Regresssão Logística - Introdução

Mini-curso de Introdução a ML e AI

Mário Olímpio de Menezes

Maio/2020

Modelos Lineares Generalizados

- Uma das chaves para se escolher o tipo certo de modelo para diferentes tipos de dados é olhar para a **variável dependente**.
  - Para Modelos Lineares, ela até pode não ter uma distribuição normal, mas ela tem que ser *contínua*, *ilimitada* e ser medida em uma escala *intervalar* ou *razão*.
- Todavia, existem muitas situações onde não é razoável assumir que estas condições sejam sempre verdadeiras para a variável dependente.

Modelos Lineares Generalizados

- A variável resposta pode ser categórica:
  - Dicotômica (por exemplo, sim/não, passou/reprovou, viveu/morreu)
  - Policotômicas (por exemplo, ruim/bom/excelente, republicano/democrata/independente)
- Estes tipos de variáveis claramente não são distribuídas normalmente.
- Também podemos ter uma variável resposta que seja uma contagem:
  - Número de veículos que passam em determinado ponto;
  - Número de doses (*de vinho*) que uma pessoa toma em um dia.
- Estas variáveis tem um número limitado de valores possíveis e nunca são negativas.
- Outra importante característica: sua média e desvio padrão são frequentemente relacionados.
  - $\Rightarrow$ **isso não ocorre para variáveis com distribuição normal**

Modelos Lineares Generalizados ampliam o *framework* do modelo linear, incluindo variáveis resposta que não seguem uma distribuição normal.

Modelos Lineares Generalizados e a função `glm()`

- No modelo linear padrão, assumimos que Y tem uma distribuição normal e que a forma do relacionamento é:

$$\mu_Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

  - Os β_j são os parâmetros especificando a mudança esperada em Y para uma mudança unitária em X_j, e β_0 é o valor esperado de Y quando todas as variáveis preditoras são 0 (*Intercept*).
- Não fizemos nenhuma suposição sobre as variáveis preditoras X_j.
- Diferentemente de Y, não há exigência de que elas sejam distribuidas normalmente. De fato, elas são frequentemente categóricas.
- Adicionalemnte, funções não lineares das preditoras são permitidas. Por exemplo, é comum se incluir preditoras do tipo X^2, ou $X_1 \times X_2$. O que é importante é que a equação **seja linear nos parâmetros** ($\beta_0, \beta_1, ..., \beta_p$).

Modelos Lineares Generalizados e a função `glm()`

- Nos modelos lineares generalizados, ajustamos modelos da forma:

$$g(\mu_Y) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

onde $g(\mu_Y)$ é uma função da média condicional (chamada de função *link*). Também *relaxamos* a suposição de que Y seja distribuida normalmente.

- Assumimos, outrossim, que Y segue uma distribuição que é membro da família exponencial.

- Especificamos que a função *link* é a distribuição de probabilidade, e os parâmetros são derivados através de um *procedimento iterativo de estimação por máxima verossimilhança*. Não utilizamos o **Método dos Mínimos Quadrados Ordinários – OLS**.

A função `glm()`

- Modelos Lineares Generalizados são ajustados no R tipicamente utilizando-se a função `glm()`.
- A forma da função é similar à `lm()` mas inclui alguns parâmetros adicionais. O formato básico é:

  `glm(`*formula*`, family=`*family*`(link=`*function*`), data=)`

  onde a distribuição de probabilidade (*family*) e a função **link** correspondente (*function*) são dadas na tabela a seguir:

Funções de probabilidade e a função `link`

Família	Função **link** padrão
`binomial`	`(link = "logit")`
`gaussian`	`(link = "identity")`
`gamma`	`(link = "inverse")`
`inverse.gaussian`	`(link = "1/mu^2")`
`poisson`	`(link = "log")`
`quasi`	`(link = "identity", variance= "contanst")`
`quasibinomial`	`(link = "logit")`
`quasipoisson`	`(link = "log")`

Funções auxiliares

Muitas das funções utilizadas em conjunto com `lm()` quando se analisa modelos lineares tem versões correspondentes para `glm()`. Algumas comumente utilizadas são dadas na tabela a seguir:

Função	Descrição
`summary()`	Mostra resultados detalhados para o modelo ajustado
`coefficients()`,`coef()`	Lista os parâmetros do modelo (deslocamente e inclinação) para o modelo ajustado
`confint()`	Provê os intervalos de confiança para os parâmetros do modelo (95%) por padrão
`residuals()`	Lista os valores dos resíduos em um modelo ajustado
`anova()`	Gera uma tabela ANOVA comparando dois modelos ajustados
`plot()`	Gera gráficos diagnósticos para avaliação do ajuste de um modelo
`predict()`	Usa um modelo ajustado para prever valores de resposta para um novo conjunto de dados
`deviance()`	Desvios para o modelo ajustado
`df.residual()`	Graus de liberdade do resíduo para o modelo ajustado

Diagnósticos da regressão e do ajuste do modelo

- A avaliação da adequabilidade de um modelo é tão importante para modelos lineares generalizados como é para modelo linear padrão (Método *OLS*).
- Há menos consenso na comunidade estatística com relação aos procedimentos de avaliação apropriados. Em geral usamos as mesmas técnicas do modelo linear padrão (ordinário), com algumas ressalvas.
- Quando se avalia a adequabilidade de um modelo, tipicamente plotamos os valores previstos na métrica da variável resposta original contra os resíduos do tipo *deviance*. Por exemplo, um gráficos de diagnóstico comum seria

  ```
  plot(predict(model, type="response"),     residuals(model, type="deviance"))
  ```

  onde `model` é o objeto retornado pela função `glm()`.
- Gráficos de diagnóstico não são úteis quando a variável resposta pode assumir apenas um número limitado de valores (por exemplo, a regressão logística).

Regressão Logística

Regressão Logística

- A regressão logística é util quando queremos prever um resultado binário de um conjunto de variáveis preditoras categóricas ou contínuas. A variável resposta é dicotômica (0 ou 1)

- O modelo assume que Y segue uma distribuição binomial e que podemos ajustar um modelo linear da forma:

$$\log_e\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

  onde $p = \mu_Y$ é a média condicional de Y (isto é, a probabilidade de $Y = 1$ dado um conjunto de valores de X).

- A razão $\left(\frac{p}{1-p}\right)$ é a chance de que $Y = 1$, e $\log(\frac{p}{1-p})$ é o log das chances, ou *logit*
- Neste caso, $\log(\frac{p}{1-p})$ é a função **link**, a distribuição de probabilidade é binomial.

Probabilidade e *Odds* – qual a diferença?

- Quais as chances (*odds*) de chover neste final de semana?
- E a probabilidade de chuva – é a mesma resposta?
  - Tomara que tenhamos dito “Não”!
- Embora usemos os termos intercambiavelmente em conversas informais, isto é um erro porque eles **não** são equivalentes!
  - Sim, eles expressam uma mesma ideia – a possibilidade (*Likelihood*) de um resultado, mas o fazem em diferentes escalas!
  - Usá-los intercambiavelmente, é como misturar *milhas* e *quilômetros* na mesma conversa sem referir a unidade empregada.
  - Pode te fazer correr muito mais do que você pretendia . . .

Probabilidade e *Odds* – qual a diferença?

- Quando medimos a possibilidade de qualquer resultado, precisamos saber duas coisas:
  - Quantas vezes alguma coisa ocorreu e quantas vezes isto poderia ter acontecido, ou equivalentemente, quantas vezes isto não ocorreu.
  - O resultado de interesse aqui é chamado **sucesso**, tanto se for um bom resultado como se não for.
  - O outro resultado é uma **falha**.
- Cada vez que um dos resultados poderia ocorrer é chamado de tentativa.
  - Cada tentativa terminará em um sucesso ou uma falha.
  - O número de sucessos e o número de falhas somados dão o número de tentativas realizadas.
- **Probabilidade** é o número de vezes que ocorreu **sucesso** comparado com o número total de tentativas.
- **Odds** (Chances) é o numero de vezes que ocorreu **sucesso** comparado com om número de falhas ocorridas.

Probabilidade e *Odds* – qual a diferença?

- Como podemos ver, os termos são relacionados, mas **não sinônimos**!

- **Probabilidades** iguais são $0.5 \implies$ 1 sucesso para cada 2 tentativas.
- **Chances** iguais são $1 \implies$ 1 sucesso para cada 1 falha. 1:1

De Probabilidade para *Odds* e para Log de *Odds*

- Digamos que a probabilidade de sucesso de um evento seja 0.8. Então a probabilidade de falha é $1 - 0.8 = .2$.
- As chances de sucesso são definidas como a razão da probabilidade de sucesso pela probabilidade de falha. No nosso exemplo, as chances de sucesso são $\frac{0.8}{0.2} = 4$. Ou seja, as chances de sucesso são **4 para 1**.
- Se a probabilidade de sucesso fosse 0.5, então as chances de sucesso seriam $\frac{0.5}{0.5} = 1$, i.e., 1 para 1.
- A transformação de probabilidade para *odds* é uma transformação monotônica, o que significa que *odds* aumenta conforme a probabilidade diminui ou vice-versa. Probabilidades variam de 0 a 1. *Odds* variam de 0 até infinito positivo.
- A transformação de *odds* para log de *odds* é uma transformação log. Novamente, esta é uma transformação monotônica – quanto maior o *odds*, maior será o log de *odds* e vice-versa.
- A tabela a seguir mostra esta transformação.

p	odds	logodds
.001	.001001	-6.906755
.01	.010101	-4.59512
.15	.1764706	-1.734601
.2	.25	-1.386294
.25	.3333333	-1.098612
.3	.4285714	-.8472978
.35	.5384616	-.6190392
.4	.6666667	-.4054651
.45	.8181818	-.2006707
.5	1	0
.55	1.222222	.2006707
.6	1.5	.4054651
.65	1.857143	.6190392
.7	2.333333	.8472978
.75	3	1.098612
.8	4	1.386294
.85	5.666667	1.734601
.9	9	2.197225
.999	999	6.906755
.9999	9999	9.21024

Interpretando o resultado da Regressão Logística

- Conforme vimos, a regressão logística é definida como

$$\log_e\left(\frac{p}{1-p}\right) = \text{logit}(p) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j$$

ou

$$\log_e\left(\frac{p}{1-p}\right) = \text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_j X_j$$

- Em termos de probabilidades, a equação acima pode ser traduzida em:

$$p = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_j X_j)}{1 + \exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_j X_j)} = \frac{\exp(\text{logit}(p))}{(1 + \exp(\text{logit}(p)))}$$

sendo p a probabilidade de Y ser 1, ou seja, p = prob(y=1).

Super Dispersão (*Overdispersion*)

- A variância esperada para os dados obtidos de uma distribuição binomial é $\sigma^2 = n\pi(1-\pi)$, onde n é o número de observações e π é a probabilidade de se pertencer ao grup $Y = 1$.
- Super Dispersão (*Overdispersion*) ocorre quando a variância observada da variável resposta é maior do que seria esperado de uma distribuição binomial.
- Super Dispersão pode levar a testes distorcidos de erros padrões e testes imprecisos de significância.
- Quando se tem super dispersão, o ajuste com uma função logística ainda é possível utilizando-se a função `glm()`, mas neste caso, é preciso utilizar a distribuição *quasibinomial* ao invés da distribuição binomial.

Super Dispersão (*Overdispersion*)

- Uma maneira de se detectar a super dispersão é comparar o desvio residual com os graus de liberdade dos resíduos no nosso modelo binomial. Se a razão

$$\phi = \frac{Desvio \;\; Residual}{GL \;\; do \;\; Residuo}$$

  é consideravelmente maior do que 1, temos evidência de super dispersão.
- Outro teste que podemos fazer para verificar se temos ou não super dispersão é ajustar o modelo duas vezes:
    - Na primeira vez utilizamos `family="binomial"`
    - Na segunda vez utilizamos `family="quasibinomial"`
- Se o objeto `glm()` retornado no primeiro caso é `fit` e o objeto retornado no segundo caso é `fit.od`, então fazemos:

```
> pchisq(summary(fit.od)$dispersion * fit$df.residual, fit$df.residual,
     lower = F)
```

- A hipótese nula deste teste é $H_0 : \phi = 1$ *versus* a hipótese alternativa $H_1 : \phi \neq 1$.

Exemplo de Regressão Logística

Vamos construir um conjunto de dados hipotético sobre autoavaliação geral de saúde (1=não boa, 0=boa) de n=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 se o indivíduo tem ou não um plano de saúde particular (1=Tem Plano de Saúde Particular, 0= Não tem Plano de Saúde). **Lembrando, são dados absolutamente hipotéticos, inventados!**

```
> 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)
> plano = c(1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1,
     0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1)
> saude = 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, 1, 1)
> autoaval <- data.frame(idade = idade, plano = plano, saude = saude)
```

Modelo

```
> modelo1 = glm(saude ~ idade + plano, family = binomial(link = "logit"),
     data = autoaval)
```

Sumário do Modelo

```
> summary(modelo1)

Call:
glm(formula = saude ~ idade + plano, family = binomial(link = "logit"),
    data = autoaval)

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 **
plano       -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
```

Examinando os coeficientes

```
> round(exp(coef(modelo1)), 3)
(Intercept)       idade       plano
      0.053       1.142       0.042
```

Intervalos de Confiança para o Log das chances (*odds*)

```
> confint.default(modelo1, level = 0.95)
                  2.5 %     97.5 %
(Intercept) -6.35684588  0.4810436
idade        0.03255546  0.2333606
plano       -6.03783801 -0.3201166
```

Agora o Intervalo de Confiança para as chances (*odds*)

```
> exp(confint.default(modelo1, level = 0.95))
                  2.5 %    97.5 %
(Intercept) 0.001734830 1.6177619
idade       1.033091190 1.2628368
plano       0.002386713 0.7260644
```

- Quando olhamos o intervalo de confiança dos parâmetros estimados, uma análise que podemos fazer é ver se o intervalo passa pelo valor nulo.
- Para a Regressão Linear, o valor nulo do intervalo de confiança é o *zero*. Ou seja, se o intervalo de confiança de um parâmetro estimado contém o *zero*, então este parâmetro não tem significância estatística.
- Quando trabalhamos com a Regressão Logística, o valor nulo do intervalo de confiança é o **um** se estamos falando das chances (*odds*) e é o **zero** se estamos falando do log das chances (*log dos odds*). Isso pode ser visto na tabela que compara os valores de Probabilidade, Odds e Log dos Odds.
- Então, percebemos que o Intercept não tem significância estatística porque o seu intervalo de confiança contém o **zero** quando tomamos o log das chances ou contém o **um** quando tomamos as chances.

Interpretação do *odds ratio*

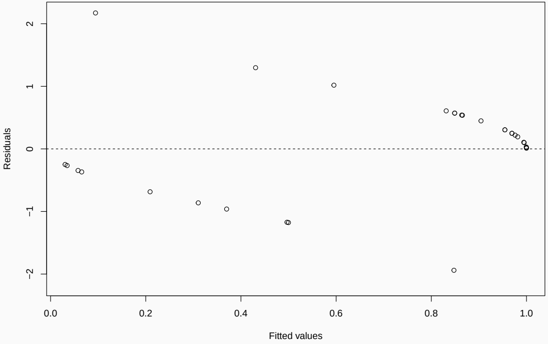
```
> round(exp(coef(modelo1)), 3)
(Intercept)       idade       plano
      0.053       1.142       0.042
```

- Tanto a idade quanto o plano de saúde tem significância estatística no nosso modelo, isto é, com a chance de a autoavaliação de saúde não boa.
- A chance do indivíduo reportar um estado de saúde não bom aumenta em 14,2% ao aumentar em 1 ano a idade, para a mesma condição de plano de saúde.
- Indivíduos com plano de saúde tem uma chance de reportar um estado de saúde não bom 95,8% menor do que os indivíduos que não tem plano de saúde, para a mesma idade.
- Os coeficientes do modelo de regressão logística tem um impacto multiplicativo nas chances (tomando o exponencial) ou nos logs das chances (tomando o valor estimado).

Diagnósticos

Fazendo o diagnóstico do modelo o gráfico dos valores ajustados (preditos) pelos resíduos – semelhante ao que obtemos no `lm`

```
> res <- residuals(modelo1, type = "deviance")
> plot(predict(modelo1, type = "response"),
     res, xlab = "Fitted values", ylab = "Residuals",
     ylim = max(abs(res)) * c(-1, 1))
> abline(h = 0, lty = 2)
```



- Os gráficos diagnósticos de um modelo de regressão logística não fazem muito sentido; por exemplo, no gráfico mostrado ao lado temos no eixo *x* os valores ajustados, onde vemos claramente o resultado da nossa variável resposta ser binária, isto é, temos duas faixas de resíduos: uma para o nível 0 e outra para o nível 1 da variável resposta. Claramente não conseguimos avaliar nada deste gráfico.
- Como nossa variável resposta não é contínua e não temos a premissa de que os resíduos sigam uma distribuição normal, não há porque fazer os demais gráficos diagnósticos.

Diagnósticos

```
> deviance(modelo1)/df.residual(modelo1)
[1] 0.6929918
```

que é próximo de 1, sugerindo que não temos super dispersão.

```
> modelo1.od <- glm(saude ~ idade + plano, family = quasibinomial(),
     data = autoaval)
> pchisq(summary(modelo1.od)$dispersion * modelo1$df.residual,
     modelo1$df.residual, lower = F)
[1] 0.7458603
```

- O resultado do *p-value* (0.746) é claramente não significante ($p > 0.05$), fortalecendo nossa crença de que super dispersão não é um problema (não podemos rejeitar a hipótese nula de que $H_0 : \phi = 1$).

Diagnósticos

A Regressão Logística (`glm`) não tem um R^2 para medirmos o *good of fitness* do modelo. Vários indicadores equivalentes têm sido propostos, sendo o *Pseudo* R^2 de McFadden um bastante utilizado. Valores entre 0.4 e 0.6 são considerados muito bons.

```
> library(DescTools)
```

```
> PseudoR2(modelo1, which = "McFadden")
 McFadden
0.5100717
```

Diagnósticos

Outro pacote que tem **um monte** de funções de diagnóstico para Regressão Logística é o `blorr`.

```
> library(blorr)
```

Uma destas funções é `blr_model_fit_stats`.

```
> blr_model_fit_stats(modelo1)
                              Model Fit Statistics
--------------------------------------------------------------------------------
Log-Lik Intercept Only:       -19.095    Log-Lik Full Model:                -9.355
Deviance(27):                  18.711    LR(2):                             19.480
                                         Prob > LR:                          0.000
MCFadden's R2                   0.510    McFadden's Adj R2:                  0.353
ML (Cox-Snell) R2:              0.478    Cragg-Uhler(Nagelkerke) R2:         0.663
McKelvey & Zavoina's R2:        0.768    Efron's R2:                         0.560
Count R2:                       0.900    Adj Count R2:                       0.700
BIC:                           28.914    AIC:                               24.711
--------------------------------------------------------------------------------
```

A interpretação de todos estes indicadores foge do nosso escopo aqui. O help da função aponta os artigos de referência para o entendimento.

Muito obrigado!

Feedbacks são muito benvindos!