Capítulo

11

Modelos de Regressão aplicados em pesquisas em Informática na Educação

Danilo Alvares
Harvard T.H. Chan School of Public Health
dalvares@hsph.harvard.edu

Objetivo do Capítulo

Este capítulo tem o objetivo de apresentar os modelos de regressão linear simples (MRLS) e múltipla (MRLM), introduzir os modelos lineares generalizados (MLG), discutir as abordagens inferenciais clássica e Bayesiana, e ilustrar um modelo ajustado a partir de um banco de dados real. Como material complementar, também é fornecido algumas das principais referências nestes tópicos e uma lista de exercícios. Ao final da leitura deste capítulo, você deve ser capaz de:

- Diferenciar os modelos de regressão apresentados.
- Identificar qual modelagem é mais apropriada em uma análise de regressão.
- Entender as principais distinções entre os paradigmas clássico e Bayesiano.
- Propor e ajustar um modelo de regressão da classe MLG (via estimação clássica ou Bayesiana) usando funções da linguagem R e interpretar os resultados.

Era uma vez... *Jessica é doutoranda no Programa de Informática na Educação e sua pesquisa está relacionada com o desenvolvimento de um APP para praticar o raciocínio verbal (capacidade de compreender e fundamentar o uso de conceitos expressos através de palavras). O APP é composto por 9 exercícios de múltipla escolha, sem distinção de nível de dificuldade, envolvendo raciocínio verbal. Jessica gostaria de saber se o desempenho na resolução dos exercícios está relacionado com o sexo do indivíduo e seu conhecimento prévio em raciocínio verbal. Para analisar essa relação, Jessica dispõe de uma base de dados com 3435 estudantes de 16 anos do Ensino Médio que testaram seu APP. Além disso, ela também tem acesso ao sexo de cada estudante e seu desempenho em uma prova de raciocínio verbal ao iniciar o Ensino Médio. No entanto, Jessica não sabe exatamente que tipo de análise estatística usar para avaliar as relações de interesse em seu estudo. Será que podemos ajudá-la?*

1 Introdução

Modelos de regressão estão amplamente presentes nas mais variadas áreas do conhecimento. A razão de sua popularidade é simples: pesquisadores experimentais sempre buscam uma expressão que relaciona satisfatoriamente a variável resposta de seus estudos (também denominada variável dependente) com informações adicionais coletadas (variáveis independentes/explicativas/regressoras), possibilitando interpretar os efeitos relacionais e também fazer predições. Tomando como exemplo o estudo apresentado na Seção "Era uma vez", nele a variável resposta é o desempenho de cada estudante na resolução dos exercícios usando o APP de Jessica e as variáveis independentes são o sexo e o conhecimento prévio em raciocínio verbal destes estudantes.

Ainda que grande parte das pesquisas científicas utilizam modelos estatísticos, é importante ter sempre em mente que os modelos são, inevitavelmente, simplificações da realidade, pois há inúmeros fatores que não são observados ou não podem ser controlados, além de erros de medições. No entanto, muitos modelos de regressão são plausíveis de representar adequadamente o comportamento de algum fenômeno envolvendo apenas um número limitado de variáveis explicativas. Motivado por esta ideia, estudaremos nas próximas seções alguns dos principais modelos de regressão e também os métodos de estimação clássico e Bayesiano.

2 Modelo de Regressão Linear Simples (MRLS)

O *modelo de regressão linear simples* (MRLS) foi, por conta da sua simplicidade, um dos primeiros modelos estatísticos largamente disseminado. Ainda hoje, este modelo é bastante útil tanto para fins aplicados quanto didáticos. Seu objetivo é relacionar uma variável resposta (em teoria, no conjunto dos reais) com uma variável explicativa através da equação de uma reta (de regressão), que pode ser descrita em um gráfico XY. Mais formalmente, a representação matemática do MRLS é dada por:

$$Y_i = \beta_0 + \beta_1 \mathrm{X}_i + \epsilon_i,$$

onde o par (Y_i, X_i) indica os valores observados das variáveis resposta e explicativa, respectivamente, para o indivíduo i, onde genericamente $i = 1, \ldots, n$. Os parâmetros β_0 e β_1 são desconhecidos e são eles que caracterizam onde a reta de regressão cruza o eixo vertical, também chamado de intercepto (β_0), e qual o coeficiente angular, também chamado inclinação da reta (β_1). Por fim, ϵ_i é uma variável aleatória residual (também conhecida como erro aleatório ou estocástico) que inclui todas as influências no comportamento de Y_i que não podem ser explicadas linearmente pelos valores de X_i. Vale ressaltar que X_i é uma variável fixa (ou seja, seu valor não altera mediante observações sucessivas), enquanto Y_i, por construção, é uma variável aleatória (ou seja, depende de fatores aleatórios). Além disso, alguns pressupostos são necessários para descrever o MRLS de forma completa:

i. Todos os pares de observações são independentes, ou seja, (Y_i, X_i) é independente de (Y_j, X_j) para quaisquer $i \neq j$;

ii. Há uma relação linear entre a variável resposta e a variável explicativa;

iii. O erro tem média zero, $E(\epsilon_i) = 0$, e variância (desconhecida) σ^2, $Var(\epsilon_i) = \sigma^2$.

Note que a suposição **i.** é algo prévio à análise de regressão, mais especificamente, trata-se de um pressuposto que deve ser considerado na fase de planejamento de experimento e coleta de dados. Por outro lado, **ii.** faz parte de uma análise descritiva dos dados (análise de correlação), onde este pressuposto pode ser verificado empiricamente através de um *diagrama de dispersão* (gráfico XY com os valores, em pares, observados) e, em caso de que haja uma tendência linear visível entre as variáveis resposta e explicativa, também podemos quantificar a "força" desta relação linear por meio do *coeficiente de correlação de Pearson* (na linguagem R, `cor(X,Y,method="pearson")`). Este coeficiente, também chamado de *coeficiente de correlação linear* ou *r de Pearson*, toma valores entre -1 e 1, onde:

- Valores próximos a -1 indicam uma forte correlação linear negativa entre X e Y;
- Valores próximos a 1 indicam uma forte correlação linear positiva entre X e Y;
- Valores próximos a 0 indicam uma fraca correlação linear entre X e Y.

Rigorosamente, o coeficiente de correlação de Pearson só pode ser calculado mediante algumas características de X e Y, onde a principal delas é que ambas sejam variáveis contínuas. Caso ao menos uma das duas variáveis seja ordinal ou categórica (por exemplo, X indica se o indivíduo acertou uma determinada questão e, portanto, X pode valer 0 (errou) ou 1 (acertou)) podemos utilizar o *coeficiente de correlação de Spearman* (na linguagem R, `cor(X,Y,method="spearman")`), onde seu intervalo de valores e intepretação da "força" de correlação são análogos ao coeficiente de Pearson.

Diferentemente dos dois primeiros pressupostos, **iii.** só pode ser checado após o ajuste do modelo, uma vez que necessitamos das estimativas de β_0 e β_1 para calcular os resíduos ($\epsilon_i = Y_i - \beta_0 - \beta_1 X_i$, para $i = 1, \ldots, n$). A primeira suposição em **iii.**, $E(\epsilon_i) = 0$, pode ser verificada através de um gráfico cruzando a variável explicativa X_i e os resíduos ϵ_i, para $i = 1, \ldots, n$, na qual o pressuposto é cumprido quando os pontos no gráfico estão aproximadamente distribuídos ao redor de zero. Outra possibilidade é calcular a média (ou a soma) dos erros e se o valor resultante é próximo a zero, a condição é satisfeita.

Ainda em **iii.**, temos a suposição de $Var(\epsilon_i) = \sigma^2$, tecnicamente denominada *homogeneidade da variância* (ou simplesmente *homocedasticidade*). A ideia por detrás deste pressuposto é que os resíduos têm variância comum e também podem ser empiricamente analisados com um gráfico de X_i por ϵ_i, para $i = 1, \ldots, n$. Para que a condição de homocedasticidade seja cumprida, os pontos deste gráfico, aleatoriamente distribuídos em torno de zero, não podem apresentar comportamento oscilatório ou alguma tendência. Tipicamente, pontos distribuídos em um formato de "funil" é o que viola esta condição, caracterizando a *heterocedasticidade* (resíduos não homocedástico). O teste de Breusch-Pagan (hipótese nula: homogeneidade da variância) é a principal ferramenta para checar a homocedasticidade e, na linguagem R, é obtido através da função `bptest` presente na biblioteca `lmtest`. Ainda que haja violação da homocedasticidade podemos utilizar um MRLS, mas incorporando uma estrutura mais complexa para a variância residual e empregando métodos de estimação ponderados.

Embora **i.**, **ii.** e **iii.** são as únicas suposições do MRLS, também é necessário definir alguma distribuição para os erros residuais. Comumente, assumimos que eles seguem uma distribuição Normal, ou seja, $\epsilon_i \sim Normal(0, \sigma^2)$, onde a média zero e a variância σ^2 garantem o pressuposto **iii.**. Ao incorporar a suposição de normalidade dos resíduos, checar sua validade após o ajuste do modelo passa a ser obrigatório. No entanto, há ferramentas gráficas e testes de significância específicos para avaliar se uma amostra é aproximadamente Normal. Estes recursos serão detalhadamente explorados na Seção "Cenário Ilustrativo".

Uma vez que temos todos os pressupostos cumpridos, finalmente podemos interpretar o modelo ajustado. Tal como definimos inicialmente, β_0 corresponde ao intercepto da reta de regressão e ele representa o valor médio da variável resposta Y_i quando X_i é igual a zero. Em algumas aplicações, a variável independente X_i só toma valores estritamente positivos (ou negativos) e, portanto, β_0 não pode ser diretamente interpretado. Por outro lado, o parâmetro de inclinação da reta β_1 é sempre interpretável, na qual ele nos diz quão sensível é Y_i ao variarmos X_i. Mais especificamente, ao aumentarmos uma unidade da variável independente X_i, a média da variável resposta Y_i aumenta β_1 unidades.

3 Modelo de Regressão Linear Múltipla (MRLM)

Na maioria das aplicações, há mais que uma variável explicativa envolvida no estudo. Em particular, quando a variável resposta é, teoricamente, contínua e definida no conjunto dos reais, podemos estender o MRLS a um *modelo de regressão linear múltipla* (MRLM), onde o termo "múltipla" faz referência a duas ou mais variáveis explicativas. A representação matemática desta modelagem, supondo q variáveis explicativas, é dada por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_q X_{qi} + \epsilon_i = \beta_0 + \sum_{j=1}^{q} \beta_j X_{ji} + \epsilon_i,$$

onde X_{ji} é uma variável explicativa, referente ao indivíduo i, com respectivo parâmetro de regressão β_j, para $j = 1, \ldots, q$. Análogo ao MRLS, Y_i representa a variável resposta, β_0 caracteriza o intercepto e ϵ_i é uma variável aleatória residual. Tipicamente, o MRLM é descrito na seguinte forma matricial:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon},$$

onde $\boldsymbol{Y} = (Y_1, \ldots, Y_n)'$, $\boldsymbol{X} = (\boldsymbol{1_n}, \boldsymbol{X_1}, \ldots, \boldsymbol{X_q})$ sendo $\boldsymbol{1_n} = (1, \ldots, 1)'$ e $\boldsymbol{X_j} = (X_{j1}, \ldots, X_{jn})'$ com $j = 1, \ldots, q$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_q)'$ e $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$.

Os pressupostos para o MRLM, assim como seus mecanismos de validação, são basicamente os mesmos que na abordagem MRLS. Adicionalmente, o número de observações deve ser maior que o número de parâmetros de regressão, ou seja, $n > q$. Além disso, também é comum assumir uma distribuição normal para os erros residuais.

A interpretação dos parâmetros de regressão $\boldsymbol{\beta}$ são equivalentes as apresentadas no MRLS. No entanto, algumas considerações adicionais devem ser levadas em conta. A primeira delas é que agora não temos mais uma relação bidimensional (uma variável resposta *versus* uma variável explicativa), onde β_0 visivelmente é o intercepto no eixo y

quando a variável explicativa é zero. A extensão natural desta interpretação é que no MRLM todas as variáveis explicativas $X_{1i}, \ldots, X_{qi}$ de um indivíduo i devem ser zero para que β_0 seja visto como intercepto. Análogo ao MRLS, em alguns casos, β_0 não é interpretado. Relembrando que agora estamos em um espaço multidimensional, os outros parâmetros da regressão, $\beta_1, \ldots, \beta_q$, não representam diretamente a inclinação da reta em um gráfico XY, mas são interpretados de forma semelhante ao MRLS. Concretamente, ao aumentarmos uma unidade da variável independente X_{ji} e mantendo as demais variáveis fixas, a média da variável resposta Y_i aumenta β_j unidades.

4 Modelos Lineares Generalizados (MLG)

Durante muito tempo os MRLS e MRLM foram as únicas opções de modelos estatísticos para descrever fenômenos aleatórios. Naquela época, quando a suposição de normalidade dos resíduos não era satisfeita, a única alternativa era transformar a variável resposta (por exemplo, usando a transformação de Box e Cox) e checar se o novo ajuste continha resíduos normalmente distribuídos.

Contudo, não era difícil encontrar situações em que a normalidade dos resíduos não era alcançada com transformações ou mesmo quando a própria natureza da variável resposta (por exemplo, discreta ou binária) não satisfazia adequadamente a suposição de um MRLS/MRLM. Estas limitações foram sanadas através dos *modelos lineares generalizados* (MLG), propostos por Nelder e Wedderburn (1972), na qual fazem parte uma extensa e flexível gama de modelos.

Por definição, o modelo para a variável (resposta) aleatória Y é dito ser um MLG se sua distribuição pertence à *família exponencial de distribuições*. Em outras palavras, a distribuição de Y pode ser escrita na forma:

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

onde θ e ϕ são parâmetros escalares, $b(\cdot)$ e $c(\cdot)$ representam funções diferenciáveis e conhecidas (ver Paula (2013) para mais detalhes). Esta classe de modelos é caracterizada por três elementos:

- **Componente aleatória**: variável resposta (Y) com distribuição pertencente à família exponencial;
- **Componente sistemática** (ou **estrutural**): variáveis explicativas ($\boldsymbol{X}$) definidas por um preditor linear, ou seja, $\eta = \boldsymbol{X\beta}$;
- **Função de ligação**: conecta as componentes aleatória e sistemática. Esta relação é feita através do valor esperado (μ) da variável resposta (Y) e o preditor linear (η), ou seja, $g(\mu) = \eta$.

A grande vantagem desta estrutura genérica que define os MLG é a flexibilidade para modelar diferentes tipos de variáveis resposta, onde nossa proposta de modelo já leva em consideração as características originais do problema. Por exemplo, com os MLG estamos aptos a modelar variáveis respostas com valores discretos, positivos, binários, categóricos, etc.. Para ilustrar alguns destes cenários, a Tabela 1 apresenta os

principais MLG, como são definidas suas funções de ligação mais populares e também uma breve descrição de quando cada um destes modelos deve ser aplicado:

Modelo	**Função de Ligação** $g(\mu)$	**Característica**
`Normal`	μ	Variável resposta no conjunto dos reais (MRLS/MRLM).
`Binomial`	$\log[\mu/(1-\mu)]$	Variável resposta (Y) representa o número de eventos bem (ou mal) sucedidos em um total de m eventos independentes. Equivalentemente, Y também pode ser visto como uma proporção de sucessos (ou falhas). Quando m=1 (ou seja, a variável resposta é binária), temos um modelo `Bernoulli`.
`Poisson`	$\log(\mu)$	Variável resposta (Y) assume valores discretos e não negativos, ou seja, dados de contagem $(0,1,2,3,...)$. Este modelo pressupõe que a média e a variância de Y são (aproximadamente) iguais. Caso esta restrição não seja cumprida, uma alternativa é migrar para a modelagem `Binomial Negativa` (que também é um MLG).
`Gamma`	$1/\mu$	Variável resposta assimétrica com valores positivos e contínuos.

Tabela 1: Exemplos de modelos lineares generalizados (MLG).

Embora as funções de ligação sejam obtidas naturalmente por construção, há também formulações alternativas que conectam satisfatoriamente a média (μ) com o preditor linear (η). Por exemplo, o modelo Binomial, além da ligação logito/logística descrita acima, também admite as funções de ligação probito e complemento log-log.

Um ponto fraco dos MLG é a exigência de erros aleatórios independentes. Dados longitudinais ou espaciais são exemplos que não satisfazem esta restrição, pois há uma estrutura de correlação inerente entre as medidas. No entanto, esta limitação pode ser contornada utilizando *modelos lineares generalizados mistos* ou *equações de estimação generalizadas*. Infelizmente, ambas as abordagens não serão vistas neste capítulo.

Esquema Básico para Modelagem Estatística

		VARIÁVEL RESPOSTA	
		Discreta	**Contínua**
VARIÁVEL EXPLICATIVA	**Discreta**	MLG (e.g., modelos Binomial e Poisson) Tabela cruzada Teste de proporção (e.g., teste Qui-quadrado)	MLG (e.g., MRLS/MRLM) ANOVA Teste de médias (e.g., test t de Student)
	Contínua	MLG (e.g., modelos Binomial e Poisson)	MLG (e.g., MRLS/MRLM)

5 Métodos de Estimação

Há apenas um caminho para tirarmos conclusões acertadas sobre os parâmetros de um modelo estatístico: aumentar o conhecimento (ou seja, inferir) sobre estes parâmetros a partir das informações disponíveis. Embora haja diferentes propostas para realizar esta tarefa, focaremos nas duas abordagens inferenciais mais utilizadas na literatura atual, denominadas *clássica* (também conhecida como *frequentista* ou *por máxima verossimilhança*) e *Bayesiana*.

Filosoficamente, as diferenças entre os paradigmas clássico e Bayesiano são extremas, mas ambas concordam que a principal fonte de informação para aprender sobre os parâmetros são os dados, tendo como quantificador a função de *verossimilhança* (as vezes, também chamada de *plausibilidade*). A verossimilhança varia de acordo com o modelo escolhido para a variável resposta, além disso, ela é uma função dos parâmetros do modelo dado todas as observações (variáveis resposta e explicativas).

A abordagem clássica afirma que os parâmetros são valores fixos (constantes), mas desconhecidos. Com isso, a forma de estimar estes valores é maximizando a verossimilhança, ou seja, encontrar os valores dos parâmetros que conjuntamente melhor explicam os dados observados. Em contrapartida, a abordagem Bayesiana considera todas as quantidades desconhecidas como aleatórias e, portanto, são caracterizadas através de uma distribuição de probabilidade, denominada *distribuição a priori*. A partir do *Teorema de Bayes*, a inferência Bayesiana é descrita por um processo de aprendizagem sobre os parâmetros do modelo, onde as fontes de informações são a função de verossimilhança e a distribuição *a priori*, e o resultado final também é uma distribuição de probabilidade, denominada *distribuição a posteriori*. Em uma forma simplificada, o que temos é:

distribuição a posteriori proporcional a ***verossimilhança*** x ***distribuição a priori***

Em muitos contextos, a expressão acima não é analiticamente manuseável, sendo necessário o uso de algum método computacional para aproximar a distribuição *a posteriori*. Infelizmente, este tema não será tratado neste capítulo, mas é importante saber que a metodologia inferencial Bayesiana mais amplamente utilizada é baseada nos métodos de Monte Carlo via cadeias de Markov (MCMC: *Markov chain Monte Carlo*) (ver Gamerman e Lopes (2006) para mais detalhes).

Outro ponto crucial da perspectiva Bayesiana é a escolha da distribuição *a priori*. Neste quesito, a análise Bayesiana pode ser dividida em *subjetiva* e *objetiva*. Quando um estudo tem disponível informações de experimentos anteriores ou mesmo a opinião de especialistas, podemos incorporar este conhecimento na distribuição *a priori* (nem sempre é uma tarefa fácil/trivial), o que torna a análise subjetiva. Porém, em alguns cenários não temos ou não queremos incluir conhecimento prévio no estudo, portanto nossa distribuição *a priori* deve ser pouco informativa (também referenciada como *distribuição a priori vaga*), caracterizando uma análise objetiva.

Nesta breve comparação entre as perspectivas clássica e Bayesiana, já identificamos diferenças entre fontes de informação para realizar a inferência e também

quanto a interpretação dos parâmetros. Ainda que haja muitas outras divergências entre ambas as abordagens, em termos práticos, quando o tamanho da amostra (número de observações) é razoavelmente grande e não incluímos informações prévias na distribuição *a priori* (ou seja, inferência Bayesiana objetiva), as conclusões das análises clássica e Bayesiana costumam ser equivalentes. Por exemplo, a estimação pontual frequentista pode ser substituída por simples medidas descritivas da distribuição *a posteriori*, tais como média ou mediana. No entanto, temos que ser cautelosos na interpretação da estimação intervalar, pois desde o ponto de vista clássico, um intervalo de confiança de 100(1-α)%, onde α é o nível de significância, nos diz que ao repetir o experimento muitas vezes, 100(1-α)% dos intervalos de confiança calculados contêm o valor verdadeiro do parâmetro em questão. Por outro lado, a abordagem Bayesiana fornece um intervalo de credibilidade de 100(1-α)%, que representa a probabilidade de 100(1-α)% de que o parâmetro em questão está contido neste intervalo para os dados observados.

Apesar das particularidades teóricas e filosóficas desses dois paradigmas, ambos seguem em desenvolvimento e com diversas propostas metodológicas implementadas em programas estatísticos. Em particular, a Tabela 2 mostra alguns dos principais pacotes/funções, na linguagem R, para ajustar os modelos de regressão estudados neste capítulo.

INFERÊNCIA CLÁSSICA		
Pacote	**Função Principal**	**Descrição**
`stats`	`lm`	Ajusta modelos de regressão linear.
`stats`	`glm`	Ajusta modelos lineares generalizados (MLG).
`biglm`	`bigglm`	Ajusta MLG para *big data* (LUMLEY, 2013).
INFERÊNCIA BAYESIANA		
Pacote	**Função Principal**	**Descrição**
`MCMCglmm`	`MCMCglmm`	Ajusta modelos lineares generalizados mistos, em particular, MLG (HADFIELD, 2010).
`brms`	`brm`	Ajusta modelos de regressão usando Stan (método inferencial). Este pacote **requer** a instalação de um **compilador C++** (BÜRKNER, 2017).
`arm`	`bayesglm`	Ajusta MLG (GELMAN et al., 2016).

Tabela 2: Funções e pacotes para ajustar modelos de regressão em R.

6 Comparação de Modelos

Relembrando que modelos são simplificações da realidade, em muitos estudos, um mesmo fenômeno pode ser descrito satisfatoriamente com modelagens distintas, onde a diferença entre modelos pode ser caracterizada pela escolha da distribuição da variável resposta e/ou pelas variáveis explicativas selecionadas. Portanto, definir uma medida de comparação de modelos é fundamental tanto na seleção do melhor ajuste

quanto na objetividade das conclusões finais de um estudo.

Tanto a literatura clássica quanto a Bayesiana apresentam diversas propostas metodológicas de comparação de modelos, no entanto focaremos nas mais comumente aceitas até o momento. Especificamente na abordagem clássica, três medidas podem ser utilizadas:

- Coeficiente de determinação (R^2);
- Critério de informação de Akaike (AIC: *Akaike Information Criterion*);
- Critério de informação Bayesiano (BIC: *Bayesian Information Criterion*).

A forma padrão do coeficiente de determinação depende fortemente da relação linear entre a variável resposta e as variáveis explicativas, por isso ele somente é válido para comparar modelos de regressão linear, tais como MRLS ou MRLM. Este coeficiente mede a proporção da variação total que é explicada pelo modelo de regressão e seu intervalo de valores é definido entre 0 e 1, onde o modelo com R^2 mais próximo de 1 é tido como melhor opção. Em um MRLS, o coeficiente de determinação pode ser obtido através do quadrado do coeficiente de correlação de Pearson (ver Seção "Modelo de Regressão Linear Simples"). Na função `lm`, o R^2 para um MRLS é nomeado como `Adjusted R-squared`, enquanto que para um MRLM é mais adequado adotar sua versão estendida, `Multiple R-squared`.

A versão original do coeficiente de determinação (R^2 ordinário) é pouco utilizada por não levar em consideração a complexidade do modelo. Esta deficiência tem um grande peso em pesquisas científicas, onde geralmente se adota o *princípio da parcimônia*, ou seja, modelos mais simples são priorizados. No entanto, a versão mais popular do coeficiente de determinação (R^2 ajustado) incorpora uma penalização baseada no número de parâmetros de regressão.

Os critérios AIC e BIC são medidas mais gerais e que podem ser calculadas para qualquer modelo de regressão. Estes critérios são baseados na verossimilhança (avalia a bondade do ajuste) e número de parâmetros do modelo (avalia a complexidade). Em ambos os critérios, menor valor indica melhor modelo. Uma informação adicional é que o BIC é um critério para estimação clássica, porém, sua construção está baseada em argumentos Bayesianos e por isso ele leva o termo "*Bayesian*" em seu nome.

Quando a abordagem Bayesiana é empregada, dois critérios de comparação de modelos podem ser utilizados:

- Critério de informação de desviância/desvio (DIC: *Deviance Information Criterion*);
- Critério de informação de Watanabe-Akaike (WAIC: *Watanabe-Akaike Information Criterion* ou *Widely Applicable Information Criterion*).

Os critérios DIC e WAIC também ponderam a qualidade do ajuste com o número de parâmetros no modelo e suas interpretações são análogas aos dois critérios frequentista apresentados, ou seja, menor valor indica melhor modelo. No entanto, o DIC é tido como a generalização dos critérios AIC e BIC, enquanto que o WAIC incorpora melhorias ao DIC, principalmente quando as comparações envolvem modelos

bastante complexos.

Por não ser o objetivo específico deste capítulo, não abordaremos temas como análises de resíduos (inferência clássica) e diagnóstico de convergência (inferência Bayesiana), mas destacamos que ambos os temas são cruciais para nos certificarmos de que nosso modelo de regressão é apropriado e que a inferência estatística fornece resultados coerentes, confiáveis e reprodutíveis.

7 Cenário Ilustrativo

Com o intuito de praticar o conteúdo apresentado na Seção "Desenvolvimento", vamos analisar uma situação-problema introduzida por Paterson (1991) no contexto educacional, onde devemos propor e ajustar modelos de regressão e interpretar os resultados. Adicionalmente, visando obter a mesma problemática da Seção "Era uma vez", faremos algumas adaptações nos dados originais.

Nossa ferramenta de trabalho para analisar estes dados será a linguagem R (R CORE TEAM, 2016). Antes de introduzir o problema, vamos dar uma olhada nos dados (disponível no pacote `mlmRev`):

```
> require(mlmRev)
> data(ScotsSec)
> head(ScotsSec)
  verbal attain primary sex social second
1     11     10       1   M      0      9
2      0      3       1   F      0      9
3    -14      2       1   M      0      9
4     -6      3       1   M     20      9
5    -30      2       1   F      0      9
6    -17      2       1   F      0      9
```

Estes dados representam informações de n=3435 alunos escoceses do ensino médio. As variáveis do problema são descritas da seguinte forma:

- `verbal`: nota obtida em uma prova de raciocínio verbal ao entrar no ensino médio;
- `attain`: nota padronizada obtida em uma prova final do ensino médio (1-10);
- `primary`: código da escola onde o aluno cursou o ensino fundamental (1-148);
- `sex`: sexo (M: homem; F: mulher);
- `social`: classe social em uma escala de quatro níveis (baixa=0, 1, 20, 31=alta).
- `second`: código da escola onde o aluno cursou o ensino médio (1-19).

O objetivo do estudo é relacionar o desempenho dos alunos na prova final (`attain`) com as demais variáveis. Fazendo um paralelo com o problema introduzido na Seção "Era uma vez", `attain` é equivalente ao desempenho do estudante na resolução dos exercícios usando o APP desenvolvido por Jessica e, para que ambos os cenários tenham a mesma escala (0-9), podemos subtrair uma unidade de `attain`. Note que a analogia pode ser ainda mais direta se utilizarmos como variáveis explicativas

somente `verbal` e `sex`, as quais representariam no estudo de Jessica o conhecimento prévio em raciocínio verbal e o sexo, respectivamente. A mudança de escala e também a exclusão das variáveis `primary`, `social` e `second` podem ser feitas através do código abaixo:

```
> ScotsSec$attain <- ScotsSec$attain - 1
> dados <- ScotsSec[,-c(3,5,6)]
> summary(dados)
     verbal            attain      sex
 Min.   :-30.000   Min.   :0.000   M:1739
 1st Qu.:-11.000   1st Qu.:2.000   F:1696
 Median : -2.000   Median :4.000
 Mean   : -2.196   Mean   :4.679
 3rd Qu.:  7.000   3rd Qu.:8.000
 Max.   : 40.000   Max.   :9.000
```

Agora as informações para nossa análise estão armazenadas em `dados` e, como podemos ver acima, uma breve descrição dos dados mostra que `verbal` varia entre -30 e 40 com média em -2.196, `attain` tem uma média de 4.679 e temos 1739 homens em um total de 3435 alunos escoceses. Esta informação pode ser complementada graficamente através de um diagrama de dispersão entre `verbal` e `attain` (Figura 1), diferenciando cada ponto de acordo com `sex`:

```
> plot(dados$verbal, dados$attain, col=dados$sex, type="p",
           pch=19, xlab="verbal", ylab="attain")
> legend("bottomright", legend=levels(dados$sex),
           col=dados$sex, pch=19)
```

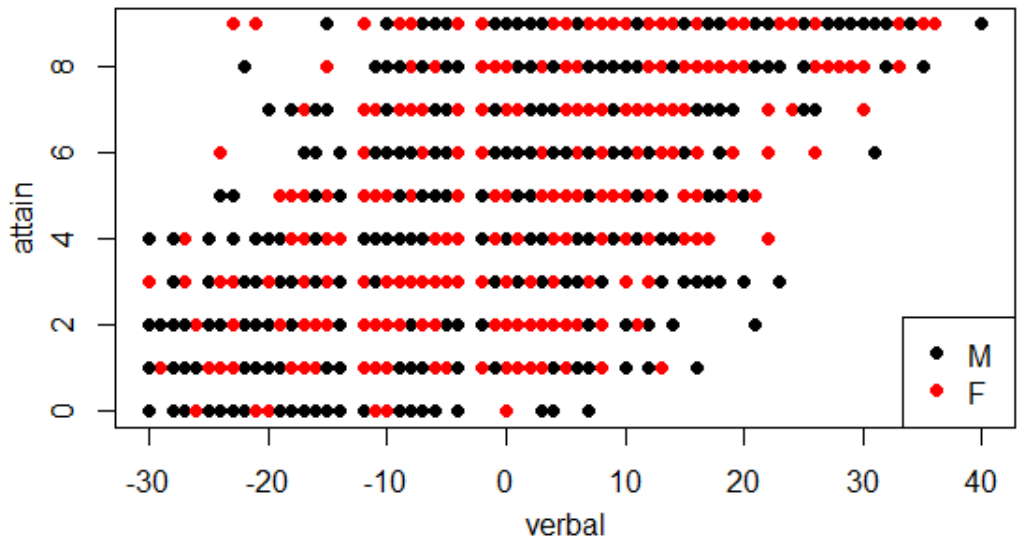

Figura 1: Diagrama de dispersão entre `verbal` e `attain`.

O gráfico acima mostra uma tendência linear crescente entre `verbal` e `attain` (ou seja, os valores de `attain` tendem a aumentar conforme `verbal` cresce) mas com muita variabilidade e com poucas evidências de que `sex` apresenta alguma influência em `attain`, pois não há um padrão explícito que diferencie homem (M) de mulher (F). Também podemos quantificar a relação linear entre cada variável explicativa e a variável resposta a partir do coeficiente de correlação de Spearman (ver Seção "Modelo de Regressão Linear Simples" para mais detalhes):

```
> cor(dados$verbal, dados$attain, method="spearman")
[1] 0.7335351
> cor(as.numeric(dados$sex), dados$attain, method="spearman")
[1] 0.08932479
```

Novamente, mas de forma mais precisa, podemos constatar que há uma relação linear positiva (0.7335351) entre `verbal` e `attain`, enquanto `sex` tem uma fraca correlação com `attain` (0.08932479). O critério para tal interpretação de correlação entre duas variáveis pode ser revisto na Seção "Modelo de Regressão Linear Simples".

Observe que a variável resposta (`attain`) assume somente valores discretos de 0 a 9, portanto, teoricamente, deveríamos assumir um modelo de regressão que se ajuste a esta condição. Na prática, podemos ignorar temporariamente esta restrição e assumir, de forma aproximada, um modelo de regressão linear simples/múltipla.

Para fins didáticos, faremos a inferência desde o ponto de vista clássico e Bayesiano (análise objetiva) utilizando as funções/pacotes na linguagem R apresentados na Seção "Métodos de Estimação". O primeiro passo é ajustar um MRLM incluindo as variáveis independentes `verbal` (X_{1i}) e `sex` (X_{2i}), com $i = 1, \ldots, n = 3435$, e assim teremos uma ideia da relevância de cada uma delas para explicar `attain` (Y_i):

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i, \qquad \text{(M1)}$$

onde o erro ϵ_i segue uma distribuição Normal com média zero e variância σ^2. Desde uma perspectiva frequentista, podemos ajustar o modelo M1 a partir da função `lm`:

```
> M1.freq <- lm(attain ~ verbal + sex, data=dados)
> summary(M1.freq)

Call:
lm(formula = attain ~ verbal + sex, data = dados)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6310 -1.4518 -0.0325  1.5237  7.7029

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.991622   0.051884  96.206   <2e-16 ***
verbal      0.164964   0.002743  60.129   <2e-16 ***
sexF        0.099606   0.072933   1.366    0.172
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.127 on 3432 degrees of freedom
Multiple R-squared:  0.5165,    Adjusted R-squared:  0.5163
F-statistic:  1833 on 2 and 3432 DF,  p-value: < 2.2e-16
```

Embora haja várias informações na saída da função `lm`, vamos focar somente na relevância das variáveis independentes do modelo M1. Claramente, `verbal` é significativo (p-valor `<2e-16`) para explicar `attain`, enquanto a contribuição da variável `sex` não aporta muita informação (p-valor `0.172`). De forma análoga, podemos ajustar o modelo M1 a partir de uma abordagem Bayesiana utilizando o pacote `MCMCglmm`:

```
> require(MCMCglmm)
> M1.bayes <- MCMCglmm(attain ~ verbal + sex, data=dados)
> summary(M1.bayes)

 Iterations = 3001:12991
 Thinning interval  = 10
 Sample size  = 1000

 DIC: 14938.7

 R-structure:  ~units

      post.mean l-95% CI u-95% CI eff.samp
units     4.528    4.327    4.746     1000

 Location effects: attain ~ verbal + sex

            post.mean l-95% CI u-95% CI eff.samp  pMCMC
(Intercept)   4.99094  4.89655  5.09589     1110 <0.001 ***
verbal        0.16481  0.15969  0.16990     1000 <0.001 ***
sexF          0.09865 -0.03536  0.24968     1000  0.188
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Aqui também temos muitas informações, mas é importante destacar que os valores da média *a posteriori* de cada parâmetro são praticamente os mesmos obtidos na abordagem frequentista. Lembrando que esta similaridade nas conclusões se deve ao fato do número de observações ser grande (n=3435) e por não estarmos incorporando informações prévias na distribuição *a priori*. Além disso, Hadfield (2010) inclui uma proposta de "p-valor Bayesiano" (`pMCMC`) nos resultados inferenciais acima. Embora esta proposta não seja amplamente aceita entre "Bayesianos", ela também aponta a variável `sex` como pouco relevante na explicação de `attain`. Portanto, temos resultados preliminares equivalentes empregando ambos os métodos de estimação.

Uma vez que a variável `sex` não contribui de forma significativa no modelo M1, podemos eliminá-la, nos levando a dar um passo atrás e utilizar um MRLS dado por:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i. \qquad \text{(M2)}$$

Novamente, primeiro ajustaremos o modelo M2 baseado em estimação por máxima verossimilhança usando a função `lm`:

```
> M2.freq <- lm(attain ~ verbal, data=dados)
> summary(M2.freq)

Call:
lm(formula = attain ~ verbal, data = dados)

Residuals:
    Min      1Q  Median      3Q     Max
-6.6868 -1.4045 -0.0497  1.4786  7.7608

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.041592   0.036792  137.03   <2e-16 ***
verbal      0.165323   0.002731   60.53   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.127 on 3433 degrees of freedom
Multiple R-squared:  0.5163,    Adjusted R-squared:  0.5161
F-statistic:  3664 on 1 and 3433 DF,  p-value: < 2.2e-16
```

A partir dos resultados apresentados acima, podemos verificar imediatamente que as estimativas de β_0 (intercepto) e β_1 para os modelos M1 e M2 (baseado na inferência clássica) quase não foram alteradas e que o `Multiple R-squared` (`0.5165`) do modelo M1 e `Adjusted R-squared` (`0.5161`) do modelo M2 são basicamente os mesmos. Esta rápida comparação já seria suficiente para escolhermos o modelo M2, uma vez que ele é mais parcimonioso, ou seja, o modelo M2 é mais simples e explica de forma equivalente ao modelo mais complexo (M1). Também podemos comparar ambos os modelos usando o critério AIC:

```
> AIC(M1.freq)
[1] 14938.68
> AIC(M2.freq)
[1] 14938.55
```

Embora saibamos que o melhor modelo tem o menor AIC, neste caso a diferença é ínfima e, portanto, podemos concluir que os modelos M1 e M2 são equivalentes. Portanto, escolhemos M2 pelo princípio da parcimônia.

Seguindo a estratégia didática de expor os métodos de estimação clássico e Bayesiano, ajustaremos o modelo M2 baseado na inferência Bayesiana:

```
> M2.bayes <- MCMCglmm(attain ~ verbal, data=dados)
> summary(M2.bayes)

 Iterations = 3001:12991
```

```
 Thinning interval  = 10
 Sample size  = 1000

 DIC: 14938.58

 R-structure:  ~units

      post.mean l-95% CI u-95% CI eff.samp
units     4.523    4.322    4.745    788.8

 Location effects: attain ~ verbal

            post.mean l-95% CI u-95% CI eff.samp  pMCMC
(Intercept)    5.0418   4.9704   5.1111     1000 <0.001 ***
verbal         0.1654   0.1604   0.1707     1000 <0.001 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note que a comparação entre os modelos Bayesianos M1 e M2 é análoga ao caso frequentista. Na abordagem Bayesiana, o DIC (*Deviance Information Criterion*) é um dos principais critérios de comparação de modelos, por meio dele podemos concluir que o modelo M2 é novamente o mais adequado, uma vez que seu DIC é de 14938.58 e o do modelo M1 é 14938.7.

Aparentemente, temos que o modelo M2 é o mais apropriado em nossa análise. No entanto, ainda devemos checar se o pressuposto de normalidade dos resíduos é cumprido. Em um primeiro momento, podemos verificar de forma gráfica (Figura 2) se os resíduos seguem uma distribuição Normal:

```
> res <- rstandard(M2.freq)
> par(mfrow=c(1,2))
> boxplot(res)
> qqnorm(res, ylab="Resíduos", xlab="", main="")
> qqline(res, col="red")
```

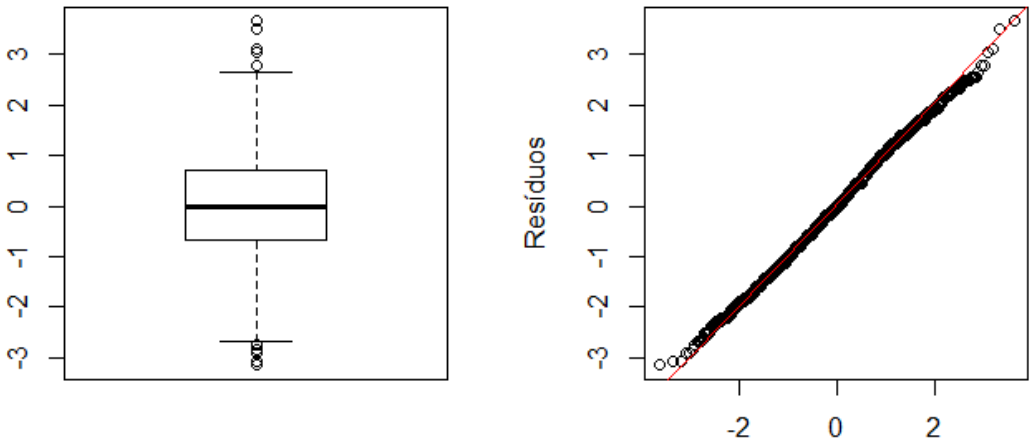

Figura 2: Diagrama de caixa e gráfico Q-Q Normal.

À esquerda, o gráfico/diagrama de caixa (*boxplot*) nos fornece informações dos resíduos quanto à mediana (linha central), simetria (distância entre a mediana e o primeiro/terceiro quartil, caracterizado pelo limite inferior/superior da caixa) e valores atípicos/discrepantes (*outliers*), representados pelos pontos. À direita e de forma mais direta, temos o gráfico Q-Q (ou Quantil-Quantil) Normal que nos ajuda a verificar se os resíduos apresentam distribuição Normal, onde a linearidade (linha vermelha como referência) dos pontos sugere que os resíduos são normalmente distribuídos. Em ambos os casos, os valores atípicos nos extremos deixam dúvidas sobre a normalidade dos resíduos. Para verificar a normalidade de forma mais precisa, podemos utilizar algum teste estatístico, tais como Shapiro-Wilk, Kolmogorov-Smirnov ou Anderson-Darling. Estes testes são similares e sua principal finalidade é checar se a distribuição dos resíduos pode ser aproximada pela distribuição Normal. O código a seguir é referente ao teste de Shapiro-Wilk, onde a hipótese nula é dada pela normalidade dos resíduos:

```
> shapiro.test(res)

    Shapiro-Wilk normality test

data:  res
W = 0.99816, p-value = 0.0005011
```

A partir do teste acima, podemos concluir que, para qualquer nível de significância maior que 0.0005011, temos evidências para rejeitar a hipótese nula.

Para contornar o problema de violação do pressuposto de normalidade e também cumprir a restrição, previamente ignorada, de que a variável resposta (`attain`) assume somente valores discretos entre 0 e 9, propomos um modelo Binomial (ou seja, um MLG) com função de ligação logito (ou logística):

$$Y_i \sim Binomial(m, p_i), \tag{M3.1}$$

$$\text{logito}(p_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}, \tag{M3.2}$$

onde m representa a nota máxima na prova final (ou seja, m=9) e mp_i é o valor esperado (nota média) de Y_i. Embora tenhamos optado pela função de ligação logito, outras opções seriam probito (*probit*) e complemento log-log (*cloglog*).

Como na primeira análise, iremos apresentar os resultados do modelo de regressão Binomial M3 usando a estimação por máxima verossimilhança e também Bayesiana (análise objetiva). Primeiro vamos empregar a inferência clássica ao modelo M3 utilizando a função `glm`:

```
> dados$m <- 9
> M3.freq <- glm(attain/m ~ verbal + sex,
            family=binomial(link="logit"), data=dados, weights=m)
> summary(M3.freq)

Call:
glm(formula = attain/m ~ verbal + sex, family = binomial(link =
"logit"), data = dados, weights = m)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.0572  -1.0099   0.0555   1.3417   5.9971

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.295105   0.018906  15.609   <2e-16 ***
verbal      0.095451   0.001254  76.101   <2e-16 ***
sexF        0.047884   0.026273   1.823   0.0684 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 17697.9  on 3434  degrees of freedom
Residual deviance:  9322.8  on 3432  degrees of freedom
AIC: 15538

Number of Fisher Scoring iterations: 4
```

Observe que nesta análise, considerando um nível de significância de 0,05, também temos que a variável `verbal` (X_1) aporta informação considerável para explicar o rendimento do aluno na prova final do ensino médio, enquanto `sex` (X_2) não é vigorosamente relevante. Análogo a modelagem apresentada anteriormente, também podemos propor um novo modelo excluindo a variável independente `sex` e comparar a qualidade do novo ajuste com os resultados do modelo M3.

$$Y_i \sim Binomial(m, p_i), \tag{M4.1}$$

$$\text{logito}(p_i) = \beta_0 + \beta_1 X_{1i}. \tag{M4.2}$$

A estimação clássica do modelo M4 é dada abaixo:

```
> M4.freq <- glm(attain/m ~ verbal,
           family=binomial(link="logit"), data=dados, weights=m)
> summary(M4.freq)

Call:
glm(formula = attain/m ~ verbal, family = binomial(link =
"logit"), data = dados, weights = m)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.0931  -1.0274   0.0558   1.3417   6.0332

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.319229   0.013520   23.61   <2e-16 ***
verbal      0.095628   0.001251   76.45   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 17697.9  on 3434  degrees of freedom
Residual deviance:  9326.1  on 3433  degrees of freedom
AIC: 15540

Number of Fisher Scoring iterations: 4
```

Similar a comparação entre os modelos M1 e M2, também selecionaremos o melhor modelo a partir do AIC, onde obtivemos 15538 para M3 e 15540 para M4. Baseado no princípio da parcimônia e considerando que a diferença entre ambos os modelos é de apenas 2 unidades, assumimos que o modelo M4 é o mais apropriado.

Usamos a função `brm`, disponível no pacote `brms`, com o intuito de ilustrar o ajuste do modelo M4 utilizando a inferência Bayesiana:

```
> require(brms)
> M4.bayes <- brm(attain | trials(m) ~ verbal,
           family=binomial(link="logit"), data=dados)
> summary(M4.bayes)
 Family: binomial
  Links: mu = logit
Formula: attain | trials(m) ~ verbal
   Data: dados (Number of observations: 3435)
Samples: 4 chains, each with iter = 2000; warmup = 1000; thin =
1;
         total post-warmup samples = 4000
    ICs: LOO = NA; WAIC = NA; R2 = NA

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Eff.Sample Rhat
Intercept     0.32      0.01     0.29     0.35       1261 1.00
verbal        0.10      0.00     0.09     0.10       4000 1.00

Samples were drawn using sampling(NUTS). For each parameter,
Eff.Sample is a crude measure of effective sample size, and
Rhat is the potential scale reduction factor on split chains
(at convergence, Rhat = 1).
```

Embora tenhamos omitido a inferência Bayesiana para o modelo M3, o comparamos com o modelo M4 utilizando outro critério Bayesiano para comparação de modelos, o WAIC, disponível no pacote `brms` na forma `WAIC(M3.bayes)` e `WAIC(M4.bayes)`. Como resultado comparativo, obtivemos 15543 (M3) e 15542 (M4), que reforça as evidências a favor do modelo M4, uma vez que o princípio da parcimônia prevalece perante a estas diferenças de WAICs.

Finalmente, podemos concluir que, dentre os modelos analisados, a melhor opção é M4 e, baseado na estimativa pontual frequentista ou na média *a posteriori*, seu valor predito é dado por:

$$\text{logito}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = 0{,}32 + 0{,}10\text{X}_{1i}.$$

De forma bastante simples, podemos obter a probabilidade de acerto (p_i), em cada questão, do aluno i no estudo de Paterson (1991) considerando como variável explicativa somente a nota deste aluno em uma prova de raciocínio verbal ao entrar no ensino médio (`verbal`=X_{1i}). No entanto, também estamos interessados em interpretar a influência de X_{1i} no desempenho deste aluno na prova final do ensino médio. Esta interpretação é feita através da razão de chances, que é definida por:

$$\psi = \frac{p(X_{1i}+1)}{[1-p(X_{1i}+1)]}\frac{[1-p(X_{1i})]}{p(X_{1i})} = \exp(\beta_1) = \exp(0{,}10) \approx 1{,}1052,$$

onde $p(X_{1i}+1) = \exp\big(\beta_0 + \beta_1(X_{1i}+1)\big)/\big[1+\exp\big(\beta_0+\beta_1(X_{1i}+1)\big)\big]$. Portanto, em nosso contexto, a razão de chances ψ igual a 1,1052 significa que a cada 1 ponto a mais na nota de raciocínio verbal ao entrar no ensino médio, o aluno aumenta as chances de acerto na prova final em 10,52%, em média. De forma geral, o cálculo da razão de chances associada com K pontos a mais na nota de raciocínio verbal é dado por $\exp(0{,}10K)$.

8 Resumo

Neste capítulo estudamos os modelos de regressão linear simples e múltipla, além de introduzir os pontos principais dos modelos lineares generalizados. Também vimos os conceitos dos paradigmas clássico e Bayesiano para métodos de estimação, funções/pacotes na linguagem R para empregar ambas as perspectivas e os principais critérios de comparação de modelos tanto para a inferência clássica quanto para a Bayesiana. Para fixar todo o conteúdo deste capítulo, a Seção "Cenário Ilustrativo" apresentou de forma prática e detalhada todo o processo de modelagem estatística, baseado nos modelos de regressão apresentados, aplicado a um exemplo real análogo ao contexto da Seção "Era uma vez".

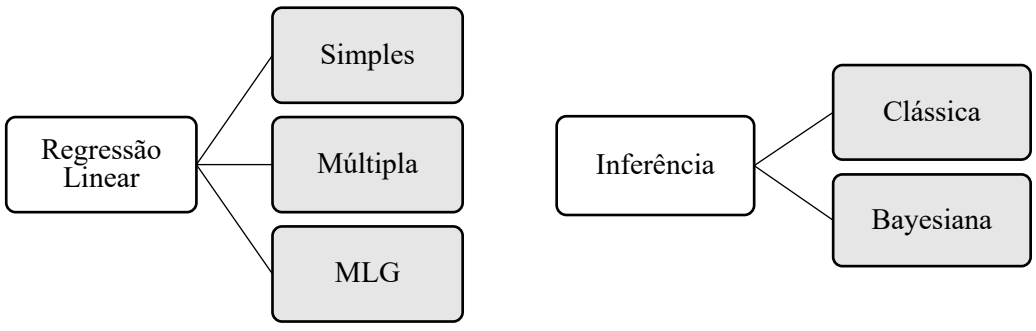

Figura 3: Mapa mental das teorias e diretrizes para modelos de regressão.

9 Leituras Recomendadas

- **Multilevel modelling of educational data** (O'CONNELL; MCCOACH, 2008). Este livro apresenta e ilustra vários modelos de regressão em aplicações no contexto de educação, facilitando o entendimento de quem trabalha nessa área.

- **Modelos de regressão com apoio computacional** (PAULA, 2013). Este livro apresenta os modelos de regressão (em especial, os modelos lineares generalizados) com mais detalhes sobre aspectos estatísticos (desde o ponto de vista da estatística clássica) e códigos computacionais em R.

- **Modelos lineares generalizados e extensões** (CORDEIRO; DEMÉTRIO, 2008). Este livro cobre todo o conteúdo teórico de cursos de Modelos Lineares Generalizados ministrados tanto na graduação quanto na pós-graduação em Estatística no Brasil. Ele é uma boa referência para leitores familiarizados com conceitos estatísticos/matemáticos.

- **Data analysis using regression and multilevel/hierarchical models** (GELMAN; HILL, 2006). Este é um dos livros mais modernos que trata modelos de regressão (incluindo tópicos além do escopo deste capítulo) desde o ponto de vista clássico e Bayesiano. Os autores utilizam uma linguagem estatística/matemática acessível e consistente, incluindo um grande número de exemplos, gráficos e códigos que tornam o conteúdo mais compreensível.

- **Generalized linear models** (MCCULLAGH; NELDER, 1989). Este é um livro de referência mundial em cursos de Modelos Lineares Generalizados. Embora seja um livro com perfil mais teórico, ele inclui exemplos e exercícios que dão suporte para uma maior compreensão do conteúdo apresentado.

10 Artigos Exemplos

- **Uma abordagem de regressão múltipla para validação de variáveis de autorregulação da aprendizagem em ambientes de LMS** (RODRIGUES et al., 2016). Este artigo ilustra perfeitamente a aplicação de um modelo de regressão linear múltipla no contexto de EAD, onde o objetivo é validar variáveis comportamentais de autorregulação da aprendizagem.
- **A study on multiple linear regression analysis** (UYANIK; GÜLERB, 2013). Este artigo apresenta uma rápida revisão de modelos de regressão linear múltipla com estimação clássica (máxima verossimilhança) e exemplifica seu uso em um contexto educacional.

- **Exploring Bayesian models to evaluate control procedures for plant disease** (ALVARES et al., 2016). Este artigo aplica vários modelos de regressão desde uma perspectiva Bayesiana e interpreta os resultados de uma forma gráfica e simplificada.

- **Métodos de estimação em regressão linear múltipla: aplicação a dados clínicos** (COELHO-BARROS et al., 2008). Este artigo apresenta de forma resumida os principais métodos de estimação para modelos de regressão linear múltipla.

- **Generalized linear mixed models: a practical guide for ecology and evolution** (BOLKER et al., 2009). Embora este artigo aborde uma modelagem mais geral que os MLG, ele realmente serve de guia para um leitor iniciante em modelos de regressão, principalmente por apresentar uma estrutura bastante didática e resumida para os métodos de estimação (clássico e Bayesiano), MLG, pacotes computacionais e uma excelente árvore de decisão para a escolha da modelagem e teste de hipótese.

11 Checklist

i. Identificar o tipo de variável resposta (contínua, discreta, positiva, etc.).

ii. Definir o modelo mais apropriado (baseado em **i.**).

iii. Escolher abordagem de estimação (por exemplo, clássica ou Bayesiana).

iv. Ajustar o modelo proposto usando uma ferramenta estatística (por exemplo, R).

v. Checar os pressupostos do modelo utilizado (se houver).

vi. Interpretar os resultados.

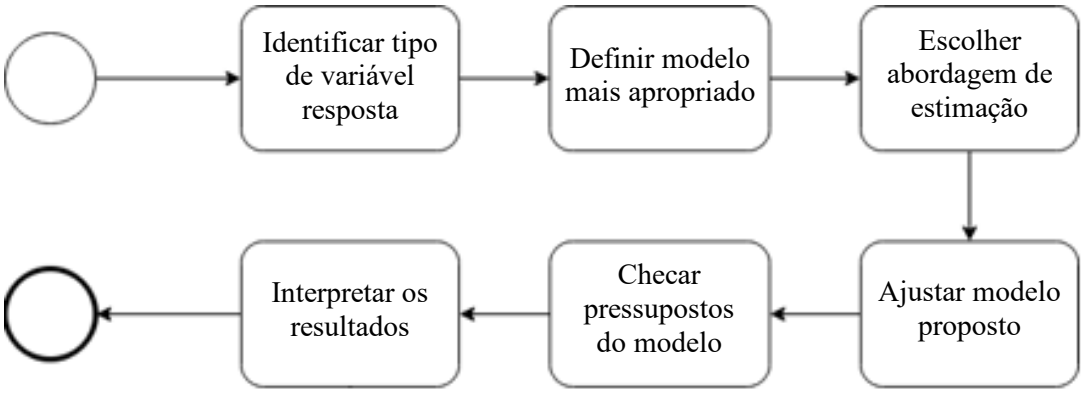

Figura 4: Fluxograma de atividades para modelos de regressão.

12 Exercícios

1) Vimos na Seção "Modelo de Regressão Linear Simples" que o MRLS é dado por:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \text{ para } i = 1, \dots, n,$$

onde, tipicamente, $\epsilon_i \sim Normal(0, \sigma^2)$. Explique as diferenças entre o paradigma clássico e Bayesiano quanto a definição dos parâmetros β_0, β_1 e σ^2 do modelo acima.

2) Uma professora de Filosofia está interessada em analisar se o tempo (em minutos) em que um estudante leva para fazer uma prova está relacionado com sua nota (0-100). Para isso, a professora dispõe dos tempos e notas de uma turma de 30 alunos:

Aluno	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Tempo	64,1	64,6	55,6	64,9	81,3	70,5	65,2	58,0	62,7	68,6	73,2	55,4	48,8	56,4	50,6
Nota	51,6	73,8	47,4	60,7	77,1	54,0	55,3	53,6	57,1	61,4	72,3	52,0	49,2	58,4	49,3
Aluno	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
Tempo	61,2	67,5	37,7	72,4	62,1	63,5	55,0	68,3	67,3	73,2	54,7	68,2	71,7	52,9	53,3
Nota	58,6	54,0	38,2	62,4	57,4	69,0	69,4	74,7	67,1	68,2	57,1	74,2	75,0	53,4	51,5

i. Faça um diagrama de dispersão do tempo em função da nota e conclua se há uma relação linear entre elas.

ii. Como seria o modelo de regressão para esta análise? Estime seus parâmetros através da abordagem frequentista e Bayesiana (análise objetiva) usando algum programa estatístico?

iii. Inclua no diagrama de dispersão (item **i.**) as retas de regressão obtidas com a estimação clássica e Bayesiana. Em ambas as abordagens, qual seria o tempo médio de um estudante que tirou zero na prova?

3) A professora do exercício anterior desconfia que o sexo do estudante (M: mulher; H: homem) também pode ser relevante na análise. Com isso, ela deseja modelar o mesmo problema do exercício **B)** incluindo o sexo dos 30 alunos:

Aluno	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
Sexo	M	H	M	M	M	M	M	H	M	M	M	H	H	H	H
Aluno	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
Sexo	M	M	H	M	M	H	H	H	M	M	H	H	H	H	H

i. Faça um diagrama de dispersão do tempo em função da nota e diferencie cada ponto de acordo com o sexo do estudante.

ii. Ajuste, via estimação por máxima verossimilhança e Bayesiana (análise objetiva), um modelo de regressão linear múltipla (MRLM) utilizando os dados do exercício **B)** e a variável sexo.

iii. Compare, empregando um dos critérios apresentados na Seção "Métodos de Estimação", a abordagem frequentista do modelo proposto no exercício **B)** com o MRLM. Qual destes modelos é mais apropriado?

iv. Inclua no diagrama de dispersão (item **i.**) as retas de regressão obtidas com a estimação Bayesiana tanto do exercício **B)** quanto deste exercício. Como conclusão, você diria que o sexo é relevante para discriminar a relação do tempo em função da nota? Por quê?

4) Um professor de Matemática está criando uma aplicação online para que seus alunos possam praticar o raciocínio lógico com perguntas de múltipla escolha.

Embora ele tenha uma infinidade de exercícios, seu objetivo é simular uma situação de prova com um número reduzido de questões. Portanto, o professor precisa definir este número considerando o tempo máximo de prova. Em uma de suas aulas, o professor resolveu fazer um teste piloto de sua aplicação com os 40 alunos presentes. Basicamente, a aplicação registrou o número de perguntas respondidas por cada aluno (independente se está correta ou não) em um intervalo de uma hora. Como informação adicional, o professor também dispõe do sexo de cada aluno (M: mulher; H: homem). A tabela abaixo mostra toda a informação coletada pelo professor:

Aluno	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**
Nº Resp	10	20	22	21	23	26	25	17	10	28	22	28	20	16	19	25	20	22	22	21
Sexo	M	M	H	H	M	M	M	M	H	H	H	M	H	H	M	H	H	H	H	M
Aluno	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**	**39**	**40**
Nº Resp	15	20	19	20	27	22	28	18	18	18	19	15	27	26	14	21	27	20	24	18
Sexo	M	M	M	H	H	M	M	H	M	H	H	M	M	H	H	M	H	M	M	H

i. Qual modelo de regressão é mais apropriado para este problema? Verifique as suposições da modelagem escolhida.

ii. A partir da perspectiva Bayesiana objetiva, ajuste o modelo proposto no item anterior incluindo a variável sexo e conclua se há diferenças entre homens e mulheres quanto ao número médio de questões respondidas em uma hora.

iii. Ajuste o modelo do item anterior via máxima verossimilhança e diga quantas questões o professor deve incluir em uma prova real usando sua aplicação online?

13 Referências

ALVARES, D.; ARMERO, C.; FORTE, A.; RUBIO, L. Exploring Bayesian models to evaluate control procedures for plant disease. **Statistics and Operations Research Transactions**, 2016. v. 40, n. 1, p. 139-152.

BOLKER, B.M.; BROOKS, M.E.; CLARK, C.J.; GEANGE, S.W.; POULSEN, J.R.; STEVENS, M.H.H.; WHITE, J.S.S. Generalized linear mixed models: a practical guide for ecology and evolution. **Trends in Ecology & Evolution**, 2009. v. 24, n. 3, p. 127-135.

BÜRKNER, P.C. brms: An R package for Bayesian multilevel models using Stan. **Journal of Statistical Software**, 2017. v. 80, n. 1, p. 1-28.

COELHO-BARROS, E.A; SIMÕES, P.A.; ACHCAR, J.A.; MARTINEZ, E.Z.; SHIMANO, A.C. Métodos de estimação em regressão linear múltipla: aplicação a dados clínicos. **Revista Colombiana de Estadística**, 2008. v. 31, n. 1, p. 111-129.

CORDEIRO, G.M.; DEMÉTRIO, C.G.B. **Modelos lineares generalizados e extensões** (ano 2008). Disponível em: http://www.ufjf.br/clecio_ferreira/files/2013/05/Livro-Gauss-e-Clarice.pdf. Acesso em: 23 de nov. 2017.

GAMERMAN, D.; LOPES, H.F. **Markov chain Monte Carlo: Stochastic simulation for Bayesian inference**. 2ª edição. Chapman and Hall/CRC, 2006.

GELMAN, A.; HILL, J. **Data analysis using regression and multilevel/hierarchical models**. 1ª edição. Cambridge University Press, 2006.

GELMAN, A.; SU, Y.S.; YAJIMA, M.; HILL, J.; PITTAU, M.G.; KERMAN, J.; ZHENG, T.; DORIE, V. **arm: Data analysis using regression and multilevel/hierarchical models** (ano 2016, versão 1.9-3). Disponível em: https://cran.r-project.org/web/packages/arm.

HADFIELD, J.D. MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. **Journal of Statistical Software**, 2010. v. 33, n. 2, p. 1-22.

LUMLEY, T. **Bounded memory linear and generalized linear models** (ano 2013, versão 0.9-1). Disponível em: https://cran.r-project.org/web/packages/biglm.

MCCULLAGH, P.; NELDER, J.A. **Generalized linear models**. 2ª edição. Chapman and Hall/CRC, 1989.

NELDER, J.A.; WEDDERBURN, R.W.M. Generalized linear models. **Journal of the Royal Statistical Society: Series A (General)**, 1972. v. 135, n. 3, p. 370-384.

O'CONNELL, A.A.; MCCOACH, D.B. (eds) **Multilevel modelling of educational data**. 1ª edição. Information Age Publishing, 2008.

PAULA, G.A. **Modelos de regressão com apoio computacional** (ano 2013). Disponível em: https://www.ime.usp.br/~giapaula/texto_2013.pdf. Acesso em: 23 de nov. 2017.

PATERSON, L. Socio-economic status and educational attainment: a multi-dimensional and multi-level study. **Evaluation & Research in Education**, 1991. v. 5, n. 3, p. 97-121.

R CORE TEAM. R: A language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, Austria, 2016. Disponível em: http://www.R-project.org.

RODRIGUES, R.; SILVA, J.; RAMOS, J.L.C.; SOUZA, F.F.; GOMES, A.S. Uma abordagem de regressão múltipla para validação de variáveis de autorregulação da aprendizagem em ambientes de LMS. **Anais do 27º Simpósio Brasileiro de Informática na Educação**, Uberlândia, 2016. p. 916-925.

UYANIK, G.K.; GÜLERB, N. A study on multiple linear regression analysis. **Procedia - Social and Behavioral Sciences**, 2013. v. 106, p. 234-240.

Sobre o Autor

Danilo Alvares da Silva

http://lattes.cnpq.br/6725871336446527

Doutor em Estatística e mestre em Bioestatística pela Universitat de València, Espanha (2017 e 2015), mestre em Ciência da Computação e Matemática Computacional e graduado em Matemática Aplicada e Computação Científica pela Universidade de São Paulo, Brasil (2013 e 2011). Atualmente, Danilo é pesquisador de pós-doutorado na Harvard T.H. Chan School of Public Health, EUA, e sua pesquisa está focada em abordagens Bayesianas para modelagem estatística e métodos computacionais.