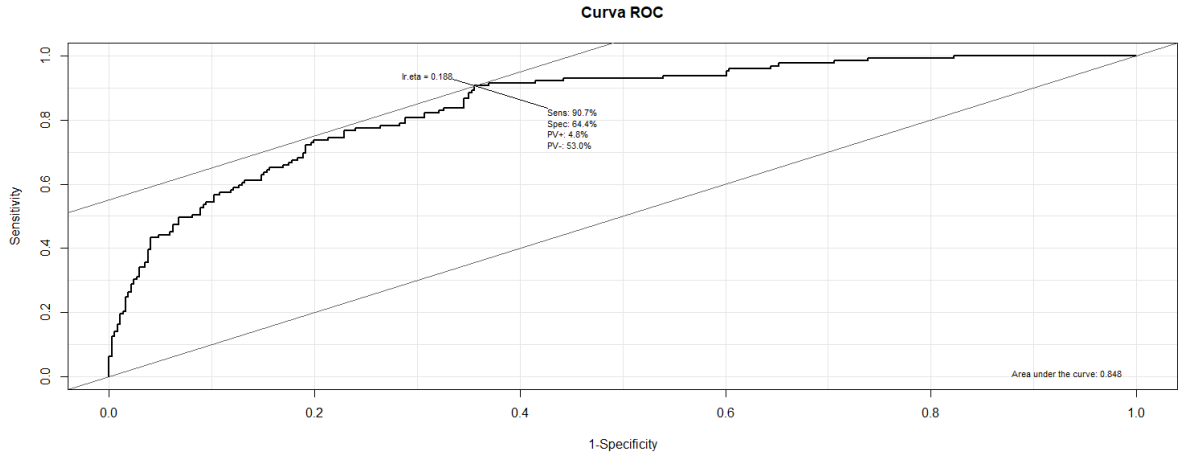

Fig. 6 Curva ROC

i) Se quisermos evitar falsos positivos, ou seja, saber realmente quais as pessoas que vão estar em incumprimento e não dizer que outra vão estar, quando realmente não vão estar, devemos maximizar a especificidade. Ao utilizarmos um ponto de corte de 0,44 obtemos uma especificidade de 0,90 e uma sensibilidade de 0,54 pelo que este será considerado um bom ponto de corte para se utilizar de modo a evitar falsos positivos.

Se quisermos evitar falsos negativos deve tentar-se maximizar-se a sensibilidade. Se for utilizado um ponto de corte de 0,142 é obtida uma sensibilidade de 0,93 e uma especificidade de 0,55, pelo que este é o ponto de corte ideal para se evitarem falsos negativos.

j)A probabilidade de incumprimento de uma pessoa com as seguintes caraterísticas

- Idade = 40 anos

- Nível de educação = 3
- Emprego atual = 3 anos
- Endereço atual = 5 anos
- Rendimento familiar anual (em milhares) = 60
- Endividamento = 17%
- Dívida do cartão de crédito (em milhares) = 70
- Outras dívidas (em milhares) = 3.

É de 1, pelo que certamente ela irá estar em incumprimento com o banco.

2.

a) Foram realizados vários modelos univariados de modo a saber quais as variáveis que são significativas para o modelo.

- y ~ dif_ano

Realizou-se um teste para a significância da variável através da razão de verosimilhanças, com as seguintes hipóteses

$$\begin{cases} H_0: \beta_{dif\_ano} = 0 \\ H_1: \beta_{dif\_ano} \neq 0 \end{cases}$$

Para o qual se obteve um $p - value \approx 0 < 0.05 = \alpha$, rejeitando-se assim H_0 com 5% de significância. Portanto, não existe evidencia estatística, para um nível de significância de 5%, de que $\beta_{dif\_ano} = 0$, ou seja, a variável dif_ano é significativa.

- y ~ sexo

Realizou-se um teste para a significância da variável através da razão de verosimilhanças, com as seguintes hipóteses

$$\begin{cases} H_0: \beta_{sexo} = 0 \\ H_1: \beta_{sexo} \neq 0 \end{cases}$$

Para o qual se obteve um $p - value = 0.04 < 0.05 = \alpha$, rejeitando-se assim H_0 com 5% de significância. Portanto, não existe evidencia estatística, para um nível de significância de 5%, de que $\beta_{sexo} = 0$, ou seja, a variável sexo é significativa.

- y ~ classesocial

Realizou-se um teste para a significância da variável através da razão de verosimilhanças, com as seguintes hipóteses

$$\begin{cases} H_0: \beta_{classesocial} = 0 \\ H_1: \beta_{classesocial} \neq 0 \end{cases}$$

Para o qual se obteve um $p - value \approx 4.800555e^{-05} < 0.05 = \alpha$, rejeitando-se assim H_0 com 5% de significância. Portanto, não existe evidencia estatística, para um nível de significância de 5%, de que $\beta_{classesocial} = 0$, ou seja, a variável classesocial é significativa.

Obteve-se por fim os seguintes resultados mostrados na Fig.7

```
> fit5 <- multinom(y ~ 1+ dif_ano + sexo + classesocial)
# weights:  18 (10 variable)
initial  value 97.776494
iter  10 value 78.019264
final  value 77.946381
converged
> summary(fit5)
Call:
multinom(formula = y ~ 1 + dif_ano + sexo + classesocial)

Coefficients:
  (Intercept)   dif_ano       sexo classesocialB classesocialC
1  -0.7211727 0.4218845 -0.1306374     0.1065587 -0.0007898174
2  -3.1140139 0.6724134  1.3822518     0.4276977  2.5037291890

Std. Errors:
  (Intercept)   dif_ano      sexo classesocialB classesocialC
1   0.5316219 0.2226198 0.5594323     0.6321577     0.8929202
2   0.8532045 0.2616009 0.6777924     0.8011360     0.8876504

Residual Deviance: 155.8928
AIC: 175.8928
> anova(fit5, fit0, test = "Chisq")
                             Model Resid. df Resid. Dev   Test    Df LR stat.      Pr(Chi)
1                                1       176   195.5304           NA       NA           NA
2 1 + dif_ano + sexo + classesocial       168   155.8928 1 vs 2     8 39.63767 3.741958e-06
```

Fig. 7 Modelo Multinomial com constante

b) Para o modelo sem constante os resultados obtidos foram exatamente os mesmos que na Fig.7.

Em ambos os modelos tem-se que:

- O risco relativo para o aumento de uma unidade na variável dif_ano é 1.5248 dos que não desejam trocar de carro vs os que gostariam de trocar de carro mas pagariam a pronto;

- O risco relativo para o aumento de uma unidade na variável dif_ano é 1.9590 dos que não desejam trocar de carro vs os que gostariam de trocar de carro mas financiariam o pagamento;
- O risco relativo para os homens que não desejam trocar de carro vs os que gostariam de trocar de carro mas pagariam a pronto é de 0.8775;
- O risco relativo para os homens que não desejam trocar de carro vs os que gostariam de trocar de carro mas financiariam o pagamento é de 3.9837.

Através do modelo tem-se também que

$$\ln\left(\frac{P(y=1)}{P(y=0)}\right) = \ln\left(\frac{P(desejam\ trocar\ de\ carro, mas\ pagariam\ a\ pronto)}{P(não\ desejam\ trocar\ de\ carro)}\right)$$
$$= -0{,}7211727 + 0{,}4218845 dif\_ano + 0{,}1065587 sexo$$
$$+ 0{,}1065587 classesocialB - 0{,}0007898174 classesocialC$$

$$\ln\left(\frac{P(y=2)}{P(y=0)}\right)$$
$$= \ln\left(\frac{P(desejam\ trocar\ de\ carro, mas\ financiariam\ o\ pagamento)}{P(não\ desejam\ trocar\ de\ carro)}\right)$$
$$= -3{,}1140139 + 0{,}674134 dif\_ano + 0{,}4276977 sexo$$
$$+ 0{,}4276977 classesocialB + 2{,}5037291890 classesocialC$$

c) O modelo de regressão logística obtido para quando Y toma os valores de 0 ou 2 foi o seguinte

$$logit(\pi) = -2{,}5558 + 0{,}5470 dif\_ano + 1{,}3736 sexo - 0{,}1840 classesocialB$$
$$+ 2{,}2326 classesocialC$$

Comparando com $\ln\left(\frac{P(y=2)}{P(y=0)}\right)$ as diferenças mais significativas que existem são em β_{sexo} e $\beta_{classesocialB}$, não havendo no geral grandes diferenças entre os dois modelos. Sendo que o AIC do modelo obtido com a regressão logística é inferior, ao do obtido pelo modelo multinomial, sendo estes respetivamente 64,401e 175,8928, pelo que o modelo obtido pela regressão logística poderá ser melhor.