Introdução à Análise de Dados Categorizados

Ivonaldo Silvestre

03/12/2020

Introdução

A franquia de jogos chamada *Pokémon*, acrônimo de *Pocket Monsters* (monstros de bolso, em tradução livre), é conhecida por ter criaturas de mesmo nome da franquia, mas com características particulares. São conhecidas 898 espécies dentre as 8 gerações existentes até agora, algumas delas apresentam múltiplas formas que interferem nas suas estatísticas. O banco de dados desse trabalho, disponível em https://www.kaggle.com/abcsds/pokemon, apesar de ter 800 observações, contém informações sobre os 721 *Pokémon* existentes até a sexta geração, já que inclui também formas diferentes de um mesmo monstro.

As variáveis presentes no banco de dados são

- **X.**: o seu número na *Pokédex*, a lista oficial dos *Pokémon*;
- **Name**: o seu nome;
- **Type.1**: o seu tipo principal. Atualmente há 18 tipos diferentes;
- **Type.2**: o seu segundo tipo;
- **Total**: a soma de todas as seis estatísticas base (as próximas variáveis);
- **HP**: a quantidade de vida base;
- **Attack**: a quantidade de ataque base;
- **Defense**: a quantidade de defesa base;
- **Sp..Atk**: a quantidade de ataque especial base;
- **Sp..Def**: a quantidade de defesa especial base;
- **Speed**: a quantidade de velocidade base;
- **Generation**: a sua geração (de 1 a 6);
- **Legendary**: uma variável categórica que representa a raridade do *Pokémon*. *False* se não for lendário e *True* se for lendário.

Este trabalho tem o objetivo de atribuir uma probabilidade em relação à raridade do *Pokémon* das seis primeiras gerações com base nas seis estatísticas, a vida, o ataque, a defesa, o ataque especial, a defesa especial e a velocidade, o que pode servir como critério de escolha para uma equipe. Já é sabido que os lendários, em média, têm estatísticas maiores que os comuns. Apesar de ser um censo das seis primeiras gerações, esse banco de dados será tratado como uma amostra aleatória simples.

Metodologia

Regressão Logística

Sejam $\boldsymbol{y} = (y_1, ..., y_n)^\top$ o vetor de variáveis aleatórias independentes com distribuição Binomial $(1, \pi_i)$, $i = 1, \ldots, n$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ a matriz de variáveis explicativas em que cada $\boldsymbol{x}_i = (1, x_1, \ldots, x_p)^\top$ é relacionado à i-ésima resposta e $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ o vetor de parâmetros associados às variáveis. Para estimar o vetor de probabilidades $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)$ com os preditores lineares $\boldsymbol{X\beta}$, usa-se a função logito, definida por $logito(x) = log\left(\frac{x}{1-x}\right)$.

Utilizando a notação $\hat{}$ para os estimadores, pode-se definir

$$logito(\hat{\boldsymbol{\pi}}) = \boldsymbol{X}\hat{\boldsymbol{\beta}},$$

ou, analogamente,

$$\hat{\boldsymbol{\pi}} = \frac{exp(\boldsymbol{X}\hat{\boldsymbol{\beta}})}{1 + exp(\boldsymbol{X}\hat{\boldsymbol{\beta}})}.$$

O β_0 pode ser interpretado como a probabilidade de sucesso quando as demais variáveis são 0 e os $\beta_i, i = 1, \ldots, p$ pode ser interpretado como a alteração no logarítmo da razão de chances (Ω) de quando sua variável correspondente x_i aumenta em uma unidade, ou seja, $\beta_i = log\left(\frac{\Omega(x_i+1)}{\Omega(x_i)}\right)$.

O vetor $\hat{\boldsymbol{\beta}}$ segue distribuição normal com média $\boldsymbol{\beta}$ e matriz de covariância igual a $(\boldsymbol{X}^\top W(\boldsymbol{\beta})\boldsymbol{X})^{-1}$, em que

$$W(\boldsymbol{\beta}) = diag\left(\frac{exp(\boldsymbol{x}_1^\top\boldsymbol{\beta})}{1 + exp(\boldsymbol{x}_1^\top\boldsymbol{\beta})}, \ \ldots, \ \frac{exp(\boldsymbol{x}_n^\top\boldsymbol{\beta})}{1 + exp(\boldsymbol{x}_n^\top\boldsymbol{\beta})}\right).$$

Usa-se $W(\hat{\boldsymbol{\beta}})$ para se estimar a matriz de covariâncias de $\hat{\boldsymbol{\beta}}$. Para testar se um componente β_i, é significativamente diferente de 0, é utilizada a estatística Wald (ou valor Z),

$$S_W = \frac{\hat{\beta}_i}{e.p.(\hat{\beta}_i)},$$

e compara-se seu módulo com o quantil da distribuição normal, ou calcula-se a probabilidade de obter valores mais extremos que o observado e compara com o nível desejado de significância.

Para construir um intervalo com 100(1-α)% de confiança, usa-se a fórmula

$$IC(\beta_i; 100(1-\alpha)\%) = [\hat{\beta}_i - Z_{1-\alpha/2}e.p.(\hat{\beta}_i), \hat{\beta}_i + Z_{1-\alpha/2}e.p.(\hat{\beta}_i)]$$

Para testar a significância de um sub-vetor de $\hat{\boldsymbol{\beta}}$ com tamanho q, pode-se utilizar a estatística da Razão da Verossimilhança (ou desvio), dada por

$$S_{LR} = 2(l(\hat{\boldsymbol{\beta}}; \boldsymbol{X}) - l(\tilde{\boldsymbol{\beta}}; \boldsymbol{X})),$$

em que $\tilde{\boldsymbol{\beta}}$ representa o estimador de máxima verossimilhança sob hipótese nula e l, a log-verossimilhança. Rejeita-se a hipótese nula se S_{LR} for maior que o quantil da χ^2_q cuja acumulada é igual a $1 - \alpha$.

Diagnóstico

Os dois primeiros tipos de resíduos a serem definidos são o resíduo de Pearson, que para a i-ésima observação é dado por

$$r_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}$$

e o resíduo de desvio, também para a i-ésima observação é dado por

$$d_i = sinal(y_i - \hat{\pi}_i)\sqrt{-2[y_i ln\ \hat{\pi}_i + (1-y_i)ln(1-\hat{\pi}_i)]}.$$

Seja $\hat{\boldsymbol{H}} = \hat{\boldsymbol{V}}^{1/2}\boldsymbol{X}[\boldsymbol{X}^\top\hat{\boldsymbol{V}}\boldsymbol{X}]^{-1}\boldsymbol{X}^\top\hat{\boldsymbol{V}}^{1/2}$, em que $\hat{\boldsymbol{V}} = diag\{\hat{\pi}_1(1-\hat{\pi}_1),\ \ldots,\ \hat{\pi}_n(1-\hat{\pi}_n)\}$. Cada elemento da sua matriz diagonal será chamado de $\hat{h}_{ii}$ e a partir deles pode-se definir o resíduo padronizado de Pearson e o resíduo padronizado de desvio para a i-ésima observação por

$$t_{r_i} = \frac{r_i}{\sqrt{1-\hat{h}_{ii}}}$$

e

$$t_{d_i} = \frac{d_i}{\sqrt{1-\hat{h}_{ii}}}.$$

Por fim, a distância de Cook, definida como

$$LD_i = \frac{\hat{h}_{ii}}{1-\hat{h}_{ii}} t_{r_i}^2$$

Quatro gráficos são usados para o diagnóstico do modelo. O primeiro deles é dos valores ajustados pela medida h. Valores altos de $\hat{h}_{ii}$ indicam que esses observações podem ser valores influentes para $\hat{\pi}_i$. O segundo gráfico é o da distância de Cook. Valores maiores que 1 indicam que essas observações também podem ser valores influentes. O terceiro gráfico é o gráfico dos resíduos de desvio padronizados. Esse gráfico tem a finalidade de mostra quantas observações tiveram resíduos altos, ou seja, observações cujo resíduo em módulo foi maior que 2. O último gráfico é o dos valores ajustados pelos resíduos de desvio. Similar ao gráfico anterior, esse gráfico também mostra as observações com valores altos e também é possível ver a tendência à medida que os valores ajustado ficam mais distantes dos valores reais.

Também é possível realizar um gráfico de envelope simulado com os resíduos padronizados de desvio. Para construir esse gráfico é necessário gerar aleatoriamente n números aleatórios entre 0 e 1 e compará-los com os valores ajustados. Caso esses números aleatórios sejam maiores, atribui-se como 0 e caso sejam menores, atribui-se 1. Com essa nova amostra simulada, ajusta-se o modelo e calcula-se os resíduos padronizados. Esse processo é repetido quantas vezes forem necessárias. Escolhe-se os quantis dos resíduos de interesse e então é criado um intervalo de confiança empírico. Se os resíduos da amostra original estiverem dentro desses limites, há indícios de que o modelo está bem ajustado.

Curva ROC

A partir da escolha de um ponto de corte, pode-se definir a especificidade como a razão entre o número de elementos com a ausência do evento de interesse que foram classificados corretamente e o número de elementos com a ausência do evento de interesse e a sensitividade como a razão entre o número de elementos com a presença do evento de interesse que foram classificados corretamente e o número de elementos com a presença do evento de interesse. A curva ROC (*Receiver Operating Characteristic*, ou Característica de Operação do Receptor) é gerada com os valores da sensibilidade e de 1 - a especificidade com pontos de corte gerados aleatoriamente e calculando a área abaixo dessa curva. Uma área igual a 1/2 indica que a escolha é totalmente aleatória. Curvas consideradas excepcionais ocorrem quando o valor da área é maior que 0,9.

Resultados

Análise Descritiva

Fazendo uma média das estatísticas dos *Pokémon* com múltiplas formas, há no total 675 *Pokémon* classificados como comuns e 46 classificados como lendários. A Figura 1 mostra os gráficos de dispersão de cada uma das estatísticas. É possível perceber que, no geral, os *Pokémon* lendários apresentam estatísticas maiores.

Também é possível perceber que as variáveis apresentam uma correlação positiva entre si, para confirmar isso, a Figura 2 mostra os valores nominais de cada correlação 2 a 2.

A Figura 3 mostra os boxplots das estatísticas por raridade de cada *Pokémon*, o que confirma visualmente a superioridade dos lendários.

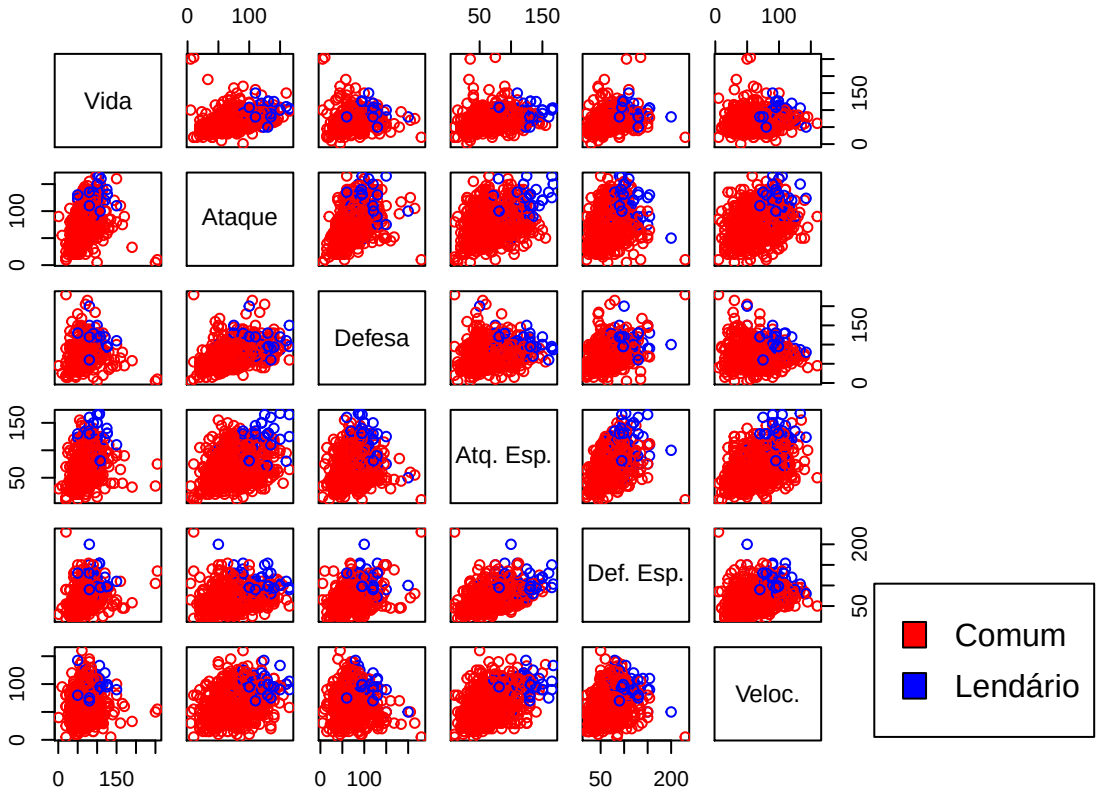

Figura 1: Gráficos de dispersão de cada estatística 2 a 2

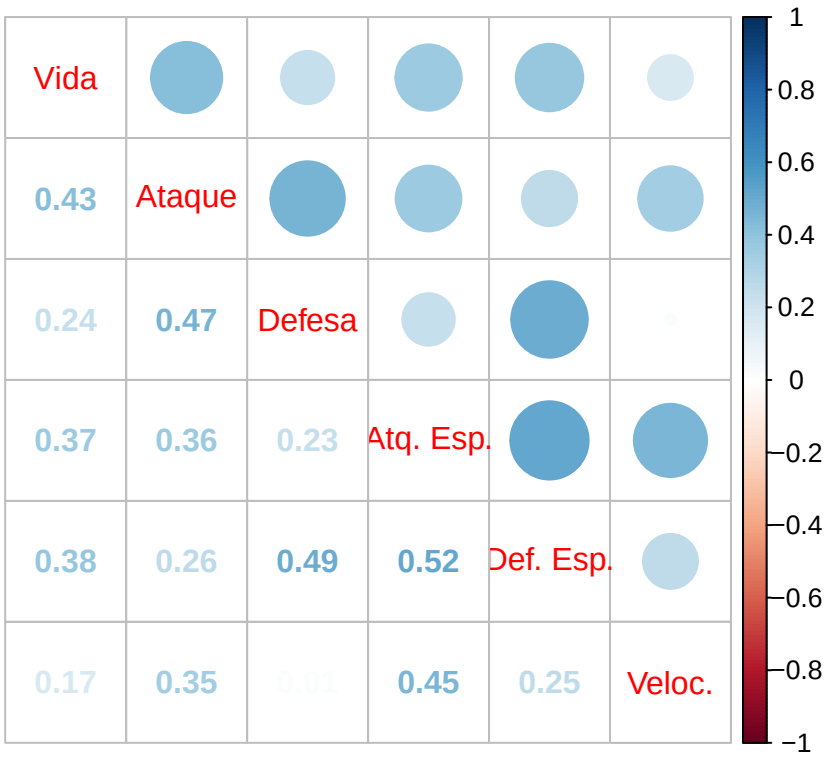

Figura 2: Gráfico da correlação entre cada estatística 2 a 2

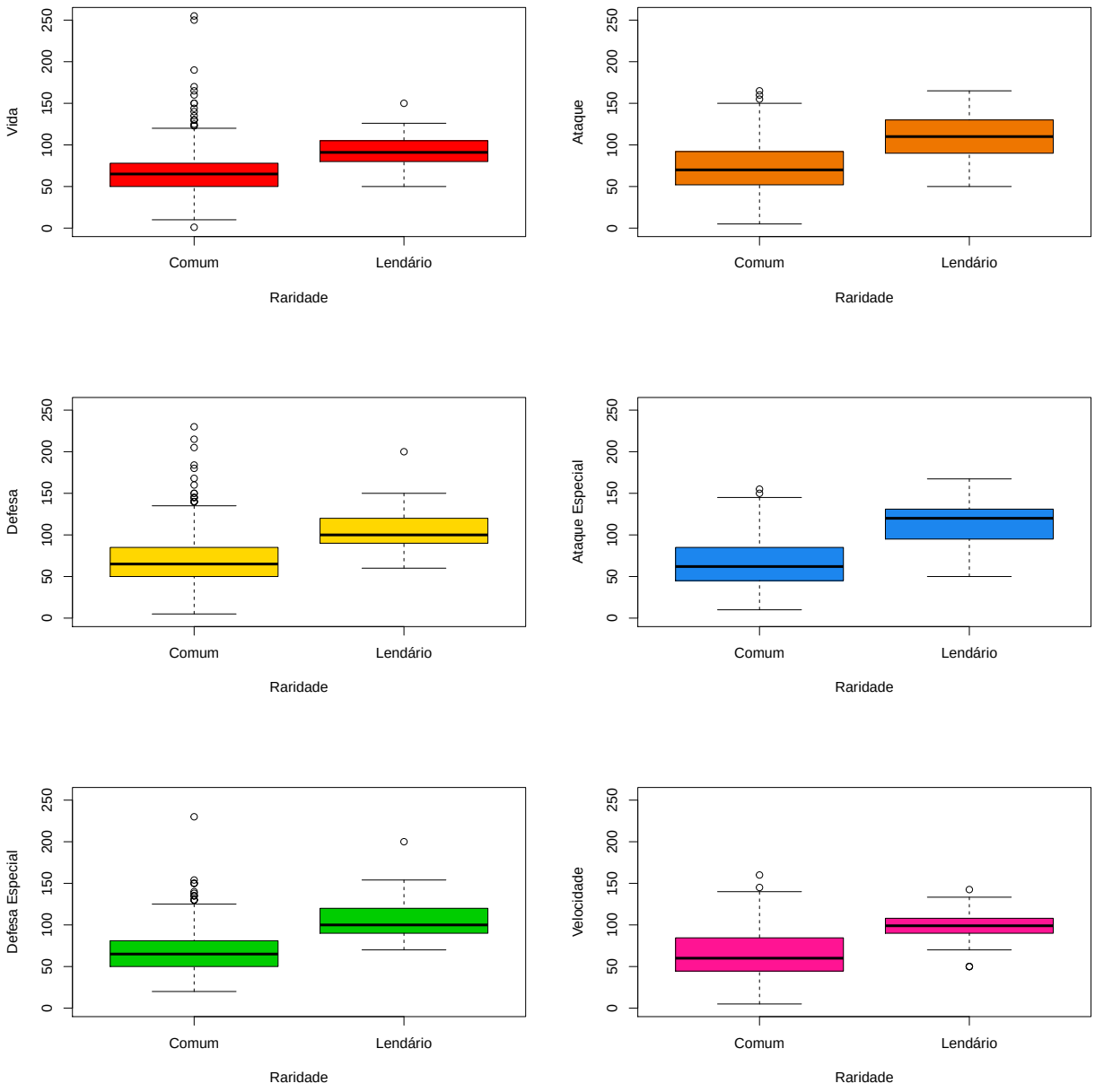

Figura 3: Boxplots das estatísticas por raridade do Pokémon

Análises individuais

Analisando separadamente cada estatística por classes, a quantidade de vida teve coeficiente positivo, 0,0289 com erro padrão igual a 0,0052, e teve seu valor-p menor que 0.05 (Tabela 1) e a tabela ANOVA mostrou sua significância (Tabela 2), entretanto, a Figura 4 mostra que nas últimas classes a probabilidade cai, ao invés de subir. Isso pode ser explicado devido ao fato de o lendário com maior Vida tem 150 pontos nessa estatística.

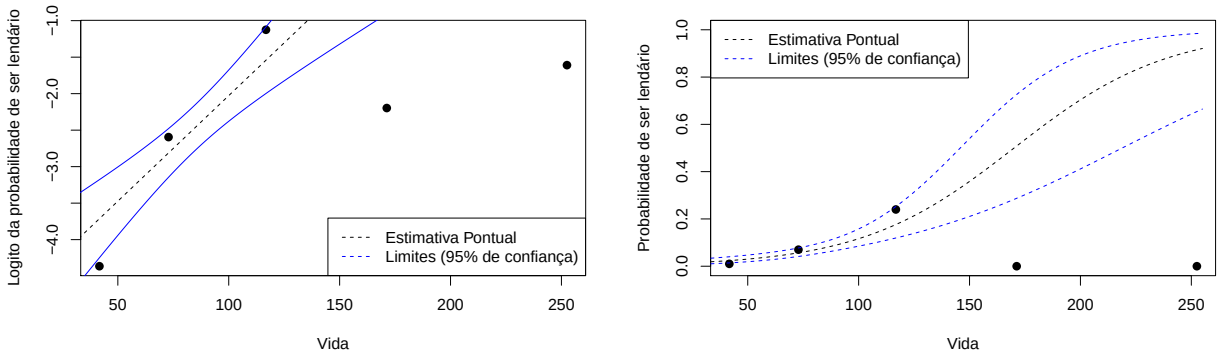

Figura 4: Gráficos da Vida pelo logito da probabilidade de ser lendário e pela probabilidade de ser lendário

	Estimativa	Erro padrão	Valor Z	Pr(>|z|)
Intercepto	-4,9189	0,4667	-10,54	<0,0001
Vida	0,0289	0,0052	5,55	<0,0001

Tabela 1: Sumário do modelo com apenas a Vida.

Modelo	G. L. Resid.	Desvio Resid.	G. L.	Desvio	Pr(>Chi)
Nulo	720	342,18			
Vida	719	306,82	1	35,36	<0.0001

Tabela 2: ANOVA do modelo com apenas a Vida.

Como pode ser visto na Tabela 3, o ataque também possui um efeito positivo (0,0427 com erro padrão igual a 0,0058), como esperado, e diferente de 0 a 5% de significância, como também pode ser visto na Tabela 4. Os gráficos da Figura 5 mostram indícios que, apesar de o último ponto ter ficado de fora dos limites de confiança, o logito da razão de chances é linear no ataque.

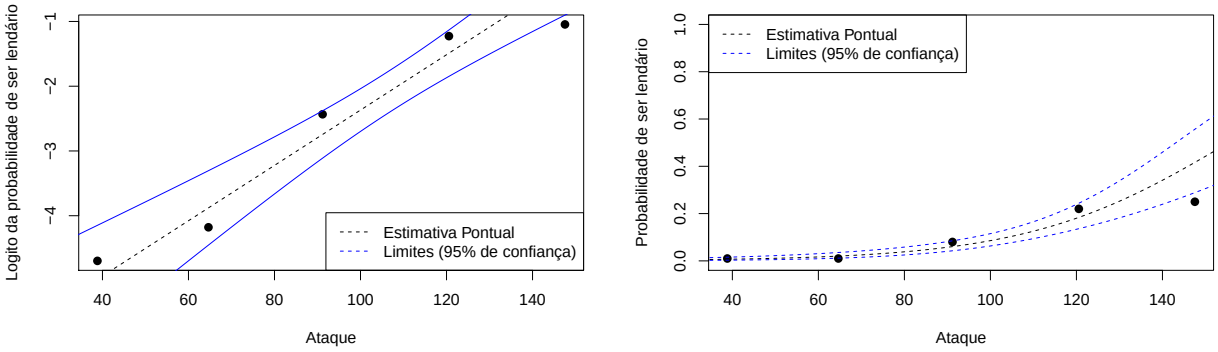

Figura 5: Gráficos do Ataque pelo logito da probabilidade de ser lendário e pela probabilidade de ser lendário

	Estimativa	Erro padrão	Valor Z	Pr(>|z|)
Intercepto	-6,6414	0,6388	-10,40	<0,0001
Ataque	0,0427	0,0058	7,38	<0,0001

Tabela 3: Sumário do modelo com apenas o Ataque.

Modelo	G. L. Resid.	Desvio Resid.	G. L.	Desvio	Pr(>Chi)
Nulo	720	342,18			
Ataque	719	274,67	1	67,52	<0.0001

Tabela 4: ANOVA do modelo com apenas o Ataque.

Os gráficos da Figura 6 mostram que, com a defesa como variável preditora, as probabilidades de cada classe, com exceção da última, estão dentro ou se aproximam dos limites de confiança, apesar de o logito delas não estar bem ajustado. A estimativa do coeficiente é igual a 0,0283 com erro padrão igual a 0,0045 e, de acordo com a Tabela 6, a estatística defesa possui significância a 5% em relação ao modelo nulo.

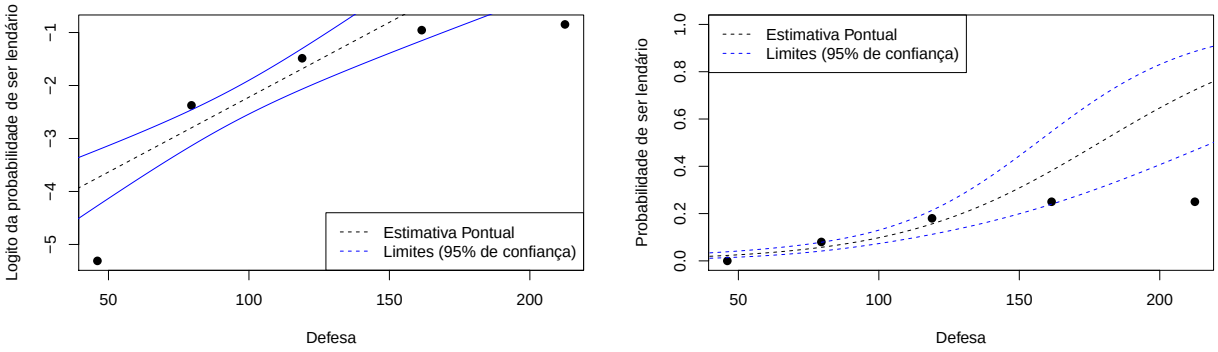

Figura 6: Gráficos da Defesa pelo logito da probabilidade de ser lendário e pela probabilidade de ser lendário

	Estimativa	Erro padrão	Valor Z	Pr(>|z|)
Intercepto	-5,0459	0,4512	-11,18	<0.0001
Defesa	0,0283	0,0045	6,30	<0.0001

Tabela 5: Sumário modelo com apenas a Defesa.

Modelo	G. L. Resid.	Desvio Resid.	G. L.	Desvio	Pr(>Chi)
Nulo	720	342,18			
Defesa	719	299,81	1	42,37	<0.0001

Tabela 6: ANOVA do modelo com apenas a Defesa.

Já com o ataque especial, a Figura 7 mostra que as classes foram bem ajustadas pelo modelo linear e com coeficiente também positivo, com valor igual a 0,0589 e erro padrão igual a 0,0070. Como esperado, essa variável também foi significativa a 5% em relação ao modelo nulo (Tabela 8).

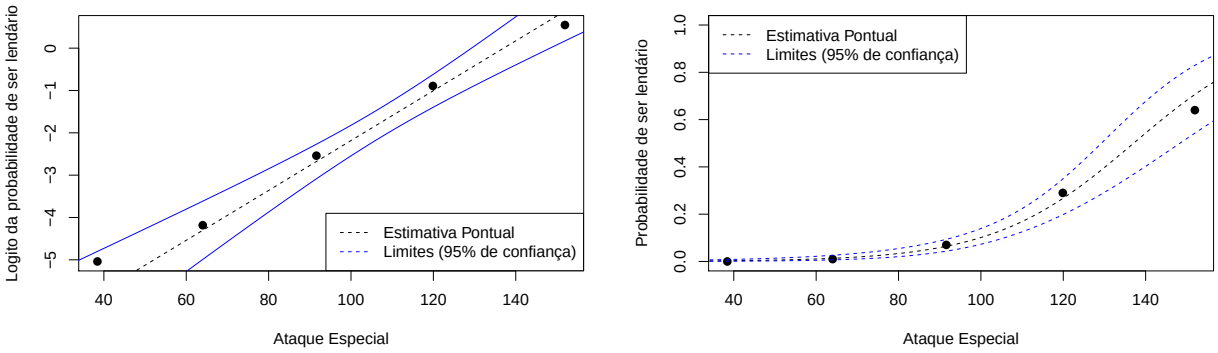

Figura 7: Gráficos do Ataque Especial pelo logito da probabilidade de ser lendário e pela probabilidade de ser lendário

	Estimativa	Erro padrão	Valor Z	Pr(>|z|)
Intercepto	-8,0724	0,7724	-10,45	<0.0001
Ataque Especial	0,0589	0,0070	8,40	<0.0001

Tabela 7: Sumário do modelo com apenas o Ataque Especial.

Modelo	G. L. Resid.	Desvio Resid.	G. L.	Desvio	Pr(>Chi)
Nulo	720	342,18			
Ataque Especial	719	228,38	1	113,81	<0,0001

Tabela 8: ANOVA do modelo com apenas o Ataque Especial.

Similar à defesa, a defesa especial mostra um bom ajuste nas primeiras classes, entretanto, a última classe ficou fora de ambos limites nos gráficos da Figura 8. Essa estatística também teve coeficiente positivo, com valor igual a 0,0479 e desvio padrão igual a 0,0062, sendo significativa a 5% em relação ao modelo nulo.

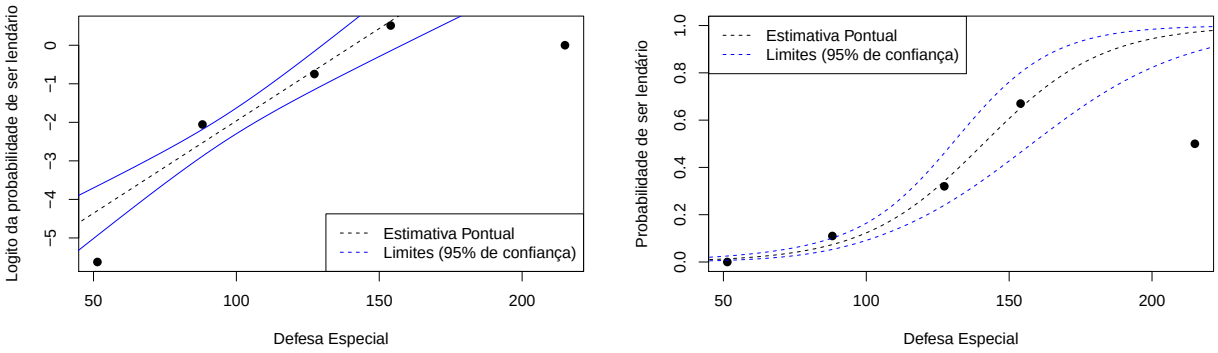

Figura 8: Gráficos da Defesa Especial pelo logito da probabilidade de ser lendário e pela probabilidade de ser lendário

	Estimativa	Erro padrão	Valor Z	Pr(>|z|)
Intercepto	-6,7538	0,6216	-10,87	<0,0001
Defesa Especial	0,0479	0,0062	7,75	<0,0001

Tabela 9: Sumário do modelo com apenas a Defesa Especial.

Modelo	G. L. Resid.	Desvio Resid.	G. L.	Desvio	Pr(>Chi)
Nulo	720	342,18			
Defesa Especial	719	261,78	1	80,41	<0.0001

Tabela 10: ANOVA do modelo com apenas a Defesa Especial.

A última estatística, a velocidade, apresentou valores sempre dentro ou ao redor dos limites de confiança, o que pode indicar que há linearidade apresente nessa estatística. Seu coeficiente foi 0,0448 com desvio padrão 0,0065 e a 5% de significância, mostrou-se significativa em relação ao modelo nulo.

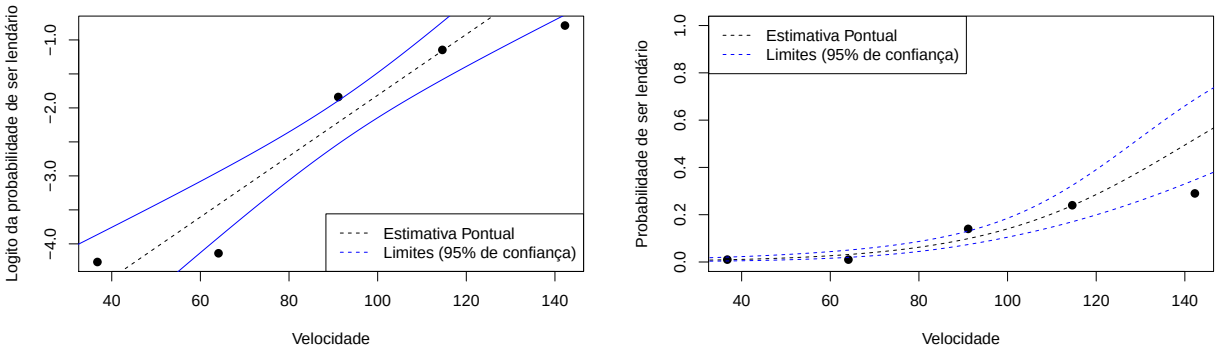

Figura 9: Gráficos da Velocidade pelo logito da probabilidade de ser lendário e pela probabilidade de ser lendário

	Estimativa	Erro padrão	Valor Z	Pr(>|z|)
Intercepto	-6,2930	0,6264	-10,05	<0,0001
Velocidade	0,0448	0,0065	6,85	<0,0001

Tabela 11: Sumário do modelo com apenas a Velocidade.

Modelo	G. L. Resid.	Desvio Resid.	G. L.	Desvio	Pr(>Chi)
Nulo	720	342,18			
Velocidade	719	282,74	1	59,45	<0.0001

Tabela 12: ANOVA do modelo com apenas a Velocidade.

Análise conjunta

A Tabela 13 mostra os coeficientes estimados, seus respectivos desvios padrão, os valores z para cada um deles e seus respectivos valores-p da regressão logística para estimar a probabilidade de um *Pokémon* ser lendário com seis estatísticas.

	Estimativa	Erro Padrão	Valor z	Pr(>|z|)
Intercepto	-31,9482	4,7421	-6,7371	<0,0001
Vida	0,0411	0,0149	2,7630	0,0057
Ataque	0,0212	0,0111	1,9135	0,0557
Defesa	0,0660	0,0138	4,7764	<0,0001
Ataque Especial	0,0552	0,0131	4,2124	<0,0001
Defesa Especial	0,0594	0,0129	4,6166	<0,0001
Velocidade	0,0781	0,0174	4,4894	<0,0001

Tabela 13: Regressão logística das seis estatísticas.

Como pode-se perceber na Tabela 13, considerando um nível de significância de 5%, apenas o ataque não foi significativo, entretanto, ao observar a Tabela 14 comparando o modelo sem ataque e com ataque e considerando o mesmo nível de significância com a razão de verossimilhanças, há evidências para rejeitar a hipótese nula de que o Ataque não possui efeito no modelo em que já estão presentes as demais estatísticas, logo, deve ser utilizado.

Modelo	G. L. Resid.	Desvio Resid.	G. L.	Desvio	Pr(>Chi)
Sem Ataque	715	104,59			
Com Ataque	714	100,62	1	3,97	0,0464

Tabela 14: ANOVA da Regressão logística comparando os modelos com e sem Ataque.

Por fim, a Tabela 15 mostra o teste de significância entre o modelo completo, ou seja, com as seis estatísticas, e o modelo nulo. Ainda considerando a significância igual a 5%, pode-se concluir, como esperado, que o modelo completo de fato explica melhor a probabilidade de um *Pokémon* ser lendário do que o modelo sem variáveis regressoras.

Modelo	G. L. Resid.	Desvio Resid.	G. L.	Desvio	Pr(>Chi)
Nulo	720	342,18			
Completo	714	100,62	6	241,56	<0,0001

Tabela 15: ANOVA da Regressão logística comparando os modelos nulo e completo.

Portanto, o logito probabilidade estimada de um *Pokémon* ser lendário, $\hat{\pi}$, pode ser escrito pela forma

$$logito(\hat{\pi}) = -31,9482 + 0,0411 \times \text{Vida} + 0,0212 \times \text{Atq.} + 0,066 \times \text{Def.} + 0,0552 \times \text{A.E.} + 0,0594 \times \text{D.E.} + 0,0781 \times \text{Vel.}$$

e a probabilidade de um *Pokémon* ser lendário pode ser escrita como

$$\hat{\pi} = \frac{exp(logito(\hat{\pi}))}{1 + exp(logito(\hat{\pi}))}.$$

Análise de Resíduos

A Figura 10 mostra os quatro gráficos da análise de resíduos. O primeiro deles, dos valores ajustados pela medida h mostra que apenas o *Pokémon* de número 213, Shuckle, que é comum, obteve uma medida h

muito mais alta que as demais. Isso pode ser explicado devido ao valor extremamente alto de defesa e defesa especial e extremamente baixo nas demais estatísticas que ele tem. O segundo gráfico, das distâncias de Cook mostra que nenhum deles teve distância maior que 1, portanto, não há pontos de alavanca. No terceiro gráfico, dos resíduos de desvio padronizados, apenas dois *Pokémon* ficaram acima do limite superior, que foram 244 (Entei) e 639 (Terrakion), indicando que eles são lendários com probabilidade estimada menor que os demais lendários, por outro lado, dois pontos ficaram abaixo do limite inferior, 373 (Salamence) e 376 (Metagross), indicando que eles têm probabilidade estimada maior que os demais comuns. Por fim, no último gráfico, dos valores ajustados pelos resíduos de desvio, o conjunto de pontos inferiores, que representa os comuns, mostra o valor residual, em módulo, aumentando de forma linear à medida que a probabilidade estimada aumenta, por outro lado, os resíduos dos lendários (grupo superior), vai diminuindo em caráter linear à medida que a probabilidade aumenta.

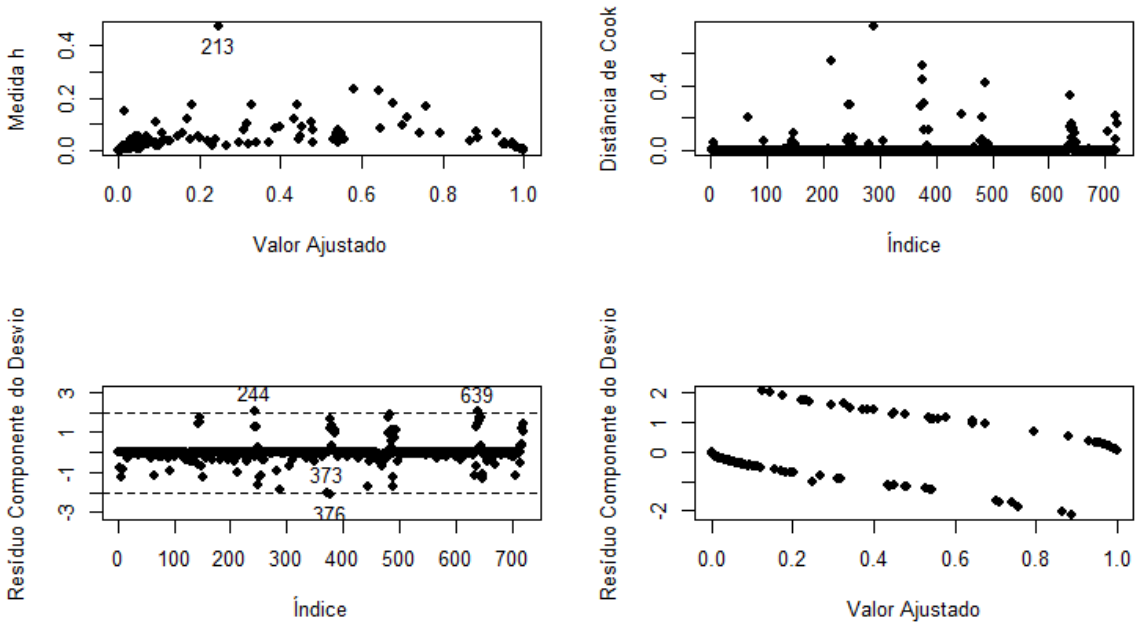

Figura 10: Análise residual

Gráfico de envelope simulado

A Figura 11 mostra o gráfico de envelope simulado com 95% de confiança. É possível perceber que, apesar de alguns pontos nas extremidades estarem fora do intervalo de confiança, a maioria deles está dentro, dando indícios de um bom ajuste.

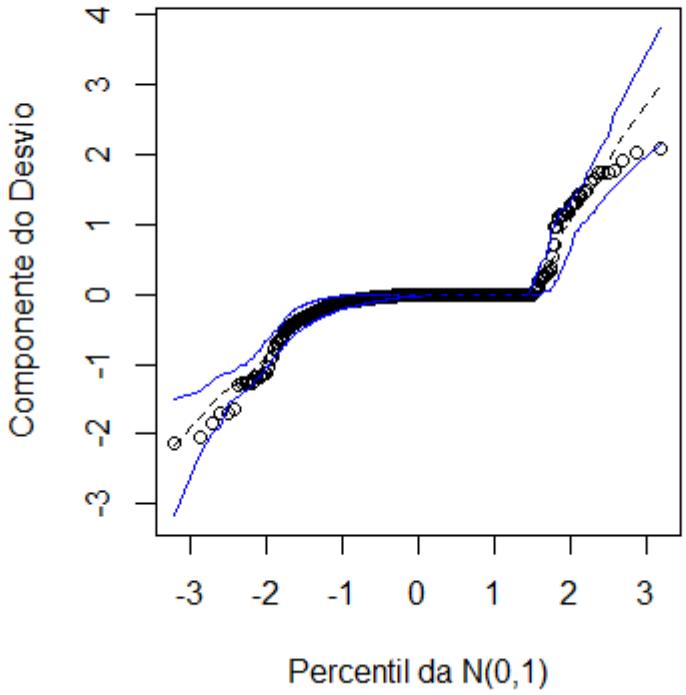

Figura 11: Gráfico de envelope com 95% de confiança

Curva ROC

A Figura 12 mostra a curva ROC e a respectiva área sobre a curva. O valor da área foi muito próxima de 1, com valor 0,9871, indicando um ajuste muito bom da regressão logística da raridade em relação às seis estatísticas.

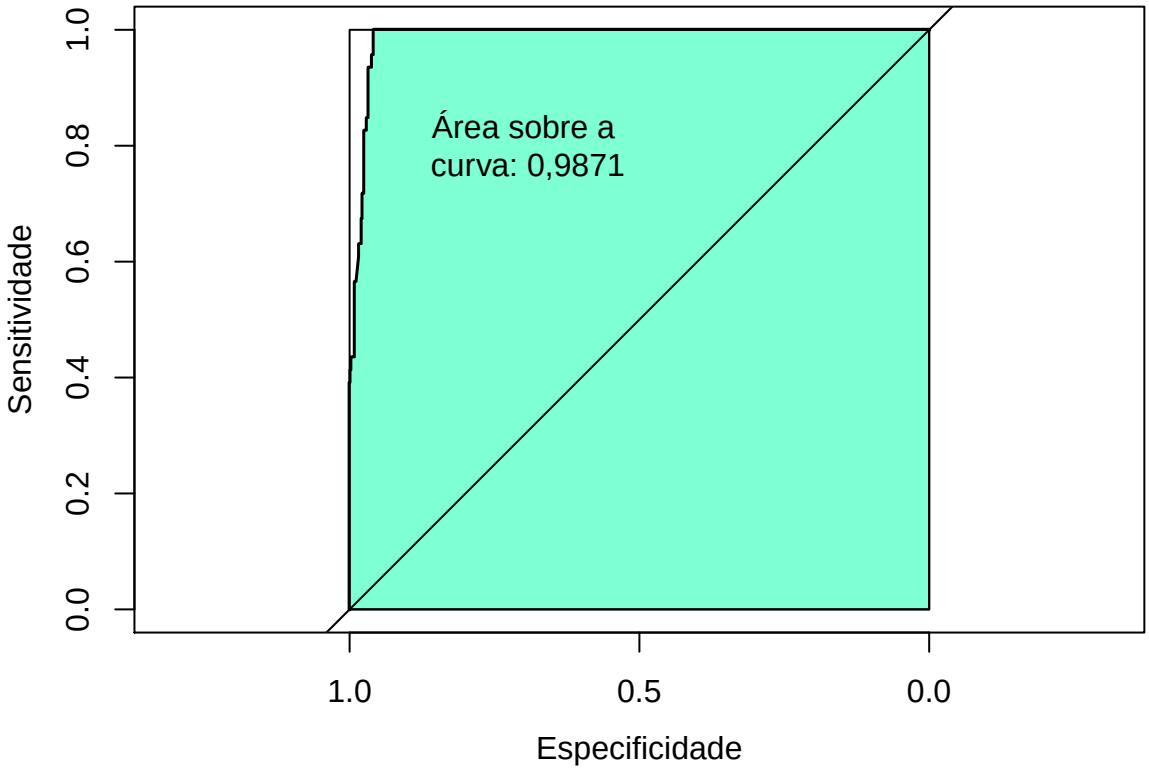

Figura 12: Curva ROC

Critério de seleção

O critério de seleção para classificar um *Pokémon* como lendário foi de deixar a especificidade e a sensibilidade com valores iguais ou próximos, tendo como resultado a probabilidade 0,1461, ou seja, se um *Pokémon* tiver probabilidade estimada de ser lendário maior que 0,1461, ele será classificado lendário, caso contrário, ele seria classificado como comum. A Tabela 16 mostra a matriz de confusão da classificação pela raridade utilizando esse critério e os valores entre parênteses indicam a porcentagem dos *Pokémon* classificados por raridade. A probabilidade de acerto é de quase 96%, indicando que esse critério possui uma taxa baixa de falsos positivos e falsos negativos. Apenas Entei (número 244) e Terrakion (número 639) foram os lendários classificados incorretamente.

		Raridade	
		Comum	Lendário
Classificação	Comum	647 (95,85%)	2 (4,35%)
	Lendário	28 (4,15%)	44 (95,65%)
	Total	675 (100%)	46 (100%)

Tabela 16: Classificação pela raridade.

Foram escolhidos três *Pokémon* da sétima geração, portanto não estão presentes no banco de dados, para testar se o funcionamento da classificação, como mostra a Tabela 17. Primarina, que é comum, e Lunala,

lendária, foram classificadas corretamente, entretanto Kommo-o, comum, foi classificado incorretamente, tendo sua probabilidade estimada bem mais alta que muitos lendários. Essa probabilidade estimada pode ser entendida também como um critério de escolha para o jogador, por exemplo, se ele tem uma vaga disponível em sua equipe e ele quer escolher entre Primarina ou Kommo-o, usando a regressão logística, ele poderia optar por Kommo-o já que, mesmo sendo comum, ele é tão forte quanto muitos lendários.

Nome	Probabilidade	Classificação	Raridade
Primarina	0,0250	Comum	Comum
Kommo-o	0,5314	Lendário	Comum
Lunala	0,9696	Lendário	Lendário

Tabela 17: Probabilidade estimada, classificação e a raridade verdadeira de três *Pokémon* adicionais.

Conclusão

Neste trabalho foi possível verificar as estatísticas que mais influenciam na probabilidade de um *Pokémon* ser lendário ou não a partir de uma regressão logística cuja finalidade é de estimar tal probabilidade, que foi igual a 0,1461, e também servir de critério de escolha para jogadores que estejam com dúvidas na escolha de sua equipe.

Referências

- Notas de aula;
- Link: https://web.stanford.edu/class/archive/stats/stats200/stats200.1172/Lecture26.pdf, acesso em 03/12/2020;
- Link: https://www.ime.usp.br/~giapaula/envel_bino, acesso em 03/12/2020.
- Link: https://www.ime.usp.br/~giapaula/diag_bino_bino, acesso em 03/12/2020.
- Link: https://www.kaggle.com/abcsds/pokemon, acesso em 03/12/2020;
- Link: https://bulbapedia.bulbagarden.net/wiki/Primarina, acesso em 03/12/2020;
- Link: https://bulbapedia.bulbagarden.net/wiki/Kommo-o, acesso em 03/12/2020;
- Link: https://bulbapedia.bulbagarden.net/wiki/Lunala, acesso em 03/12/2020.

Anexos

```
## Carregando os pacotes
library(corrplot)
library(dplyr)
```

```
library(xtable)
library(mgcv)
library(pROC)

## Carregando o banco de dados
pokemon=read.csv("Pokemon.csv")

## Escolhendo as variáveis
status=pokemon%>%
select(HP,Attack,Defense,Sp..Atk,Sp..Def,Speed)
status=apply(status,2,function(x){tapply(x,pokemon$X.,mean)})
lendario=ifelse(pokemon$Legendary=="True",1,0)
lendario=tapply(lendario,pokemon$X.,mean)
colnames(status)=c("Vida","Ataque","Defesa","Atq. Esp.",
"Def. Esp.","Veloc.")
status=as.data.frame(status)


## Gráficos de pares e correlação
names(status)=c("Vida","Ataque","Defesa","Atq. Esp.","Def. Esp.","Veloc.")
pairs(status,col=ifelse(lendario==0,"red","blue"),oma=c(3,3,3,15))
par(xpd = TRUE)
legend("bottomright",fill=c("red","blue"),legend=c("Comum","Lendário"))
cors=cor(status)
corrplot.mixed(cors)

## Boxplots
boxplot(status$Vida~lendario,xlab="Raridade",ylab="Vida",
names=c("Comum","Lendário"),ylim=c(0,255),col="red")
boxplot(status$Ataque~lendario,xlab="Raridade",ylab="Ataque",
names=c("Comum","Lendário"),ylim=c(0,255),col="darkorange2")
boxplot(status$Defesa~lendario,xlab="Raridade",ylab="Defesa",
names=c("Comum","Lendário"),ylim=c(0,255),col="gold1")
boxplot(status$`Atq. Esp.`~lendario,xlab="Raridade",ylab="Ataque Especial",
names=c("Comum","Lendário"),ylim=c(0,255),col="dodgerblue2")
boxplot(status$`Def. Esp.`~lendario,xlab="Raridade",ylab="Defesa Especial",
names=c("Comum","Lendário"),ylim=c(0,255),col="green3")
boxplot(status$Veloc.~lendario,xlab="Raridade",ylab="Velocidade",
names=c("Comum","Lendário"),ylim=c(0,255),col="deeppink1")

## Análises individuais
loginv=function(x){
exp(x)/(1+exp(x))
}
attach(status)
fitnulo=glm(lendario~1,family=binomial)

### Vida
lmt.cls <- c(min(status$Vida)-0.5,seq(min(status$Vida[lendario==1]),
max(status$Vida),len=5))
classes <- cut(status$Vida, breaks=lmt.cls)
nn <- tapply(lendario, classes, length)
pt.med <- tapply(status$Vida, classes, mean)
yy <- round(tapply(lendario, classes, sum), 2)
```

```
prop <- round(tapply(lendario, classes, mean), 2)
logito <- log((yy + 0.5)/(nn - yy + 0.5))

fitVida=glm(lendario~Vida,family=binomial)
xtable(summary(fitVida))
xtable(anova(fitnulo,fitVida,test="Chisq"))

plot(pt.med, logito, xlab="Vida",
ylab="Logito da probabilidade de ser lendário", pch=16, cex=1.2)
res <- gam(lendario ~ Vida, family=binomial(link=logit),data=status)
xvals <- seq(min(status$Vida), 255, .1)
pvals <- predict(res, newdata=data.frame(Vida=xvals), type="link",se.fit=T)
lines(x=xvals, y=pvals$fit, lty=2)
lines(x=xvals, y=pvals$fit-1.96*pvals$se, col="blue")
lines(x=xvals, y=pvals$fit+1.96*pvals$se, col="blue")
legend("bottomright",legend=c("Estimativa Pontual",
"Limites (95% de confiança)"),
col=c("black","blue"),lty=2)

plot(pt.med, prop, xlab="Vida",
ylab="Probabilidade de ser lendário", pch=16, cex=1.2, ylim=c(0,1))
lines(x=xvals, y=loginv(pvals$fit), lty=2)
lines(x=xvals, y=loginv(pvals$fit-1.96*pvals$se), lty=2, col="blue")
lines(x=xvals, y=loginv(pvals$fit+1.96*pvals$se), lty=2, col="blue")
legend("topleft",legend=c("Estimativa Pontual",
"Limites (95% de confiança)"),
col=c("black","blue"),lty=2)

### Ataque
lmt.cls <- c(min(status$Ataque)-0.5,seq(min(status$Ataque[lendario==1]),max(status$Ataque),len=5))
classes <- cut(status$Ataque, breaks=lmt.cls)
nn <- tapply(lendario, classes, length)
pt.med <- tapply(status$Ataque, classes, mean)
yy <- round(tapply(lendario, classes, sum), 2)
prop <- round(tapply(lendario, classes, mean), 2)
logito <- log((yy + 0.5)/(nn - yy + 0.5))

fitataque=glm(lendario~Ataque,family=binomial)
xtable(summary(fitataque))
xtable(anova(fitnulo,fitataque,test="Chisq"))
plot(pt.med, logito, xlab="Ataque",
     ylab="Logito da probabilidade de ser lendário", pch=16, cex=1.2)
res <- gam(lendario ~ Ataque, family=binomial(link=logit),data=status)
xvals <- seq(min(status$Ataque), 255, .1)
pvals <- predict(res, newdata=data.frame(Ataque=xvals), type="link",se.fit=T)
lines(x=xvals, y=pvals$fit, lty=2)
lines(x=xvals, y=pvals$fit-1.96*pvals$se, col="blue")
lines(x=xvals, y=pvals$fit+1.96*pvals$se, col="blue")
legend("bottomright",legend=c("Estimativa Pontual","Limites (95% de confiança)"),
col=c("black","blue"),lty=2)

plot(pt.med, prop, xlab="Ataque",
     ylab="Probabilidade de ser lendário", pch=16, cex=1.2, ylim=c(0,1))
lines(x=xvals, y=loginv(pvals$fit), lty=2)
```

```
lines(x=xvals, y=loginv(pvals$fit-1.96*pvals$se), lty=2, col="blue")
lines(x=xvals, y=loginv(pvals$fit+1.96*pvals$se), lty=2, col="blue")
legend("topleft",legend=c("Estimativa Pontual","Limites (95% de confiança)"),
col=c("black","blue"),lty=2)

### Defesa
lmt.cls <- c(min(status$Defesa)-0.5,seq(min(status$Defesa[lendario==1]),max(status$Defesa),len=5))
classes <- cut(status$Defesa, breaks=lmt.cls)
nn <- tapply(lendario, classes, length)
pt.med <- tapply(status$Defesa, classes, mean)
yy <- round(tapply(lendario, classes, sum), 2)
prop <- round(tapply(lendario, classes, mean), 2)
logito <- log((yy + 0.5)/(nn - yy + 0.5))

fitdefesa=glm(lendario~Defesa,family=binomial)
xtable(summary(fitdefesa))
xtable(anova(fitnulo,fitdefesa,test="Chisq"))

plot(pt.med, logito, xlab="Defesa",
     ylab="Logito da probabilidade de ser lendário", pch=16, cex=1.2)
res <- gam(lendario ~ Defesa, family=binomial(link=logit),data=status)
xvals <- seq(min(status$Defesa), 255, .1)
pvals <- predict(res, newdata=data.frame(Defesa=xvals), type="link",se.fit=T)
lines(x=xvals, y=pvals$fit, lty=2)
lines(x=xvals, y=pvals$fit-1.96*pvals$se, col="blue")
lines(x=xvals, y=pvals$fit+1.96*pvals$se, col="blue")
legend("bottomright",legend=c("Estimativa Pontual","Limites (95% de confiança)"),
col=c("black","blue"),lty=2)

plot(pt.med, prop, xlab="Defesa",
     ylab="Probabilidade de ser lendário", pch=16, cex=1.2, ylim=c(0,1))
lines(x=xvals, y=loginv(pvals$fit), lty=2)
lines(x=xvals, y=loginv(pvals$fit-1.96*pvals$se), lty=2, col="blue")
lines(x=xvals, y=loginv(pvals$fit+1.96*pvals$se), lty=2, col="blue")
legend("topleft",legend=c("Estimativa Pontual","Limites (95% de confiança)"),
col=c("black","blue"),lty=2)

### Ataque Especial
lmt.cls <- c(min(status$`Atq. Esp.`)-0.5,
             seq(min(status$`Atq. Esp.`[lendario==1]),max(status$`Atq. Esp.`),len=5))
classes <- cut(status$`Atq. Esp.`, breaks=lmt.cls)
nn <- tapply(lendario, classes, length)
pt.med <- tapply(status$`Atq. Esp.`, classes, mean)
yy <- round(tapply(lendario, classes, sum), 2)
prop <- round(tapply(lendario, classes, mean), 2)
logito <- log((yy + 0.5)/(nn - yy + 0.5))

atksp=`Atq. Esp.`

fitatksp=glm(lendario~atksp,family=binomial)
xtable(summary(fitatksp))
xtable(anova(fitnulo,fitatksp,test="Chisq"))

plot(pt.med, logito, xlab="Ataque Especial",
```

```
     ylab="Logito da probabilidade de ser lendário", pch=16, cex=1.2)
res <- gam(lendario ~ atksp, family=binomial(link=logit))
xvals <- seq(min(status$`Atq. Esp.`), 255, .1)
pvals <- predict(res, newdata=data.frame(atksp=xvals), type="link",se.fit=T)
lines(x=xvals, y=pvals$fit, lty=2)
lines(x=xvals, y=pvals$fit-1.96*pvals$se, col="blue")
lines(x=xvals, y=pvals$fit+1.96*pvals$se, col="blue")
legend("bottomright",legend=c("Estimativa Pontual","Limites (95% de confiança)"),
col=c("black","blue"),lty=2)

plot(pt.med, prop, xlab="Ataque Especial",
     ylab="Probabilidade de ser lendário", pch=16, cex=1.2, ylim=c(0,1))
lines(x=xvals, y=loginv(pvals$fit), lty=2)
lines(x=xvals, y=loginv(pvals$fit-1.96*pvals$se), lty=2, col="blue")
lines(x=xvals, y=loginv(pvals$fit+1.96*pvals$se), lty=2, col="blue")
legend("topleft",legend=c("Estimativa Pontual","Limites (95% de confiança)"),
col=c("black","blue"),lty=2)

### Defesa Especial
lmt.cls <- c(min(status$`Def. Esp.`)-0.5,
             seq(min(status$`Def. Esp.`[lendario==1]),max(status$`Def. Esp.`),len=5))
classes <- cut(status$`Def. Esp.`, breaks=lmt.cls)
nn <- tapply(lendario, classes, length)
pt.med <- tapply(status$`Def. Esp.`, classes, mean)
yy <- round(tapply(lendario, classes, sum), 2)
prop <- round(tapply(lendario, classes, mean), 2)
logito <- log((yy + 0.5)/(nn - yy + 0.5))


defsp=`Def. Esp.`
fitdefsp=glm(lendario~defsp,family=binomial)
xtable(summary(fitdefsp))
xtable(anova(fitnulo,fitdefsp,test="Chisq"))

plot(pt.med, logito, xlab="Defesa Especial",
     ylab="Logito da probabilidade de ser lendário", pch=16, cex=1.2)
res <- gam(lendario ~ defsp, family=binomial(link=logit))
xvals <- seq(min(status$`Def. Esp.`), 255, .1)
pvals <- predict(res, newdata=data.frame(defsp=xvals), type="link",se.fit=T)
lines(x=xvals, y=pvals$fit, lty=2)
lines(x=xvals, y=pvals$fit-1.96*pvals$se, col="blue")
lines(x=xvals, y=pvals$fit+1.96*pvals$se, col="blue")
legend("bottomright",legend=c("Estimativa Pontual","Limites (95% de confiança
col=c("black","blue"),lty=2)

plot(pt.med, prop, xlab="Defesa Especial",
     ylab="Probabilidade de ser lendário", pch=16, cex=1.2, ylim=c(0,1))
lines(x=xvals, y=loginv(pvals$fit), lty=2)
lines(x=xvals, y=loginv(pvals$fit-1.96*pvals$se), lty=2, col="blue")
lines(x=xvals, y=loginv(pvals$fit+1.96*pvals$se), lty=2, col="blue")
legend("topleft",legend=c("Estimativa Pontual","Limites (95% de confiança)"),
col=c("black","blue"),lty=2)

### Velocidade
```

```
lmt.cls <- c(min(status$Veloc.)-0.5,
             seq(min(status$Veloc.[lendario==1]),max(status$Veloc.),len=5))
classes <- cut(status$Veloc., breaks=lmt.cls)
nn <- tapply(lendario, classes, length)
pt.med <- tapply(status$Veloc., classes, mean)
yy <- round(tapply(lendario, classes, sum), 2)
prop <- round(tapply(lendario, classes, mean), 2)
logito <- log((yy + 0.5)/(nn - yy + 0.5))

fitveloc=glm(lendario~Veloc.,family=binomial)
xtable(summary(fitveloc))
xtable(anova(fitnulo,fitveloc,test="Chisq"))

plot(pt.med, logito, xlab="Velocidade",
     ylab="Logito da probabilidade de ser lendário", pch=16, cex=1.2)
res <- gam(lendario ~ Veloc., family=binomial(link=logit))
xvals <- seq(min(status$Veloc.), 255, .1)
pvals <- predict(res, newdata=data.frame(Veloc.=xvals), type="link",se.fit=T)
lines(x=xvals, y=pvals$fit, lty=2)
lines(x=xvals, y=pvals$fit-1.96*pvals$se, col="blue")
lines(x=xvals, y=pvals$fit+1.96*pvals$se, col="blue")
legend("bottomright",legend=c("Estimativa Pontual","Limites (95% de confiança)"),
col=c("black","blue"),lty=2)

plot(pt.med, prop, xlab="Velocidade",
     ylab="Probabilidade de ser lendário", pch=16, cex=1.2, ylim=c(0,1))
lines(x=xvals, y=loginv(pvals$fit), lty=2)
lines(x=xvals, y=loginv(pvals$fit-1.96*pvals$se), lty=2, col="blue")
lines(x=xvals, y=loginv(pvals$fit+1.96*pvals$se), lty=2, col="blue")
legend("topleft",legend=c("Estimativa Pontual","Limites (95% de confiança)"),
col=c("black","blue"),lty=2)

## Modelo
fit=glm(lendario~.,family=binomial,data=status)
fit2=glm(lendario~.-Ataque,family=binomial,data=status)
summary(fit)
anova(fit2,fit,test="Chisq")
anova(fitnulo,fit,test="Chisq")
xtable(summary(fit))
xtable(anova(fit2,fit,test="Chisq"))
xtable(anova(fitnulo,fit,test="Chisq"))

## Análise de Resíduos
diag_bino=function(fit.model){
  X <- model.matrix(fit.model)
  n <- nrow(X)
  p <- ncol(X)
  w <- fit.model$weights
  W <- diag(w)
  H <- solve(t(X)%*%W%*%X)
  H <- sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
  h <- diag(H)
  ts <- resid(fit.model,type="pearson")/sqrt(1-h)
  td <- resid(fit.model,type="deviance")/sqrt(1-h)
```

```
  di <- (h/(1-h))*(ts^2)
  a <- max(td)
  b <- min(td)
  par(mfrow=c(2,2))
  plot(fitted(fit.model),h,xlab="Valor Ajustado",
       ylab="Medida h", pch=16)
  identify(fitted(fit.model), h, n=1)
  #
  plot(di,xlab="Índice", ylab="Distância de Cook",pch=16)
  #identify(di, n=1)
  #
  plot(td,xlab="Índice", ylab="Resíduo Componente do Desvio",
       ylim=c(b-1,a+1), pch=16)
  abline(2,0,lty=2)
  abline(-2,0,lty=2)
  identify(td, n=4)
  #
  plot(fitted(fit.model), td,xlab="Valor Ajustado",
       ylab="Resíduo Componente do Desvio", pch=16)
  #identify(fitted(fit.model), td, n=1)
  par(mfrow=c(1,1))
}
diag_bino(fit)

## Envelope
envel_bino=function(fit.model){
  par(mfrow=c(1,1))
  X <- model.matrix(fit.model)
  n <- nrow(X)
  p <- ncol(X)
  w <- fit.model$weights
  W <- diag(w)
  H <- solve(t(X)%*%W%*%X)
  H <- sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
  h <- diag(H)
  td <- resid(fit.model,type="deviance")/sqrt(1-h)
  e <- matrix(0,n,100)
  #
  for(i in 1:100){
    dif <- runif(n) - fitted(fit.model)
    dif[dif >= 0 ] <- 0
    dif[dif<0] <- 1
    nresp <- dif
    fit <- glm(nresp ~ X, family=binomial)
    w <- fit$weights
    W <- diag(w)
    H <- solve(t(X)%*%W%*%X)
    H <- sqrt(W)%*%X%*%H%*%t(X)%*%sqrt(W)
    h <- diag(H)
    e[,i] <- sort(resid(fit,type="deviance")/sqrt(1-h))}
  #
  e1 <- numeric(n)
  e2 <- numeric(n)
  #
```

```
  for(i in 1:n){
    eo <- sort(e[i,])
    e1[i] <- (eo[2]+eo[3])/2
    e2[i] <- (eo[97]+eo[98])/2}
  #
  med <- apply(e,1,mean)
  faixa <- range(td,e1,e2)
  par(pty="s")
  qqnorm(td,xlab="Percentil da N(0,1)",
         ylab="Componente do Desvio", ylim=faixa, main="")
  #
  par(new=TRUE)
  #
  qqnorm(e1,axes=F,xlab="",ylab="",type="l",ylim=faixa,lty=1, main="",col="blue")
  par(new=TRUE)
  qqnorm(e2,axes=F,xlab="",ylab="", type="l",ylim=faixa,lty=1, main="",col="blue")
  par(new=TRUE)
  qqnorm(med,axes=F,xlab="", ylab="", type="l",ylim=faixa,lty=2, main="")

}
envel_bino(fit)

## Curva ROC
prob=predict(fit,type="response")
croc=roc(lendario~prob)
plot(croc,xlab="Especificidade",ylab="Sensitividade")
polygon(c(croc$specificities,0,0),c(croc$sensitivities,0,1),
        col="aquamarine")
polygon(c(0,1,1,0),c(0,0,1,1))
abline(1,-1)
text(0.7,0.8,"Área sobre a\n curva: 0,9871")

## Tabela para usar o critério de classificação
classificacao=function(prob){
ifelse(fitted(fit)>prob,1,0)
}

## Encontrar o valor que deixa as taxas de acertos iguais
class2=function(prob){
sum(diag(table(classificacao(prob),lendario))/c(735,-65))
}
uniroot(class2,c(0,0.2))
xtable(table(classificacao(0.1461),lendario))
```