L5P3 - Regressão romântica

Rodolfo Bolconte

24/06/21

Neste problema, utilizaremos dados românticos descritos e disponíveis aqui: <https://github.com/nazareno/ciencia-de-dados-1/tree/master/5-regressao/speed-dating>

Especialmente, atente para uma coluna chamada dec, que diz se houve match entre os dois participantes do encontro – isso é: ambos disseram que gostariam de se encontrar novamente depois: <https://github.com/nazareno/ciencia-de-dados-1/blob/master/5-regressao/speed-dating/speed-dating2.csv>

Sua missão é utilizar regressão logística em um conjunto de variáveis explicativas que você escolher (com no mínimo 4 variáveis e sem usar as variáveis like e prob) para responder o seguinte com esses dados em um RMarkdown:

Que fatores nos dados têm efeito relevante na chance do casal ter um match? Descreva se os efeitos são positivos ou negativos e sua magnitude.

Lembre que temos apenas uma amostra de encontros. Faça inferência para suas conclusões.

Lembre de fazer um descritivo das variáveis antes, e de escrever o relatório de maneira que alguém que saiba sobre regressão mas não sabe nada sobre os dados entenda.

## Carregamento dos Dados

O repositório com os dados: <https://github.com/nazareno/ciencia-de-dados-1/tree/master/5-regressao/speed-dating>

dataset = read\_csv(here::here("data/speed-dating2.csv"))

glimpse(dataset)

## Rows: 4,918  
## Columns: 44  
## $ iid <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3~  
## $ gender <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~  
## $ order <dbl> 4, 3, 10, 5, 7, 6, 1, 2, 8, 9, 10, 9, 6, 1, 3, 2, 7, 8, 4, 5,~  
## $ pid <dbl> 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 11, 12, 13, 14, 15, 1~  
## $ int\_corr <dbl> 0.14, 0.54, 0.16, 0.61, 0.21, 0.25, 0.34, 0.50, 0.28, -0.36, ~  
## $ samerace <dbl> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1~  
## $ age\_o <dbl> 27, 22, 22, 23, 24, 25, 30, 27, 28, 24, 27, 22, 22, 23, 24, 2~  
## $ age <dbl> 21, 21, 21, 21, 21, 21, 21, 21, 21, 21, 24, 24, 24, 24, 24, 2~  
## $ field <chr> "Law", "Law", "Law", "Law", "Law", "Law", "Law", "Law", "Law"~  
## $ race <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2~  
## $ from <chr> "Chicago", "Chicago", "Chicago", "Chicago", "Chicago", "Chica~  
## $ career <chr> "lawyer", "lawyer", "lawyer", "lawyer", "lawyer", "lawyer", "~  
## $ sports <dbl> 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3~  
## $ tvsports <dbl> 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 8~  
## $ exercise <dbl> 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7~  
## $ dining <dbl> 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 10, 10, 10, 10, 10, 10, 10, 10,~  
## $ museums <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 5~  
## $ art <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 5~  
## $ hiking <dbl> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 8~  
## $ gaming <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 4~  
## $ clubbing <dbl> 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 5~  
## $ reading <dbl> 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 10, 10, 10, 10, 10, 10, 10, 10,~  
## $ tv <dbl> 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 8~  
## $ theater <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 7~  
## $ movies <dbl> 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 8, 8, 8, 8, 8, 8, 8, ~  
## $ concerts <dbl> 10, 10, 10, 10, 10, 10, 10, 10, 10, 10, 7, 7, 7, 7, 7, 7, 7, ~  
## $ music <dbl> 9, 9, 9, 9, 9, 9, 9, 9, 9, 9, 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 5~  
## $ shopping <dbl> 8, 8, 8, 8, 8, 8, 8, 8, 8, 8, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 8~  
## $ yoga <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 7~  
## $ attr <dbl> 6, 7, 5, 7, 5, 4, 7, 4, 7, 5, 5, 8, 5, 7, 6, 8, 7, 5, 7, 6, 7~  
## $ sinc <dbl> 9, 8, 8, 6, 6, 9, 6, 9, 6, 6, 7, 5, 8, 9, 8, 7, 5, 8, 6, 7, 9~  
## $ intel <dbl> 7, 7, 9, 8, 7, 7, 7, 7, 8, 6, 8, 6, 9, 7, 7, 8, 9, 7, 8, 8, 1~  
## $ fun <dbl> 7, 8, 8, 7, 7, 4, 4, 6, 9, 8, 4, 6, 6, 6, 9, 3, 6, 5, 9, 7, 7~  
## $ amb <dbl> 6, 5, 5, 6, 6, 6, 6, 5, 8, 10, 6, 9, 3, 5, 7, 6, 7, 9, 4, 9, ~  
## $ shar <dbl> 5, 6, 7, 8, 6, 4, 7, 6, 8, 8, 3, 6, 4, 7, 8, 2, 9, 5, 5, 8, 9~  
## $ like <dbl> 7, 7, 7, 7, 6, 6, 6, 6, 7, 6, 6, 7, 6, 7, 8, 6, 8, 5, 5, 8, 8~  
## $ prob <dbl> 6, 5, NA, 6, 6, 5, 5, 7, 7, 6, 4, 3, 7, 8, 6, 5, 7, 6, 6, 7, ~  
## $ match\_es <dbl> 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, N~  
## $ attr3\_s <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~  
## $ sinc3\_s <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~  
## $ intel3\_s <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~  
## $ fun3\_s <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~  
## $ amb3\_s <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~  
## $ dec <chr> "yes", "yes", "yes", "yes", "yes", "no", "yes", "no", "yes", ~

## Análise Exploratória de Dados

Como o propósito da atividade é selecionar variáveis explicativas do conjunto de dados original para descobrir o efeito na ocorrência de match em encontros, foram selecionadas 5 variáveis, sendo elas:

* attr: quão atraente p1 achou p2
* sinc: quão sincero p1 achou p2
* intel: quão inteligente p1 achou p2
* fun: quão divertido p1 achou p2
* amb: quão ambicioso p1 achou p2

Em que p1 é o primeiro participante do encontro e p2 é o segundo participante.

variaveis\_descritivas = dataset %>%  
 filter(attr!="NA", sinc!="NA", intel!="NA", fun!="NA", amb!="NA", dec!="NA") %>%   
 select(iid, attr, sinc, intel, fun, amb, dec)  
  
count(variaveis\_descritivas)

## # A tibble: 1 x 1  
## n  
## <int>  
## 1 4451

count(variaveis\_descritivas %>% group\_by(dec))

## # A tibble: 2 x 2  
## # Groups: dec [2]  
## dec n  
## <chr> <int>  
## 1 no 2585  
## 2 yes 1866

atributos\_yes = melt(filter(variaveis\_descritivas, dec=="yes"), id.vars = 'iid',  
 measure.vars=c("attr", "sinc", "intel", "fun", "amb"))  
  
atributos\_no = melt(filter(variaveis\_descritivas, dec=="no"), id.vars = 'iid',  
 measure.vars=c("attr", "sinc", "intel", "fun", "amb"))

grafico\_yes = atributos\_yes %>%   
 ggplot(aes(value, variable)) +  
 geom\_jitter(alpha=.2, height=.3, color='chartreuse3') +  
 geom\_boxplot(fill=.01, coef = 1000) +  
 scale\_x\_continuous(breaks=seq(0,10,1)) +  
 theme(text=element\_text(size=16)) +  
 labs(x=NULL, y='', subtitle = "Match = Yes")  
  
grafico\_no = atributos\_no %>%   
 ggplot(aes(value, variable)) +  
 geom\_jitter(alpha=.2, height=.3, color='red') +  
 geom\_boxplot(fill=.01, coef = 1000) +  
 scale\_x\_continuous(breaks=seq(0,10,1)) +  
 theme(text=element\_text(size=16)) +  
 labs(x=NULL, y='', subtitle = "Match = No")  
   
grid.arrange(grafico\_yes, grafico\_no,  
 left = textGrob("Variáveis",  
 gp = gpar(fontsize=16), r=90),  
 bottom = textGrob("Valor",  
 gp = gpar(fontsize=16)))

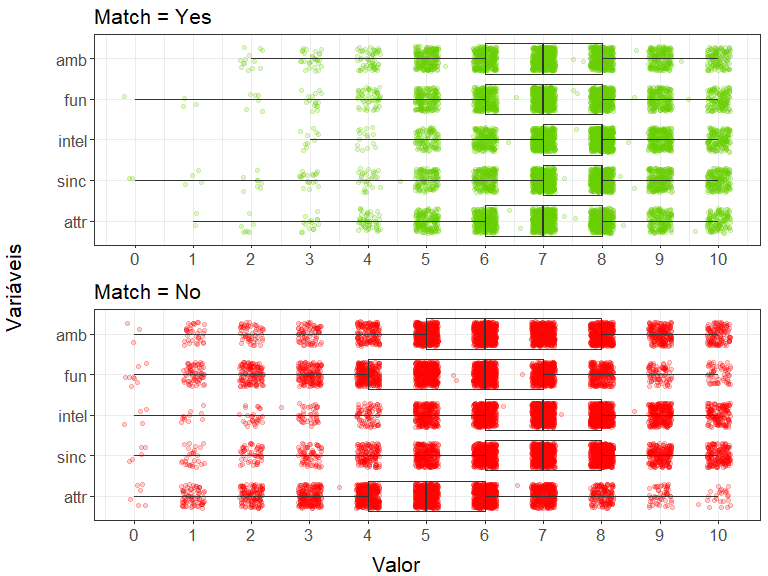


Gráfico 1: Distribuição das avaliações para cada variável separadas por encontros que deram certo ou não.

O Conjunto de Dados original possui 4918 amostras de encontros, porém para a análise foram retiradas amostras que tenham pelo menos uma variável com valor inexistentes (NA), sendo assim, o número total de amostras utilizadas para a análise foi de 4451, sendo 1866 amostras de encontro que deram certo e 2585 amostras de encontro que não deram certo.

Através do Gráfico 1, é possível visualizar que para encontros que deram certo, há pouco valores abaixo de 5 para as variáveis, visto que pela coloração do gráfico, há uma presença maior de valores acima de 5. Na variável intel, que representa o quanto o primeiro participante achou o segundo participante inteligente, apresenta o menor intervalo de avaliações, de 3 a 10, com metade dos valores se concentrando entre 8 e 10, o que pode ser uma ótima variável que explique um encontro dar certo.

Para encontros que não deram certo, todas as variáveis possuem avaliações entre 0 e 10, sendo attr (quão o primeiro participante achou o segundo participante atraente), a variável com uma concentração de valores de avaliações baixa, entre 4 e 6.

## Questão de Atividade

* Que fatores nos dados têm efeito relevante na chance do casal ter um match? Descreva se os efeitos são positivos ou negativos e sua magnitude.

variaveis\_descritivas = variaveis\_descritivas %>% mutate(dec=if\_else(dec=='yes', 1, 0))  
  
modelo = glm(dec ~ attr + sinc + intel + fun + amb, variaveis\_descritivas, family="binomial")  
  
estimativas = tidy(modelo, conf.int=T) %>%   
 select(-statistic, -p.value)  
  
estimativas

## # A tibble: 6 x 5  
## term estimate std.error conf.low conf.high  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) -4.97 0.226 -5.42 -4.54   
## 2 attr 0.573 0.0264 0.522 0.625   
## 3 sinc -0.0831 0.0290 -0.140 -0.0262  
## 4 intel 0.0166 0.0362 -0.0543 0.0875  
## 5 fun 0.363 0.0270 0.311 0.417   
## 6 amb -0.120 0.0277 -0.175 -0.0660

Para responder a questão, foi realizada a criação de um modelo de regressão logística utilizando 5 variáveis (attr, sinc, intel, fun, amb) para tentar explicar quando um encontro dá certo ou não, através da variável categórica “dec”. Com o modelo de regressão criado, foi utilizado o odds (chance) para que o efeito das variáveis explicativas seja analisado em relação à variável de resposta. Sendo assim, o modelo gerou a seguinte equação do nível de chance para cada coeficiente:

O coeficiente de multiplicação para a variável attr (quão atraente p1 achou p2) atingiu o maior valor, de 0,57, sendo assim, se aumentar o valor do atributo attr em uma unidade, aumenta em 57% a chance do primeiro participante achar o segundo um pouco mais atraente. Essa mesma lógica segue para a variável com o segundo maior coeficiente, fun (quão divertido p1 achou p2), com valor de 0,36. Com relação aos coeficientes negativos, a variável amb (quão ambicioso p1 achou p2) atingiu o menor valor, em torno de -0,12, sendo assim, quando o valor da variável aumenta em uma unidade (quando p1 acha p2 mais ambicioso), diminui a chance do encontro dar certo em 12%. Os coeficientes das variáveis intel (quão inteligente p1 achou p2) e sinc (quão sincero p1 achou p2) atingiram valores baixos, alterando as chances de o encontro dar certo por volta de 8% e 1% respectivamente.

estimativas %>% filter(term != "(Intercept)") %>%   
 ggplot(aes(y=reorder(term, conf.low))) +  
 geom\_errorbar(aes(xmin=conf.low, xmax=conf.high), width=.3) +  
 geom\_point(aes(x=estimate), size=3) +  
 geom\_vline(xintercept=0) +  
 geom\_vline(xintercept=.3, lty=2) +  
 geom\_vline(xintercept=-.3, lty=2) +  
 theme(text=element\_text(size=16)) +  
 labs(x='\nEstimativa', y='Variáveis\n')

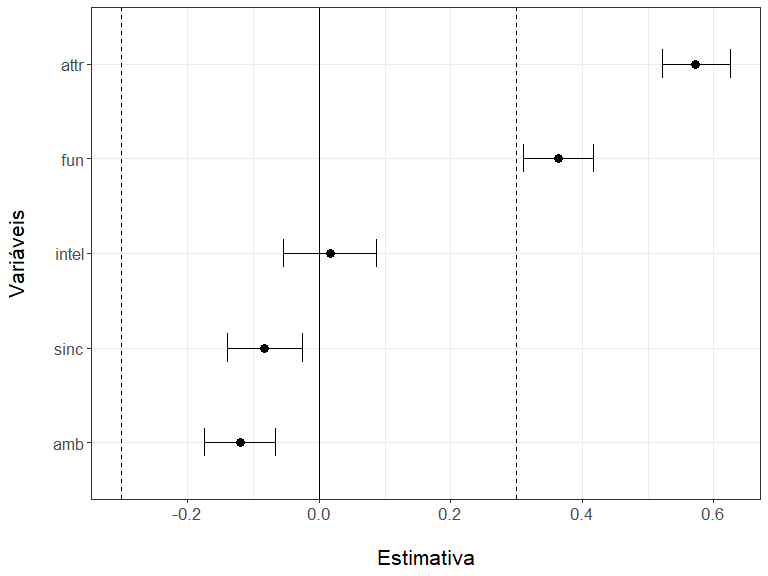


Gráfico 2: Intervalos de Confiança dos coeficientes das variáveis explicativas analisadas.

No Gráfico 2 é possível visualizar os intervalos de confiança do odds para os coeficientes das variáveis, sendo intel a variável com o coeficiente que não é possível inferir se terá efeito ou não no odds do encontro dar certo, enquanto os coeficientes de fun e attr se mostraram altos, indicando um maior efeito no impacto do odds para os encontros, sendo attr atingindo o maior valor de intervalo, até 0,62, ou 62%.

pscl::pR2(modelo)

## fitting null model for pseudo-r2

## llh llhNull G2 McFadden r2ML   
## -2339.2122810 -3026.8704132 1375.3162645 0.2271845 0.2658122   
## r2CU   
## 0.3575815

Por fim temo o valor do pseudo , em que se baseando no valor de McFadden, o modelo é capaz de explicar cerca de 22,7% da variação da variável de resposta, no caso a variável dec, quando utilizando apenas as 5 variáveis apresentadas anteriormente.