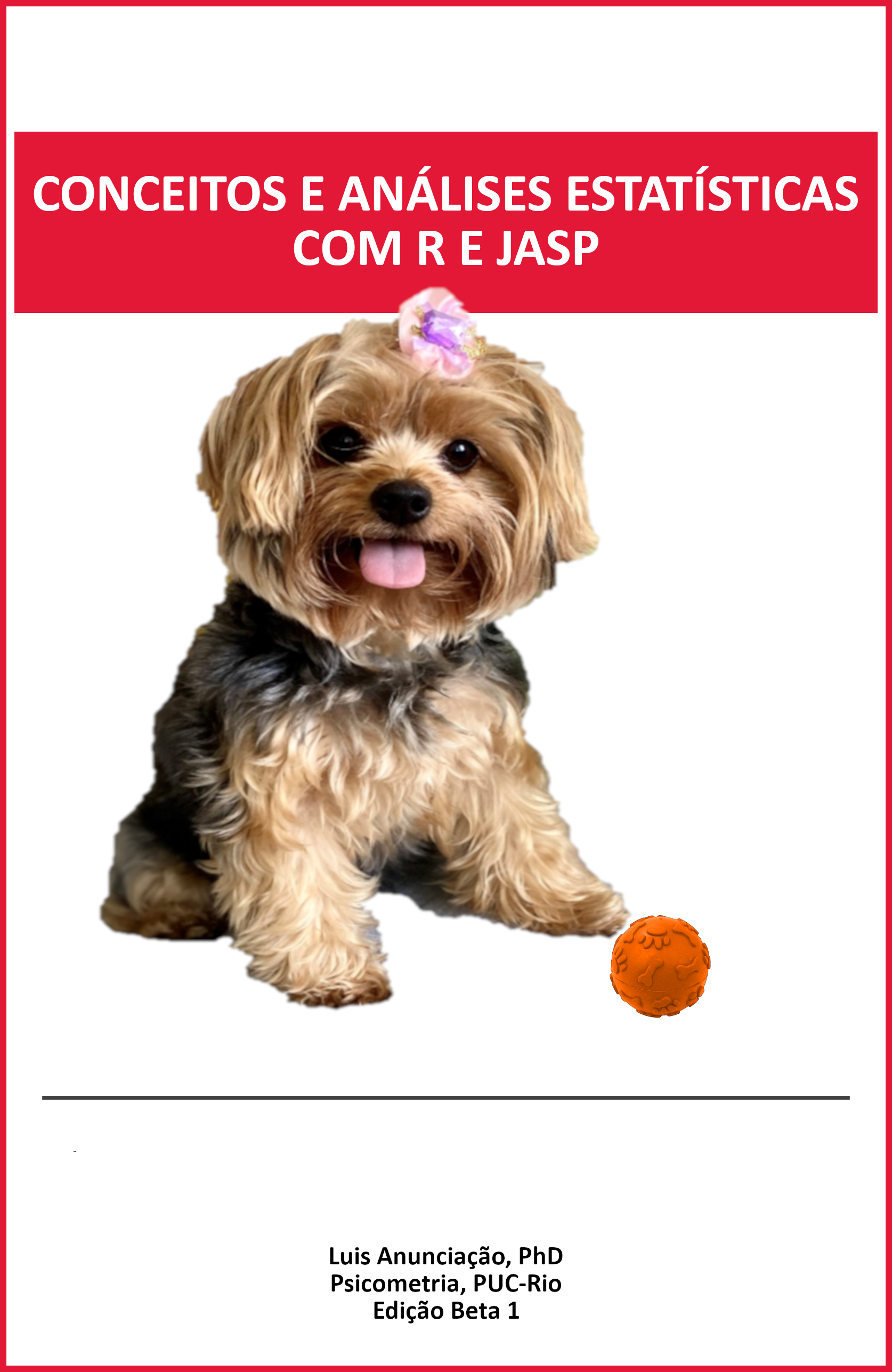
Conceitos e análises estatísticas com R e JASP

[Luis Anunciação (PUC-Rio), PhD](mailto:%20luisfca@puc-rio.br)

# 1 Prefácio



## 1.1 Atenção

Atenção: Leia com cuidado. Este livro ainda está em sua fase de revisão.  
Última modificação: 05 March, 2021 às 21:31

## 1.2 A proposta

Este livro nasceu como um dos principais e mais frutíferos resultados das aulas de graduação e pós-graduação ministradas por mim em alguns locais, mas com maior intensidade na PUC-Rio, UFRJ e IBNeuro. Por bastante tempo, nas aulas de estatística aplicada à Psicologia e Bioestatística, eu recorri a diferentes livros que, cada qual a sua maneira, apresentavam conceitos de pesquisa, técnicas estatísticas e análise de dados.

No entanto, acabei percebendo (ou tendo a impressão) de que eles apresentam a estatística por diferentes atalhos pedagógicos, (1) sugerindo que pesquisa e estatística eram áreas distantes, (2) que toda estatística podia ser resumida por testes de hipóteses independentes entre si e que (3) situações envolvendo dados reais não tinham tanto interesse. No geral, parece-me que para eles apresentarem a estatística na ciência, era necessário se distorcer pesadamente a ciência da estatística.

Em função disso, nos últimos anos, eu fui sentindo necessidade de apresentar os conceitos de pesquisa e técnicas estatísticas de forma integrada, contanto com dados reais e seguindo por uma metodologia de aula que pudesse ser pragmática, mas sem reforçar vícios inadequados sobre conceitos de estatística.

Conciliar essas condições em um único livro de maneira adequada é bastante improvável. Dessa forma, esse livro opta por uma abordagem majoritariamente pragmática, mas que evita se distanciar de conceitos teóricos. O pragmatismo é fundamental para que o estudante consiga, rapidamente, entender os procedimentos relacionados à análise de dados e implementar técnicas estatísticas para tomar decisões. Quão antes o estudante entender a utilidade da estatística para resolver problemas, maior é a probabilidade dele vir a gostar da área. Por sua vez, os aspectos teóricos são os alicerces para que o estudante possa perceber também que a utilidade que a estatística tem na ciência só é possível por ela ser uma disciplina sólida e robusta e que veio se aprimorando nas últimas décadas.

Isso posto, este livro é fruto de um grande esforço que tem a proposta de ser um manual técnico, em que são apresentados conceitos de pesquisa e análises estatísticas realizadas no R e no JASP e com especial aplicação em Psicologia e Bioestatística. Em cada capítulo, o estudante terá a oportunidade de acessar:

1. Uma pesquisa científica, explicitando o problema e as hipóteses que a guiaram
2. O artigo publicado com os resultados
3. A base de dados em formato R ou CSV para reprodução das análises
4. O conjunto de procedimentos estatísticos utilizados
5. Recursos extras para aprofundamento em tópicos específicos
6. Exercícios que auxiliem no entendimento do conteúdo, quase sempre retirados de provas externas

Com isso, o livro oferece ao estudante um ambiente em que ele possa resolver um problema real, utilizando as técnicas e métodos estatísticos como ferramentas para tomada de decisão. Apesar do foco ser mais no problema de pesquisa do que nas ferramentas analíticas, em todos os capítulos, a aplicação da estatística na ciência será reforçada pela apresentação de alguns conceitos da ciência da estatística.

Espero que este livro possa ser útil a estudantes de graduação e pós-graduação, agradável a leitores de estatística como de Psicologia e um recurso importante para outros docentes que, eventualmente, precisem de um material de apoio.

## 1.3 Objetivo

O livro tem como objetivos (1) apresentar, (2) discutir e (3) operacionalizar conceitos de pesquisa e análises estatística de dados a partir de pesquisas publicadas e dados reais. Espera-se que qualquer o estudante consiga realizar todas as ações descritas no decorrer dos capítulos de maneira guiada e intuitiva. As sintaxes utilizadas no ambiente R e as telas de execução do JASP são integralmente disponíveis.

## 1.4 Público-alvo

Este livro foi desenvolvido de maneira mais focada a estudantes de Psicologia e Bioestatística. As pesquisas e exemplos utilizados são mais aderentes a essas duas áreas. No entanto, como parte dos conceitos e análises implementadas no livro são interdisciplinares, espera-se que estudantes de áreas como educação, administração e economia também possam também ter proveito do livro.

## 1.5 Formato do livro

O livro foi pensado para ter uma estrutura (1) organizada, (2) linear e (3) formada por capítulos autossuficientes escritos para responder questões específicas. Acredito que, assim, ele possa atender tanto estudantes interessados em ler a obra inteira, como aqueles que buscam informações mais específicas sobre um tópico particular.

Esse formato adotado tende a gerar uma percepção diferente entre aqueles que consultarem apenas um capítulo ou outro e aqueles que lerem o conteúdo por completo. Há uma maior chance disso ocorrer em capítulos sobre testes estatísticos. Uma vez que diversos testes estatísticos são casos particulares de outros, alguns assuntos que parecem destoantes em uma leitura inicial, tornam-se articulados em outros capítulos.

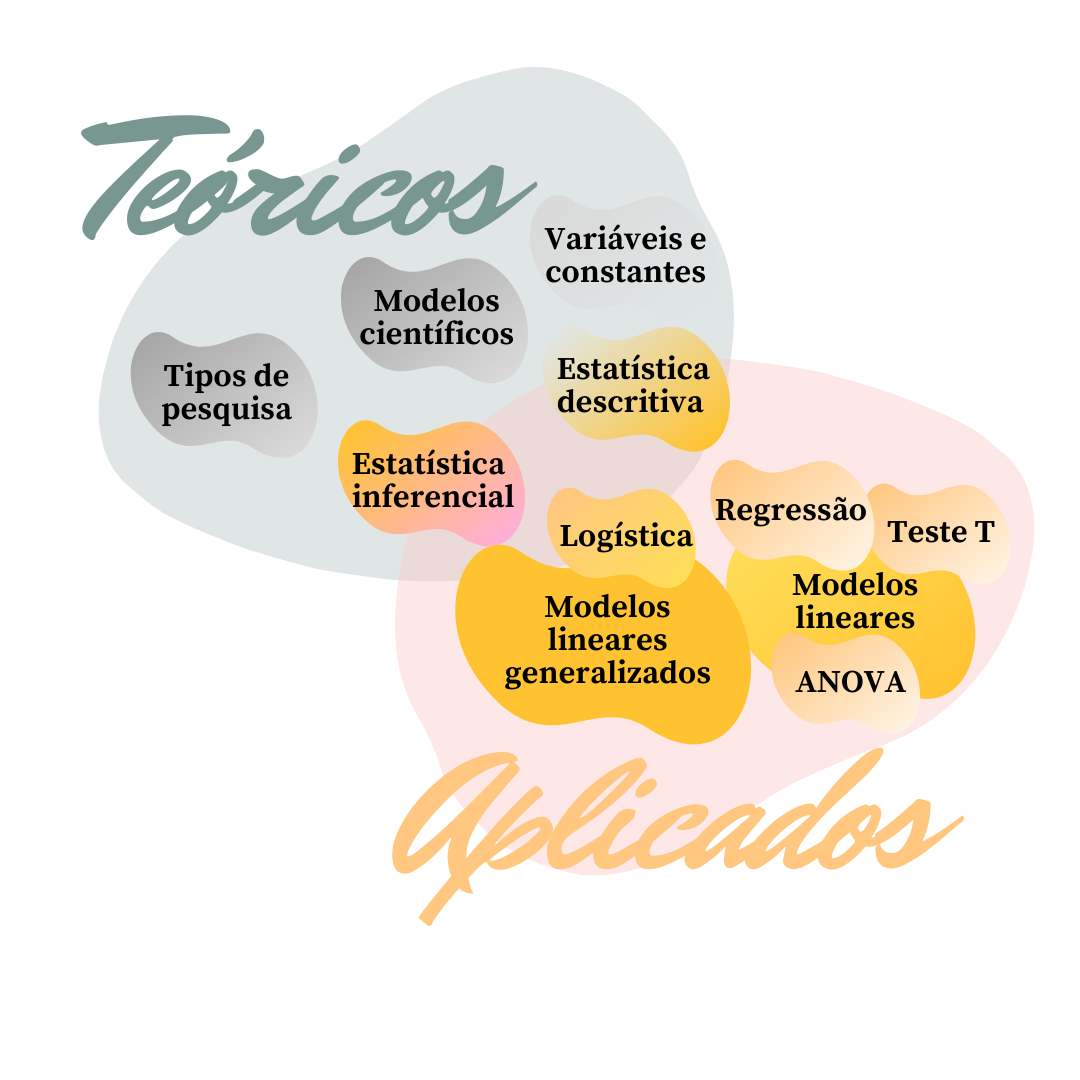
Muitos capítulos recebem o nome de testes de hipóteses (ex: Teste T ou Regressão). Isso foi intencional e visa auxiliar estudantes que precisem apenas de informações pontuais, bem como tende a enfraquecer a ideia de uma relação ponto a ponto tipicamente feita entre testes estatísticos e delineamentos específicos.

## 1.6 Como usar este livro

O livro é formado por dois componentes: capítulos teóricos e capítulos voltados à análise de dados. Os capítulos teóricos reúnem alguns conceitos fundamentais de pesquisa e estatística, tais como tipos de variáveis, delineamento de pesquisa e técnicas de amostragem. Estes capítulos foram escritos pensando em alunos de graduação do curso de Psicologia. Tenho a impressão que esses capítulos serão pouco acessados, apesar de importantes.

Os capítulos analíticos são focados em testes de hipóteses e contam com uma metodologia direta ao ponto, em que atividades similares às realizadas nos artigos são demonstradas. Estes capítulos foram desenvolvidos para estudantes de pós-graduação. Acredito que esses capítulos serão bastante acessados.

A figura abaixo diagrama os dois componentes de forma aproximada.



## 1.7 Pesquisas e dados

Neste livro, as seguintes pesquisas e seus materiais são utilizados:

* [“Depression and Anxiety Symptoms in a Representative Sample of Undergraduate Students in Spain, Portugal, and Brazil”](https://doi.org/10.1590/0102.3772e36412)
* [“Confirmatory analysis and normative tables for the Brazilian Ages and Stages Questionnaires: Social–Emotional”](https://onlinelibrary.wiley.com/doi/abs/10.1111/cch.12649)
* [Psychometric properties of a short-term visual memory test (MEMORE)"](https://www.neuropsicolatina.org/index.php/Neuropsicologia_Latinoamericana/article/view/545)
* [“A relação entre o nível de Empreendedorismo (TEG) e os aspectos sociodemográficos dos Taxistas cooperados da cidade de Santo André/São Paulo, Brasil”](https://www.metodista.br/revistas/revistas-metodista/index.php/REGS/article/view/6453)
* [“Avaliação psicométrica em português do indicador de dor crônica de Helsinki em cães com sinais crônicos de osteoartrite”](https://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-09352019000100109)
* [“Aspects Related to Body Image and Eating Behaviors in Healthy Brazilian Undergraduate Students”](https://www.researchgate.net/publication/323729370_Aspects_Related_to_Body_Image_and_Eating_Behaviors_in_Healthy_Brazilian_Undergraduate_Students)
* “Parent-reported diagnosis of Attention Deficit Hyperactivity Disorder and psychostimulant use among children and adolescents: a population-based nationwide study”
* “Resilience and vulnerability in adolescents with primary headaches: a cross-sectional population-based study”

As bases são *Open Science*. Isso significa que elas são gratuitas e universalmente acessíveis para finalidades acadêmicas. Em cada capítulo, as bases irão aparecer na seção “Pesquisa”, da seguinte maneira:

A base desta pesquisa está disponível em formato **R (Rdata)** e em **CSV**, que é lido pelo JASP. Clique na opção desejada.

Base R: Base R  
Base JASP: Base CSV

As bases em R tem formato .RData e as bases para o JASP tem formato .CSV.

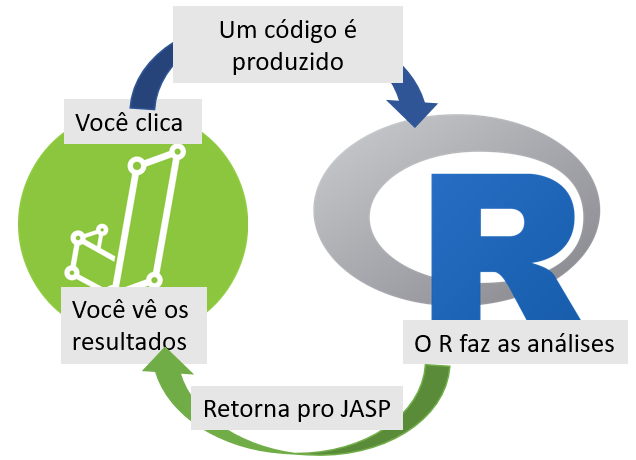
## 1.8 O R e os pacotes

O livro é integralmente desenvolvido pelo recurso de “programação letrada” no R Markdown, ou seja, ele entrelaça aspectos textuais e linhas de código. Em todos os capítulos, as funções nativas do R e do Tidyverse serão utilizadas. Caso alguém queira reproduzir as análises, será necessário apenas executar as linhas de código disponíveis no decorrer do livro.

O tidyverse costuma ter atualizações frequentes. Caso um alerta de deprecated seja apresentado, isso significa que a função utilizada foi parcialmente desativada, o que não costuma impactar nas análises.

## 1.9 JASP

O JASP é um programa gratuito que tem sido cada vez mais utilizado em Psicologia. Ele é feito integralmente por código aberto e sua interface é bastante amigável e intuitiva. Ao instalar o JASP, o R também será instalado em seu computador e ficará no pano de fundo. Dessa maneira, todas as ações feitas por *Point and Click* no JASP, serão convertidas em linhas de código no R e apresentadas de maneira dinâmica no JASP.



Em todos os capítulos, telas do JASP serão apresentadas para que seja possível a reprodução integral de algumas análises. Da mesma forma que qualquer pacote estatístico, o JASP é atualizado frequentemente. Esse livro contou com a versão 0.14.1 e espero que futuras atualizações não comprometam a proposta do livro.

## 1.10 Outros recursos

Em cada um dos capítulos, aplicações da estatística e referências bibliográficas serão apresentadas. Tenha em mente que há um debate intenso em diferentes conceitos de estatística, da mesma forma que muitas condições computacionais podem aparecer durante a execução das análises propostas. Eu recomendo fortemente a comunidade [stackoverflow](https://stackoverflow.com/) como um recurso pedagógico para auxiliar em ambos os cenários.

Questões relacionadas aos capítulos são listadas de forma a conectar o conteúdo do livro com exigências balizadas por critérios externos, tal como o ENADE e bancas de concurso.

## 1.11 Capa

Por tradição, livros de Ciência de Dados e Estatística utilizam a imagem de algum animal na capa. Há livros com cachorros, papagaios, peixes, carangueijos, lagartos, etc. Esse livro não foge dessa regra e tem como capa a Jolie, a minha cachorrinha com a Anna. Ela foi indispensável para o atraso ao término deste livro.

## 1.12 Versão do livro

Como todos os livros, este também tem uma história de desenvolvimento. A tabela abaixo apresenta a versão, a data de lançamento e algumas características importantes.

|  |  |  |
| --- | --- | --- |
| Versão | Data | Características |
| Beta 1 | Fevereiro, 2020 | Primeira versão. Baixa revisão textual e dos conceitos estatísticos. Erros são esperados. A utilização deve ser feita apenas de maneira incipiente |

## 1.13 Autor

[Luis Anunciação](http://lattes.cnpq.br/3982200733248687) é doutor em Psicometria pela Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), com intercâmbio na University of Oregon, mestre em saúde pública pela Universidade do Estado do Rio de Janeiro e especialista em Neuropsicologia (IBNeuro) e Bioestatística (Johns Hopkins University). Atualmente, é professor do Departamento de Psicologia da PUC-Rio, coordenador da ANOVA e psicometrista da Nila Press, uma editora especialzada no desenvolvimento de instrumentos psicológicos.

## 1.14 Revisão técnica

Este livro contou com a revisão técnica do Dr. J. Landeira-Fernandez (Doutor em Neurociências pela University of California - UCLA, Pesquisador 1A do CNPq) e Ms. Regina Albanense (Graduada em Matemática e Estatística e Mestre em Ciências). Dr. Landeira foi fundamental em tópicos envolvendo aspectos de pesquisa e análises estatísticas e Ms. Regina foi fundamental para revisão da modelagem matemática e revisão conceitual.

## 1.15 Agradecimentos

Nenhum homem é uma ilha. Este livro só foi possível graças a um conjunto de pessoas que auxiliaram e fizeram uma profunda revisão do texto. Meus sinceros agradecimentos a (ao):

J. Landeira-Fernandez, PUC-Rio  
Regina Albanense, CONRE  
Cristiano Fernandes, PUC-Rio  
Danilo Assis Pereira, IBNeuro  
Anna Carolina de Almeida Portugal, UFRJ  
Emanuel Cordeiro, UFPE  
Alunos da PUC-Rio, UFRJ, IBNeuro e ANOVA

# 2 Programas estatísticos

**Objetivos do capítulo**  
1. Apresentar os programas estatísticos utilizados durante o livro  
2. Discutir características do R, tidyverse e seus pacotes  
3. Discutir características do JASP

Programas estatísticos são ferramentas indispensáveis tanto na gestão, como na análise dos dados resultantes de uma pesquisa. Eles servem para otimizar o tempo gasto nas etapas analíticas de uma pesquisa, apesar de, em menor escala, permitirem a execução de análises inadequadas. No dia a dia de um pesquisador, apenas muito raramente as análises são feitas manualmente. Dessa maneira, o conhecimento de programas de análise de dados faz parte das competências esperadas para quem deseja ou precisa trabalhar com estatística.

Atualmente, há muitos programas e pacotes estatísticos disponíveis para uso. Acredito que a maioria deles tenha mais similaridades do que diferenças e produzam resultados confiáveis. Neste livro, o R e o JASP serão utilizados e algumas de suas características serão descritas neste capítulo.

## 2.1 O R

O R é uma linguagem de programação focada em análises estatísticas, que vem ganhando popularidade entre pesquisadores e cientistas. Este livro foi integralmente feito e baseado no ambiente R que, apesar de ainda não ser o programa mais frequente em Psicologia, apresenta diversas vantagens em comparação aos programas mais usuais.

* o R e todos os seus pacotes e otimizações são gratuitos.
* o R é uma linguagem de programação desenvolvida especificamente para Estatística.
  + Diferente de uma linguagem mais geral (por exemplo, Python) ou de um programa *point and click*, o R é uma linguagem focada em análise estatística. Ao se programar utilizando o R, o usuário tem controle total das ações realizadas e dos resultados obtidos. Assim, raramente o R apresentará resultados excessivos e distantes das análises solicitadas. Apesar disso poder assustar no início, acredito que essa característica seja essencial e, inclusive, serve como um excelente auxílio pedagógico para que o estudante planeje adequadamente as análises de interesse, em vez de apenas selecionar parte de um *output* padronizado, como ocorre com o SPSS.
* O R é um programa de nicho em Estatística. Essa característica faz com que ele seja absolutamente adaptado para o dia a dia em estatística, incluindo não apenas análises descritivas e inferenciais, mas também análises para simulação e controle de resultados.
* O R permite o desenvolvimento de interfaces web e aplicativos.
* O R tem diversos pacotes.
  + Pacotes são complementos que permitem otimizar as análises que o R faz. A comunidade R tem um exército de pacotes, que além de gratuitos, foram verificados publicamente. O ambiente CRAN (*The Comprehensive R Archive Network*) é o local em que estes pacotes estão alocados. Neste livro, todos os capítulos contam com pacotes específicos, que permitiram que análises complexas fossem realizadas com poucos comandos.
* O R possui uma enorme comunidade de apoio.
  + Os usuários do R formam uma rede muito dinâmica e que oferece grande apoio em caso das mais diversas dúvidas. A comunidade [stackoverflow](https://stackoverflow.com/) talvez seja a mais voluma e reúne pessoas de todas as nacionalidades.

Para baixar o R, é necessário ir no site <https://cran.r-project.org/>. Em seguida, para baixar o R Studio, é necessário acessar <https://rstudio.com/>.

## 2.2 Tidyverse

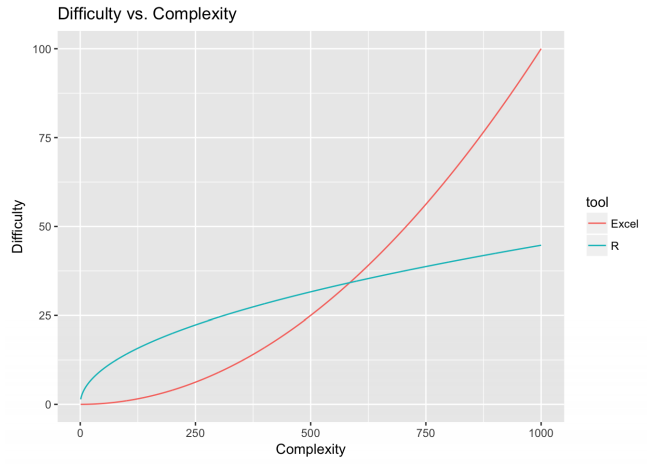
O Tidyverse é um ambiente de pacotes. Eles funcionam de maneira totalmente integrada e permitem que a estrutura da programação seja mais intuitiva e próxima à forma pela qual pensamos. No Tidyverse, os códigos seguem a lógica de sujeito + verbo e permitem uma programação encadeada, ao se utilizar a ligação *pipe* (%>%). Ao instalar o tidyverse install.package("tidyvese"), os pacotes abaixo ficam disponíveis no R.



## 2.3 Dificuldades esperadas

Algumas dificuldades são esperadas quando se trabalha programação no geral e com o R especificamente. Apesar do R e seus pacotes oferecem excelentes ferramentas para análise de dados, algumas condições descritivas são demasiadamente custosas. Por exemplo, enquanto realizar algumas tabelas e gráficos no Excel é tremendamente fácil, com alguma frequência o R exige diversas linhas de código para isso.

Nesse sentido, na relação entre dificuldade e complexidade, o R sai na frente em tarefas complexas (como exemplo, estimar os coeficientes de um modelo não-linear), mas talvez perca em tarefas fáceis (por exemplo, gerar uma tabela de contingência). A Figura a seguir apresenta esta relação comparando as análises feitas no R e no Excel.



Além disso, o R traz uma outra barreira importante em aspectos que envolvam o ensino de estatística, especialmente na graduação. Como o R é uma linguagem de programação, o estudante teria de aprender a programar antes de conseguir entender conceitos de pesquisa e a utilidade da estatística. Isso poderia impactar negativamente na motivação do estudante, principalmente àqueles com uma aversão *a priori* da matéria,

Em situações como esta, talvez o ideal seja começar motivando o estudante a entender como a estatística é uma ferramenta importante para tomar decisões para, só depois e lentamente, apresentar aspectos matemáticos e computacionais.

## 2.4 Verbos do dplyr

Entre os pacotes do ambiente tidyverse, o dplyr é o que será mais utilizado. Este pacote funciona de maneira muito intuitiva, em que as funções dependem de uma estrutura sujeito %>% verbo(complemento). Essa é uma diferença importante em relação ao R Base. Por exemplo, no R Base é necessário executar names(dataset) para verificar as variáveis de um conjunto de dados. Pelo dplyr deve-se utilizar dataset %>% names.

O dplyr funciona a partir de verbos declarativos e os principais estão listados na tabela a seguir. As sintaxes deixadas no decorrer do livro também permitem uma melhor apreensão das funcionalidades.

|  |  |
| --- | --- |
| Verbo | Ação |
| glimpse | Inspeciona os dados |
| count | Conta os níveis de uma variável |
| select | seleciona uma variável específica |
| filter | Filtra os resultados por um nível específico de uma variável |
| group\_by | Agrupa os resultados por níveis de uma variávei específica |
| summarise | Apresenta sumários (com medidas estatísticas) |
| mutate | Cria novas variáveis ou altera as existentes |
| arrange | Organiza a apresentação dos resultados |
| left\_join | Junta bases ou colunas |
| pivot\_longer | Transforma uma base larga em longa |
| pivot\_wider | Transforma uma base longa em larga |

*Nota: Em alguns momentos, em função da praticidade computacional, algumas sintaxes vão contar com o formato base do R.*

É importante ficar atento às atualizações do dplyr e do sistema tidyverse como um todo. Eventualmente, mudanças podem ocorrer e impossibilitar (ou dificultar) a reprodução de rotinas antigas.

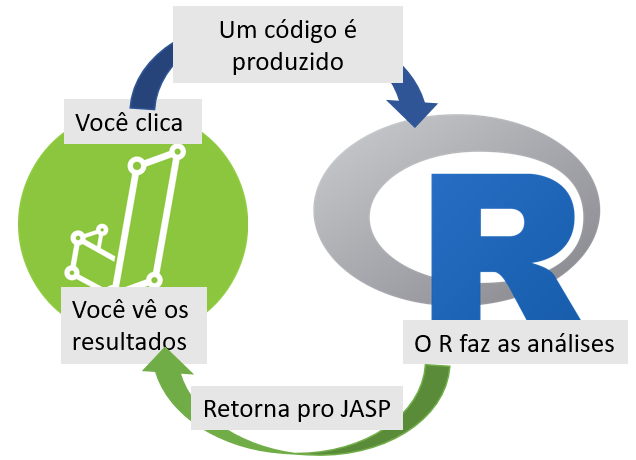
## 2.5 O JASP

O JASP é um programa gratuito, com uma interface amigável, *point and click* e altamente versátil para a maioria das análises realizadas. A versão utilizada neste livro é a 0.14.1. O JASP foi desenvolvido por um time de psicólogos e estatísticos liderados pelo Prof. Eric-Jan Wagenmakers, da Universidade de Amsterdam. Isso talvez explique o motivo pelo qual o JASP vem sendo cada vez mais utilizado em análises de dados psicológicos e na docência de matérias relacionadas a métodos estatísticos.

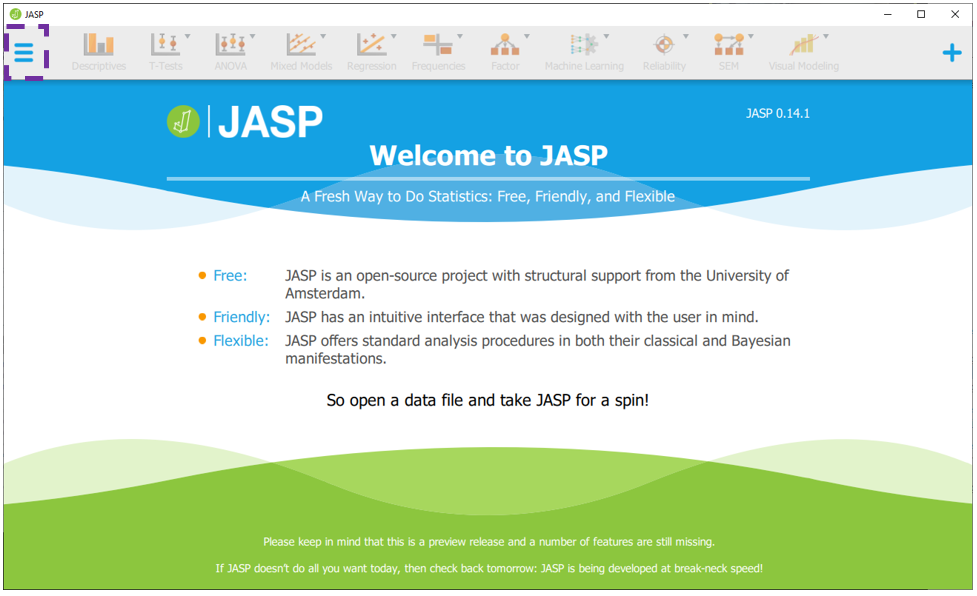
Para baixar o JASP, é necessário acessar <https://jasp-stats.org/download/>.



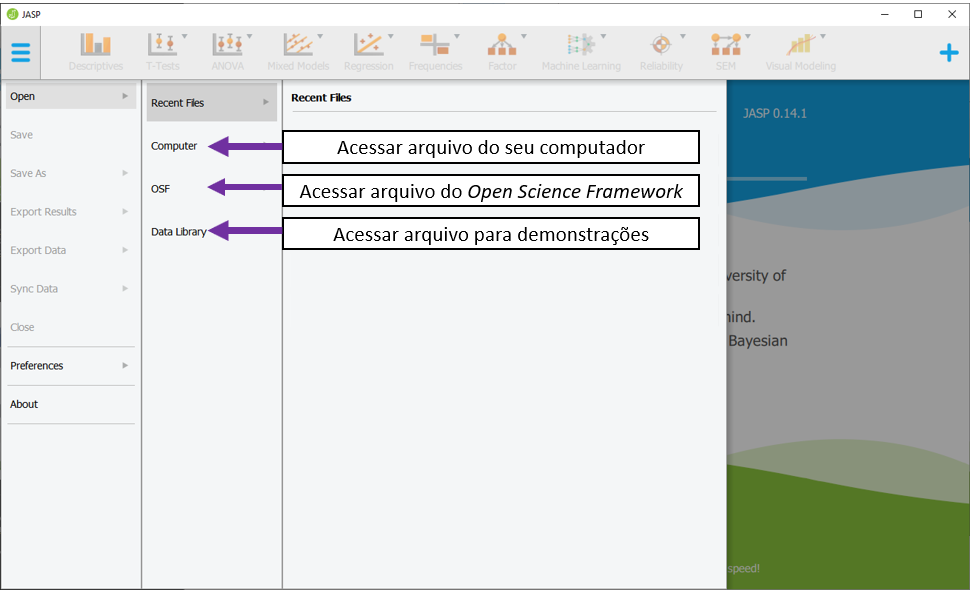
Quando se instala o JASP no computador, se instala também o R. Todas as ações feitas no JASP se transformam em linhas de código que são enviadas ao R e, em seguida, retornam pro JASP e são apresentadas na tela. Isso ocorre de maneira instantânea e não há nenhum incômodo para o usuário. Na maioria das vezes, computadores pessoais conseguem rodar o JASP sem grandes problemas.



O idioma oficial do JASP é o inglês e, nesta versão, não pode ser alterado. Para fazer qualquer análise, é necessário carregar um arquivo de dados. Isso pode ser feito no símbolo das três linhas, localizado na parte superior à esquerda.

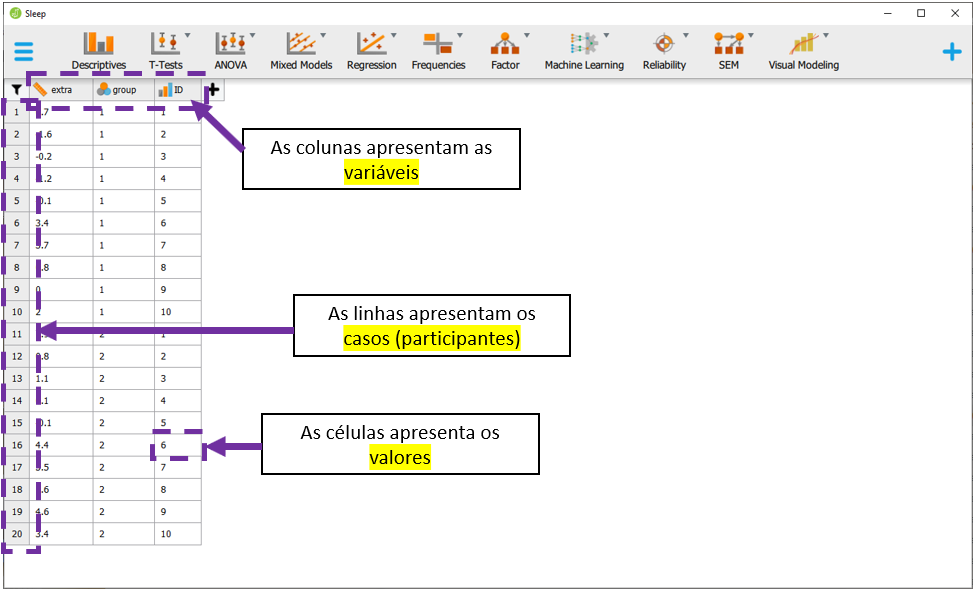


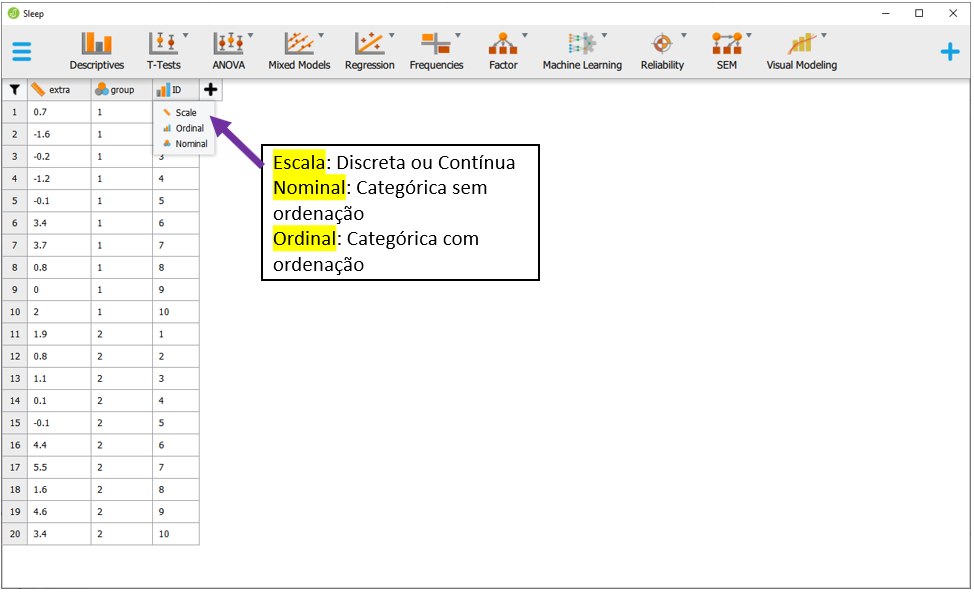
A opção open permite que se acesse algum diretório local específico ou se baixe os dados diretos da plataforma *Open Science Framework (OSF)*. Esta versão do JASP aceita, majoritariamente, arquivos em CSV, que indica que os dados são separados por vírgulas.



Ao abrir algum arquivo de dados, o JASP irá apresentar os dados no centro do programa e as opções de análise na parte superior. Tenha atenção que essas telas podem variar de versão para versão.

Em relação aos dados, quase sempre o formato utilizado é o largo. Neste formato, cada coluna representa uma variável, cada linha representa um caso e cada célula (ou vetor) apresenta um valor específico.

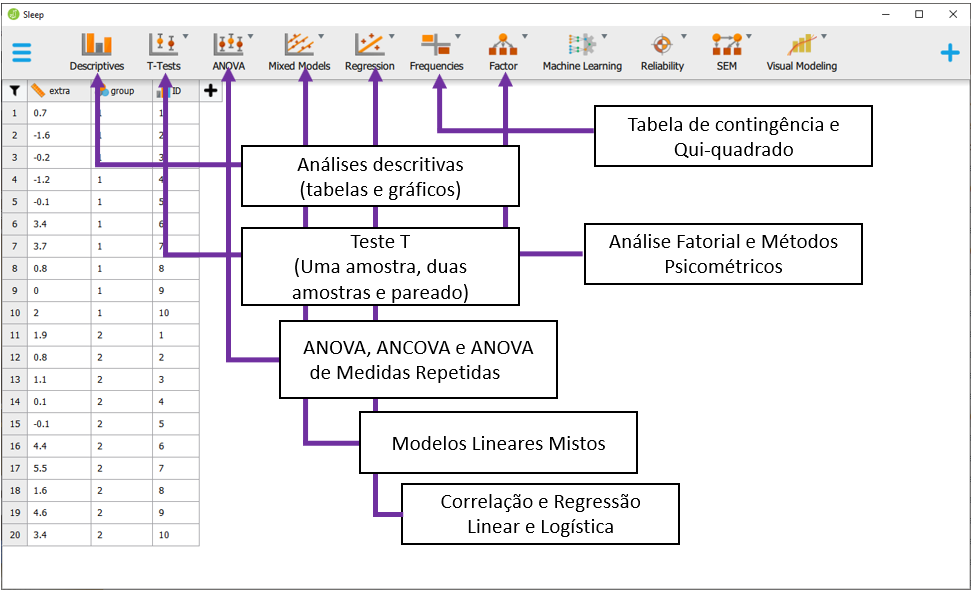
 O JASP adota uma simbologia específica para definir a escala (ou nível) de medida das variáveis. Isso pode ser visto no símbolo ao lado dos nomes das variáveis. Para alterar o nível, basta clicar sobre o símbolo.



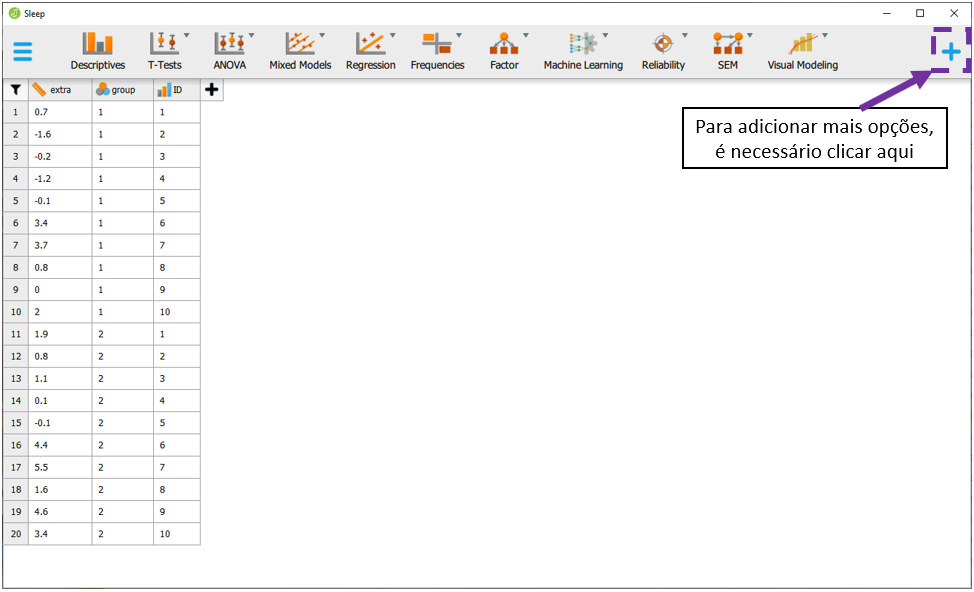
É possível também alterar valores e rótulos das variáveis. Para isso, basta clicar no centro da variável. Uma seção na parte de cima do programa será exibida. Para fechar, é necessário clicar no X ao lado direito.



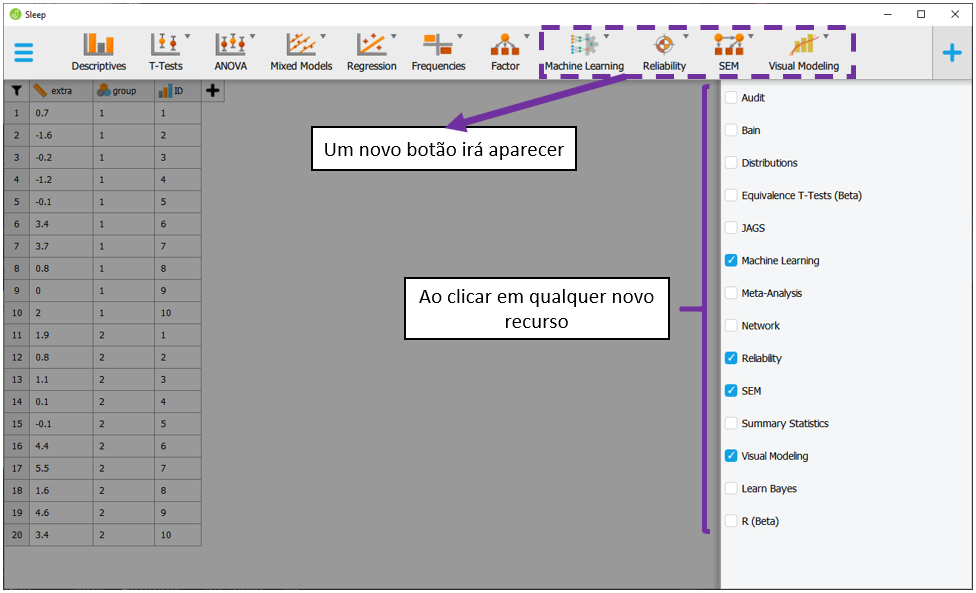
Na parte superior, o JASP oferece as análises possíveis. As opções podem variar de versão para versão, mas as principais tendem a ser as seguintes.



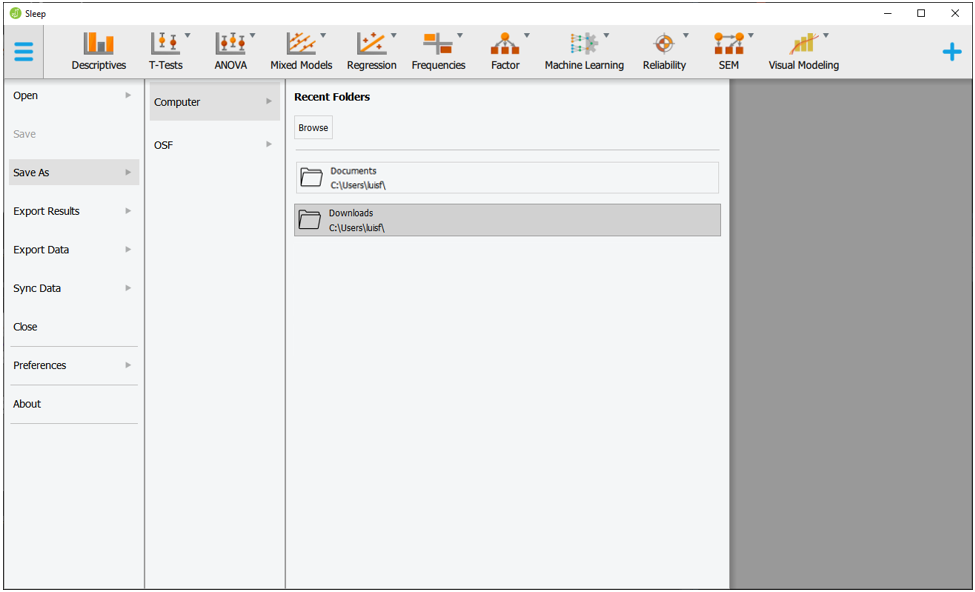
É possível adicionar módulos e complementos no JASP. Essas adições funcionam de maneira análoga aos pacotes do R. Para fazer isso, basta clicar na cruz azul ao lado direito.



Uma barra ao lado direito irá ser exibida. Qualquer opção pode ser selecionada e, ao fazer isso, novos botões irão aparecer na parte superior do programa.



Finalmente, para salvar o trabalho realizado em uma sessão, deve-se clicar na símbolo de três linhas azuis e, em seguida, clicar em Save as e selecionar a pasta.. Por padrão, o formato do arquivo sera \*.jasp.



## 2.6 Resumo

1. Há muitos programas e pacotes de estatística. Neste livro, o R e o JASP serão utilizados.
2. O R é gratuito e pode ser complementado por pacotes adicinais.
3. O tidyverse servirá como principal ecossistema de programação.
4. O JASP tem recebido grande atenção especialmente em Psicologia e as análises também serão feitas nele.

# 3 Regressão logística binária

**Objetivos do capítulo**  
1. Apresentar características da regressão logística  
2. Diferenciar o conceito de probabilidade e chance  
3. Integrar o teste Qui-quadrado, **Odds Ratio** e regressão logística

**GLOSSÁRIO**

**Modelo Linear Generalizado**: Classe de modelos compostos por uma função de ligação, preditores lineares e uma distribuição de probabilidades. São ampliações de modelos lineares.  
**Função de ligação**: Termo que associa os valores esperados da resposta aos preditores lineares no modelo.  
**Logit**: Uma função de ligação que transforma probabilidades em chances.  
**Risco**: Probabilidade ou proporção.  
**Chance ou Odds**: Caso particular de uma razão em que o numerador não está contido no denominador. Tradução para *Odds*.  
**Razão de Chances ou Odds Ratio**: Medida de tamanho de efeito que indica a chance de ocorrência de um desfecho em um grupo quando comparado a outro. Seus valores variam entre 0 e infinito.

A regressão logística é um modelo estatístico que permite estimar a chance da ocorrência de um determinado desfecho categórico (Y) em função de um ou mais preditores (X), que podem ser contínuos ou categóricos. Quando a variável dependente apresenta apenas dois níveis ou classes, a regressão é chamada de binária. Quando há mais níveis ou classes, é chamada de multinomial.

Desta maneira, é possível entender a regressão logística como um complemento da regressão linear aplicada a variáveis categóricas a partir de uma função de ligação, uma generalização do teste Qui-quadrado ou, de maneira geral, um caso particular da família dos modelos lineares generalizados (GLM), que implementa uma ligação logit.

Conceitualmente, neste modelo há os seguintes termos:

$$\overbrace{y}^{VD} = \underbrace{ln\left ( \frac{P}{1-P} \right)}\_\text{logit} \\= b\_0+b\_1X\_1+ \dots + b\_iX\_i$$

Onde:  
 representa a probabilidade de um desfecho ocorrer  
 representa a probabilidade de um desfecho não ocorrer  
 representa uma transformação logit variável dependente  
 é o intercepto  
 indica os preditores

Os seguintes pressupostos devem ser investigados:

*(i)* Os dados são aleatórios e representativos da população  
*(ii)* A variável dependente é dicotômica/binária  
*(iii)* Os preditores não apresentam alta correlação entre eles  
*(iv)* Há uma relação linear entre preditores contínuos e o logit do desfecho

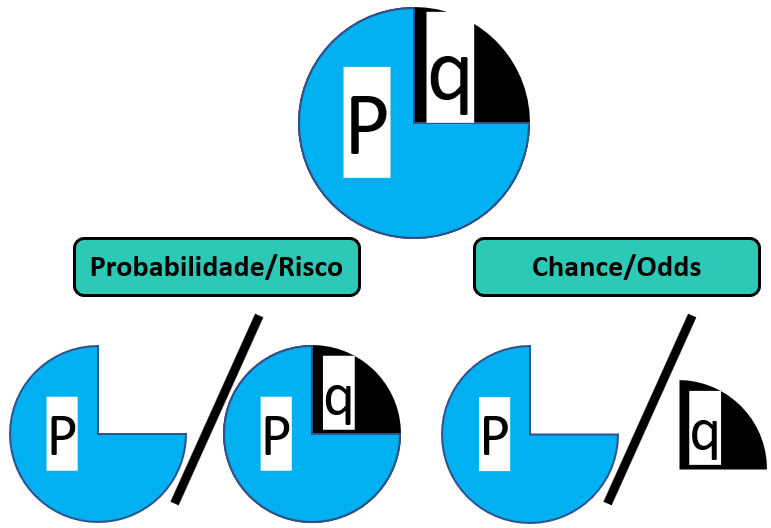
Há muitas pesquisas em que modelos logísticos são utilizados. Em áreas de saúde, eles tendem a ser os modelos de primeira escolha para verificar condições clínicas e de agravo à saúde. Em Psicologia, por exemplo, ele é útil para descrever e investigar os possíveis preditores de condições diagnósticas bem definidas, tal como TDAH e outros transtornos psiquiátricos. Em epidemiologia, eles são fundamentais em estudos do tipo caso-controle.

Algumas condições são importantes:

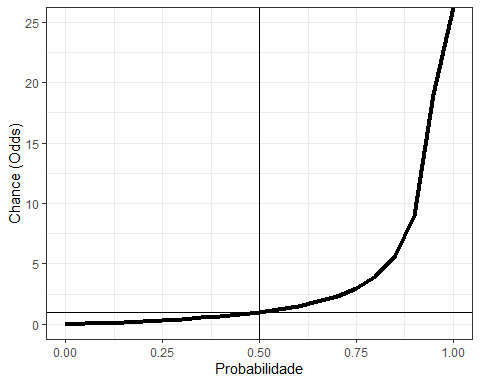
1. A variável dependente nesta equação é uma transformação logit do desfecho. Ela não é, por definição, uma probabilidade, mas sim uma função.
2. Não há a definição de um termo de erro (tal como nos modelos de regressão linear) e
3. O desfecho é assumido seguir uma distribuição Bernoulli.

Essas características fazem com que os resultados da regressão logística informem sobre chances (*Odds*) e Razão de chances (*Odds Ratio*) e não sobre probabilidades (Riscos), diretamente. Existem muitas formas de demonstrar esta diferença. Entretanto, a forma mais simples é por sua estrutura matemática.

Uma probabilidade é uma razão entre uma parte contra o todo (*Tudo o que tenho sobre tudo o que posso quero*, em jargão pedagógico). Quando se fala sobre riscos, implicitamente está se falando sobre probabilidades. Por sua vez, a chance é uma razão em que o numerador é uma probabilidade e o denominador é seu complemento. A figura a seguir ilustra essa diferença.



É legítimo ter dificuldade na interpretação do conceito, além de questionar o motivo pelo qual ele é tão utilizado nestas áreas, o que será apresentado ao fim do capítulo. No entanto, é importante perceber que quanto maior for a probabilidade, maior também serão as chances de um determinado evento e vice-versa. A relação entre ambos tem o seguinte formato.



Por sua vez, a Razão de chances (*Odds Ratio* ou OR) é uma medida de tamanho de efeito que se dá pela comparação entre dois grupos. Para computá-la, é necessário verificar a chance do desfecho ocorrer em cada um dos grupos e, em seguida, dividir os resultados.

Como em áreas de saúde, quase sempre, o interesse é verificar situações clínicas, quanto maior for o OR, maior será a chance da ocorrência de um desfecho negativo em um determinado grupo em comparação à outra e, consequentemente, maior será o risco.

A tabela a seguir apresenta aos principais resultados de OR e suas interpretações:

|  |  |  |
| --- | --- | --- |
| Chance | Probabilidade | Interpretação |
| < 1 | < 50% | Fator de risco |
| = 1 | 50% | Fator neutro |
| > 1 | > 50% | Fator de proteção |

## 3.1 Pesquisa

A base desta pesquisa está disponível em formato **R (Rdata)** e em **CSV**, que é lido pelo JASP. Clique na opção desejada.

Base R: [Base R - Headache anonymous](https://github.com/anovabr/mqt/raw/master/bases/Base%20R%20-%20Headache%20anonymous.RData)  
Base JASP: [Base CSV - headache (no names)](https://github.com/anovabr/mqt/raw/master/bases/bases_csv_jasp.zip)

Neste capítulo, vamos utilizar a pesquisa intitulada [“Resilience and vulnerability in adolescents with primary headaches: a cross-sectional population-based study”](www.google.com), publicada em 2021 no Headache, que sou coautor.

Essa é uma pesquisa censitária, que contou com todos os 339 estudantes de uma cidade pequena no interior do Brasil. O estudo visou mapear as possíveis condições de apresentar baixos recursos psicológicos em adolescentes portadores de enxaqueca de diferentes tipos e intensidades. Na literatura médica internacional, são raros os estudos com finalidade epidemiológica descritiva sobre enxaqueca. No Brasil, por sua vez, este nosso estudo foi o primeiro. Esse fato marcou o caráter inovador da pesquisa e também permitiu impactar positivamente o trabalho de clínicos que, em seu dia a dia, lidam com jovens com tais características.

Para definir o tipo e a frequência da dor de cabeça, os participantes responderam a diferentes questionários médicos. Por sua vez, para verificar o possível baixo recurso psicológico, eles foram submetidos à escala *Resiliency Scales for Children and Adolescents*, que foi desenvolvida justamente para medir esta condição.

Um dos objetivos que tivemos na execução das análises foi verificar quais eram os preditores significativos que poderiam impactar na apresentação (ou não) de baixos recursos psicológicos. Os preditores de interesse foram a idade da criança (age), sua etnia (race), sexo (sex), classe socioeconômica (ses), problemas de sono (sleeping), prematuridade (premature), fumo durante a gestão (smoking), uso de álcool durante a gestação (alcohol) e possível TDAH (sdq\_risk).

Para isso, modelos de regressão logística binária foram utilizados.

## 3.2 Execução no R

Inicialmente, é necessário carregar a base de dados ao R. Após isso feito, a apresentação de tabelas e gráficos auxilia na interpretação dos resultados. De maneira similar à feita em outros capítulos, abaixo há uma tabela descritiva feita com o pacote arsenal.

arsenal::tableby(risk\_cefaleia\_resources ~   
 age + race + sex + ses + sleeping + premature + smoking + alcohol + sdq\_risk,   
 test = FALSE,  
 base\_uso) %>% summary()

|  |  |  |  |
| --- | --- | --- | --- |
|  | 0 (N=312) | 1 (N=27) | Total (N=339) |
| **age** |  |  |  |
| Mean (SD) | 13.929 (2.068) | 13.370 (2.060) | 13.885 (2.070) |
| Range | 10.000 - 18.000 | 11.000 - 18.000 | 10.000 - 18.000 |
| **race** |  |  |  |
| N-Miss | 10 | 0 | 10 |
| white | 220 (72.8%) | 19 (70.4%) | 239 (72.6%) |
| Other | 82 (27.2%) | 8 (29.6%) | 90 (27.4%) |
| **sex** |  |  |  |
| Male | 150 (48.1%) | 8 (29.6%) | 158 (46.6%) |
| Female | 162 (51.9%) | 19 (70.4%) | 181 (53.4%) |
| **ses** |  |  |  |
| AB | 83 (26.6%) | 10 (37.0%) | 93 (27.4%) |
| C | 189 (60.6%) | 14 (51.9%) | 203 (59.9%) |
| DE | 40 (12.8%) | 3 (11.1%) | 43 (12.7%) |
| **sleeping** |  |  |  |
| N-Miss | 7 | 0 | 7 |
| no | 296 (97.0%) | 25 (92.6%) | 321 (96.7%) |
| yes | 9 (3.0%) | 2 (7.4%) | 11 (3.3%) |
| **premature** |  |  |  |
| N-Miss | 4 | 2 | 6 |
| no | 272 (88.3%) | 21 (84.0%) | 293 (88.0%) |
| yes | 36 (11.7%) | 4 (16.0%) | 40 (12.0%) |
| **smoking** |  |  |  |
| N-Miss | 1 | 1 | 2 |
| no | 243 (78.1%) | 20 (76.9%) | 263 (78.0%) |
| yes | 68 (21.9%) | 6 (23.1%) | 74 (22.0%) |
| **alcohol** |  |  |  |
| N-Miss | 2 | 1 | 3 |
| no | 278 (89.7%) | 23 (88.5%) | 301 (89.6%) |
| yes | 32 (10.3%) | 3 (11.5%) | 35 (10.4%) |
| **sdq\_risk** |  |  |  |
| no | 287 (92.0%) | 19 (70.4%) | 306 (90.3%) |
| yes | 25 (8.0%) | 8 (29.6%) | 33 (9.7%) |

Essa tabela oferece uma primeira informação sobre os dados. Gráficos também poderiam ser feitos com a mesma finalidade.

Os resultados apresentados nas tabelas são úteis e permitem calcular vários testes Qui-quadrado de independência, gerando um indicador útil sobre o perfil de associação entre cada uma das variáveis, quando elas são categóricas. Entretanto, os resultados obtidos por este teste indicariam apenas o possível relacionamento bivariado sem, no entanto, controlar os resultados por todas as outras variáveis de interesse, uma vez que elas não foram analisadas todas em conjunto.

Apesar de ser possível fazer alguns ajustes particulares ou utilizar o teste de Cochran–Mantel–Haenszel, a regressão logística é a modelagem mais adaptada para lidar com esta situação. Esta modelagem permite analisar todas as variáveis simultaneamente e também incluir preditores contínuos.

Para executar esse tipo de regressão no R, é necessário usar a função nativa glm e definir a família como binomial. Isso indicará que o desfecho é binário e que a função logit deverá ser implementada na equação.

O vetor mod\_risco será criado e reunirá os resultados do modelo de regressão logística.

mod\_risco <- glm(risk\_cefaleia\_resources ~  
 age + race + sex + ses + sleeping + premature + smoking + alcohol + sdq\_risk,   
 family = "binomial", data = base\_uso)

Há muitas formas de apresentar os resultados e o pacote sjPlot tem bons recursos para isso. A função tab\_model produz uma tabela com o *Odds Ratio* de cada preditor, seu intervalo de confiança, a estatística de teste, o valor de P, bem como a quantidade de observações utilizada e uma medida de ajuste ().

sjPlot::tab\_model(mod\_risco, show.stat = TRUE)

risk cefaleia resources

Predictors

Odds Ratios

CI

Statistic

p

(Intercept)

0.56

0.02 – 15.56

-0.34

0.730

age

0.82

0.64 – 1.03

-1.63

0.104

race [Other]

1.13

0.36 – 3.16

0.23

0.817

sex [Female]

3.07

1.16 – 9.34

2.15

0.032

ses [C]

0.51

0.18 – 1.44

-1.29

0.197

ses [DE]

0.67

0.12 – 2.91

-0.50

0.617

sleeping [yes]

1.18

0.13 – 7.05

0.17

0.868

premature [yes]

1.59

0.41 – 4.97

0.75

0.456

smoking [yes]

0.88

0.22 – 2.93

-0.20

0.840

alcohol [yes]

0.96

0.18 – 3.75

-0.06

0.951

sdq\_risk [yes]

7.53

2.51 – 22.36

3.67

<0.001

Observations

314

R2 Tjur

0.101

A interpretação dos resultados costuma ser feita de maneira gradual, em que cada elemento é apresentado e discutido.

Inicialmente, é importante verificar se o modelo testado é significativamente mais informativo do que um modelo nulo. Isso é feito por uma estatística de desvio (*Deviance*) e pelos critérios de informação de Akaike e Bayesiano. Por padrão, esta tabela assume esta condição, mas não a apresenta formalmente. Isso pode ser feito pela função nativa anova, tal como demonstrado a seguir. É importante ter atenção que esse tipo de diagnóstico é particularmente importante em modelos com menos variáveis.

#criar um null model  
mod\_nulo <- base\_uso %>%   
 filter(across(c(age,race,sex,ses,sleeping,premature,smoking,alcohol,sdq\_risk), ~!is.na(.x))) %>%   
 glm(risk\_cefaleia\_resources ~ 1, family = "binomial", data = .)  
#comparar modelos  
anova(mod\_nulo, mod\_risco, test = "LRT")

O segundo momento é relacionado à análise do , disposto na parte inferior da tabela. Existem diferentes maneiras de calculá-lo e sua interpretação não é exatamente igual à feita na regressão linear. Enquanto na regressão linear, o se refere à proporção da variância explicada, em modelos logísticos, no geral, o indica o quanto o modelo testado é próximo de um modelo com ajuste perfeito aos dados (saturado) (Portugues, [2020](#ref-portugues2020)).

O é a medida utilizada nesta apresentação e seus valores variam entre 0 e 1. Ele computa a diferença entre os dados e os valores previstos pelo modelo testado. Quão maior o valor, mais discriminativo é o modelo testado. Neste caso, o o foi de 0.10. Não há uma regra geral para interpretar o . Em Psicologia, por exemplo, é bastante difícil que ele chegue a 0.5.

A terceira interpretação é baseada nos *Odds Ratio* (OR) e sugere alguma com atenção: o OR indica a chance de ocorrência de um desfecho caso o preditor analisado tenha ocorrido, em comparação com sua não ocorrência. Ele não indica diretamente sobre a probabilidade, apesar disso ser possível de ser feito.

Os preditores sex[female] e sqd\_risk foram significativos. A interpretação pode ser feita da seguinte maneira:

* A chance do participante vir a apresentar baixo recurso psicológico é 3.07 maior em meninas do que em meninos.
* A chance do participante vir a apresentar baixo recursos psicológico é 7.53 maior em participantes com TDAH do que em participantes sem TDAH.

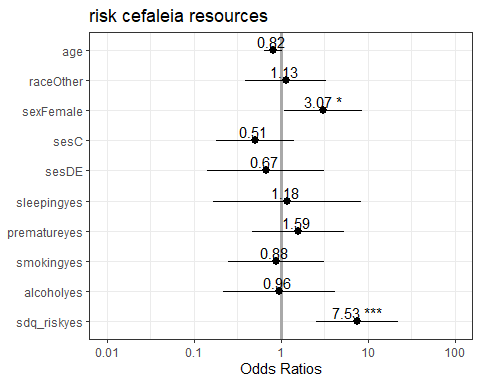
Eventualmente, variáveis contínuas também podem ser significativas. Caso a variável age tivesse sido significativa, sua interpretação poderia ser:

* O aumento de 1 ano de idade impacta a chance de vir a apresentar baixos recursos psicológicos em 0.82, indicando uma característica protetiva.

É importante ter em mente que esses resultados são baseados nas observações válidas. Por definição, o R não utiliza casos ausentes, o que ocasionou uma redução de 7% das observações (de 339 para 314, neste caso). É provável que os achados fossem diferentes caso estratégias de imputação de casos ausentes tivessem sido implementadas.

De maneira análoga a outros modelos estatísticos, a apresentação dos resultados é muito beneficiada por gráficos, o que é feito pelo *Forest plot* neste caso. Neste gráfico, o eixo horizontal apresenta os valores possíveis de OR () e é centralizado em 1. O eixo vertical lista todos os preditores analisados e seus intervalos de confiança. A interpretação é feita ancorada no valor 1, que quando está contida no intervalo de um determinado resultado, indica que os resultados não são significativos. Valores de OR acima de 1 se situam à direita e indicam um possível fator de risco. Valores de OR abaixo de 1 se situam à esquerda e indicam um possível fator protetivo.

sjPlot::plot\_model(mod\_risco,  
 grid = FALSE,   
 show.values = TRUE, value.offset = .3,   
 colors = "bw",  
 vline.color = "darkgray") +  
 scale\_x\_discrete(labels = function(x) str\_wrap(x, width = 10)) +  
 theme\_bw() +  
 theme(legend.position = "top")



Caso haja interesse, é também possível computar as probabilidades e adicioná-las à tabela aplicando . Ao fazer isso, uma nova informação será apresentada e poderá ter alguma utilidade para interpretar preditores variáveis categóricas. Tradicionalmente, isso não é feito.

**Atenção**: A validade das inferências dos resultados depende da adequação ou não dos pressupostos dos testes estatísticos. A avaliação destas condições é parte de um procedimento diagnóstico que deve ser sempre feito.

Um aspecto importante é que a validade da interpretação dos resultados depende dos pressupostos do modelo estatístico. A violação destes pressupostos distorce, limita ou invalida as interpretações teóricas propostas, uma vez que tanto o aumento do erro do tipo 1 (falso positivo), como do tipo 2 (falso negativo) podem ocorrer (Barker & Shaw, [2015](#ref-Barker2015); Ernst & Albers, [2017](#ref-Ernst2017); Lix et al., [1996](#ref-Lix1996)). Corriqueiramente, testar os pressupostos é uma etapa anterior à própria realização do teste inferencial. Entretanto, pedagogicamente a apresentação deles após a execução do teste parece mais adequada. Assim, eles serão testados a seguir.

Multicolinearidade: A multicolinearidade pode ser investigada pela análise chamada *Variance Inflaction Factor* (VIF). Essa análise verifica o quão correlacionados são os preditores e gera um resultado numérico. Valores abaixo de 4 são tipicamente utilizados para indicar que os preditores não são fortemente correlacionados e, consequentemente, considerar este pressuposto atendido.

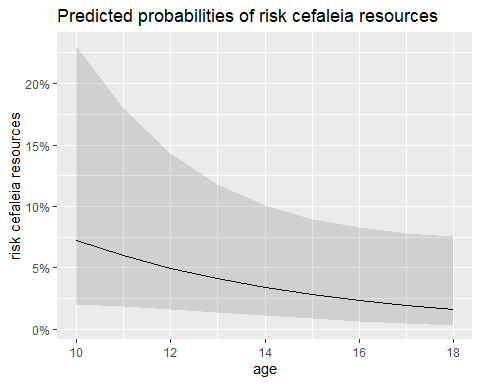
car::vif(mod\_risco) %>% pander()

Registered S3 methods overwritten by ‘car’: method from influence.merMod lme4 cooks.distance.influence.merMod lme4 dfbeta.influence.merMod lme4 dfbetas.influence.merMod lme4

|  |  |  |  |
| --- | --- | --- | --- |
|  | GVIF | Df | GVIF^(1/(2\*Df)) |
| **age** | 1.035 | 1 | 1.017 |
| **race** | 1.112 | 1 | 1.055 |
| **sex** | 1.078 | 1 | 1.038 |
| **ses** | 1.294 | 2 | 1.067 |
| **sleeping** | 1.17 | 1 | 1.082 |
| **premature** | 1.017 | 1 | 1.009 |
| **smoking** | 1.365 | 1 | 1.168 |
| **alcohol** | 1.155 | 1 | 1.075 |
| **sdq\_risk** | 1.162 | 1 | 1.078 |

Relação linear entre preditores contínuos e o logit: A regressão logística assume que há um relacionamento linear entre os preditores contínuos e o desfecho logit. O pacote sjplot também pode ser utilizado para testar este pressuposto. Não há um teste formal específico para este pressuposto e a decisão é baseada na visualização deste relacionamento.

sjPlot::plot\_model(mod\_risco, type = "pred", terms = "age")



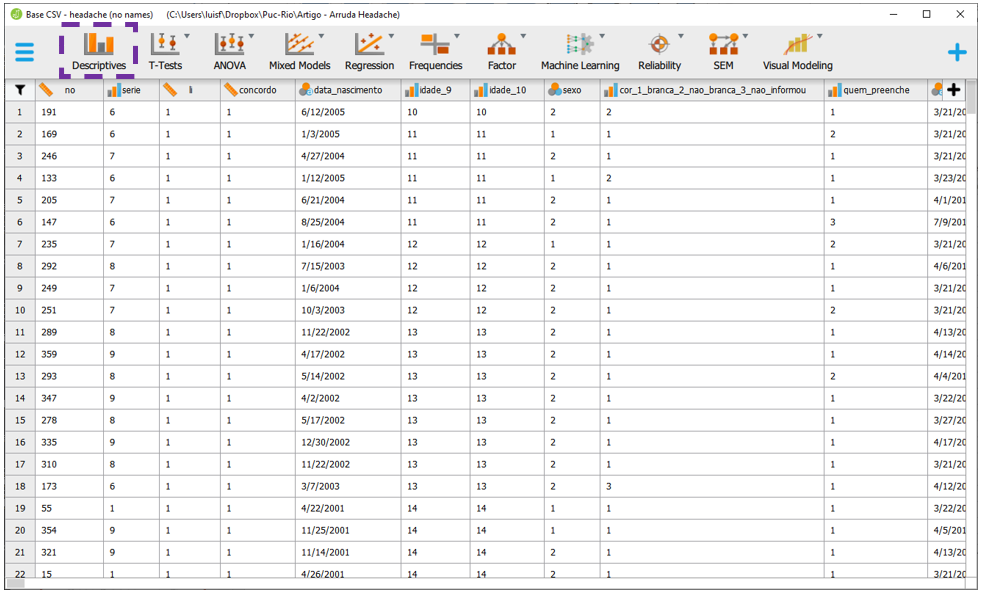
O gráfico não sugere desvio da linearidade. Com isso, esse pressuposto é também considerado mantido.

Após essas análises, é possível interpretar os resultados.

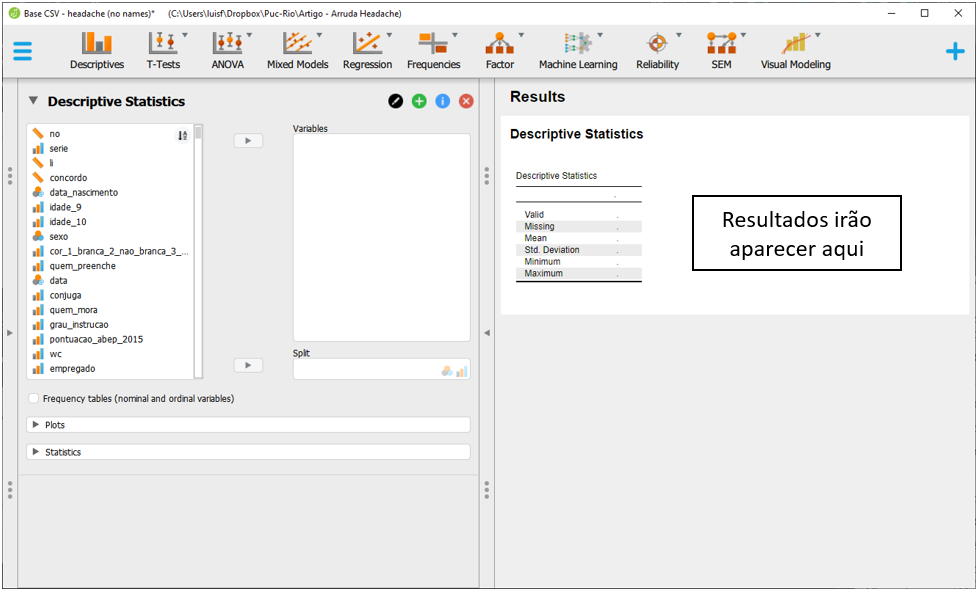
## 3.3 Execução no JASP

Inicialmente, é necessário carregar a base intitulada “Base CSV - headache (no names)” para o ambiente JASP. Essa base apresenta as variáveis que foram pesquisadas, bem como um conjunto de variáveis auxiliares que serviram para responder a outras perguntas.

A apresentação de tabelas e gráficos é sempre muito importante e pode ser feito ao clicar na opção Descriptives.

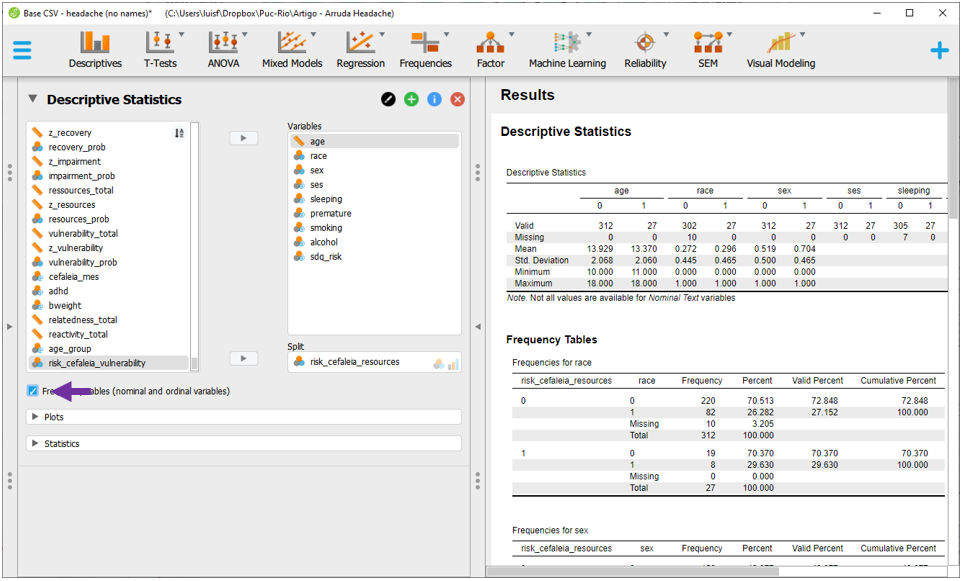


Ao clicar nesta opção, a interface do JASP ser próxima à exposta a seguir. É possível eleger as variáveis que irão ser analisadas e as variáveis que irão funcionar como agrupadores.



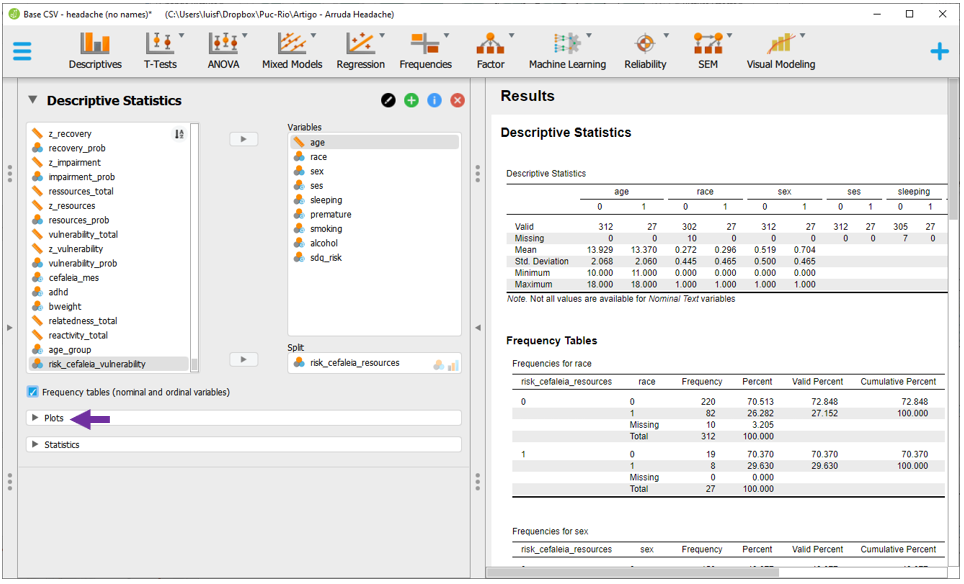
Neste caso, o interesse é agrupar os resultados em função de apresentar baixos recursos psicológicos. Neste sentido, a variável risk\_cefaleia\_resources deve ser inserida na parte Split e as variáveis age, race, sex, ses, sleeping, premature, smoking, alcohol e sdq\_risk deverão ser inseridas em Variables.

Para a apresentação ficar adequada à escala de medida, é necessário marcar a opção Frequency tables (nominal and ordinal variables), na parte inferior esquerda da interface.

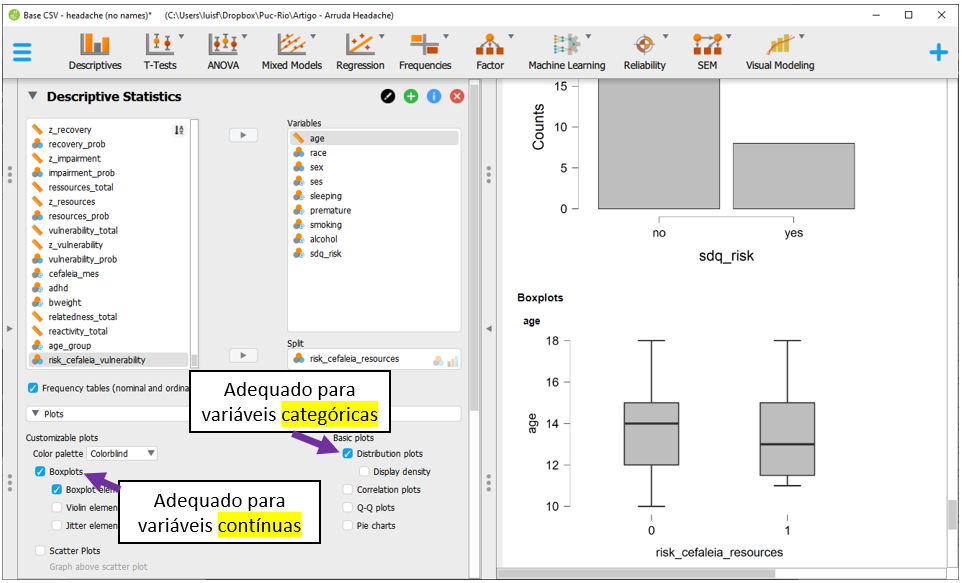


A tabela apresentada será bastante informativa e poderá ser utilizada também para uma apreensão inicial dos dados. Variáveis contínuas serão resumidas por suas médias e desvios-padrão e variáveis categóricas serão apresentadas por contagens e proporções.

Gráficos tendem a ser úteis também para verificar o formato da distribuição dos dados. o JASP permite realizá-los clicando em Plots.



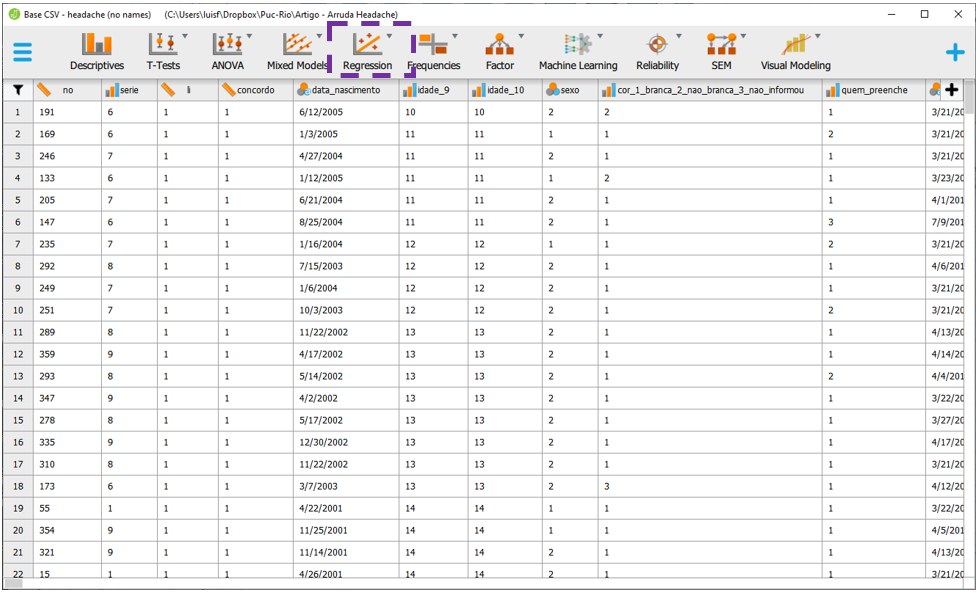
Como há variáveis contínuas (age) e variáveis categóricas (sex, etc) o ideal é apresentar tanto boxplots como gráficos de barras. Ao clicar na opção Boxplot e Distribution plots, o JASP irá realizar todos os gráficos e apresentá-los ao lado direito.



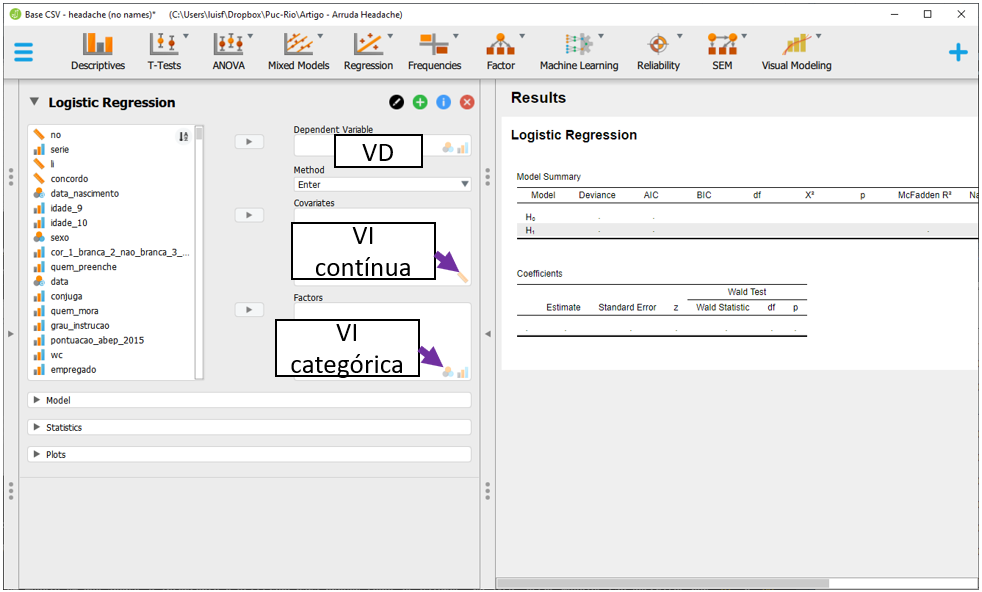
A análise conjunta da tabela previamente realizada e dos gráficos pode ser feita. Apesar deste procedimento inicial não ser uma etapa fundamental, ele tem sua aplicação para apresentar resultados descritivos da amostra.

Com isto feito, é possível fazer a regressão logística.

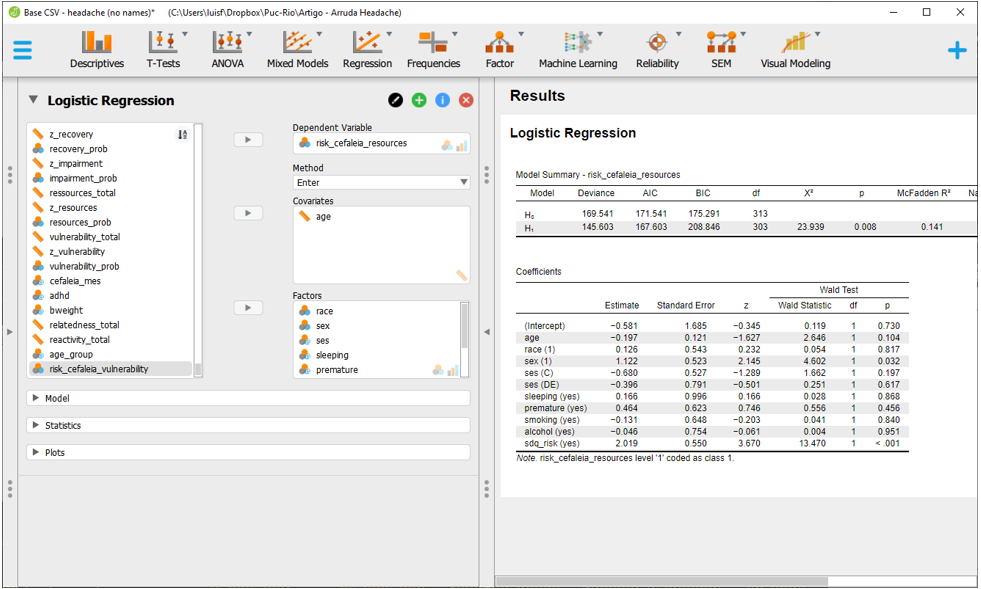
Para execução da regressão logística, será necessário clicar em Regression e Logistic Regression.



A tela do JASP irá apresentar algumas opções. O local Dependent variables é onde a VD será inserida. Repare que o JASP somente permitirá inserir uma única variável nominal ou ordinal neste espaço. Covariates é o local onde as VIs contínuas serão colocadas e Factors é onde as VIs categóricas serão colocadas.



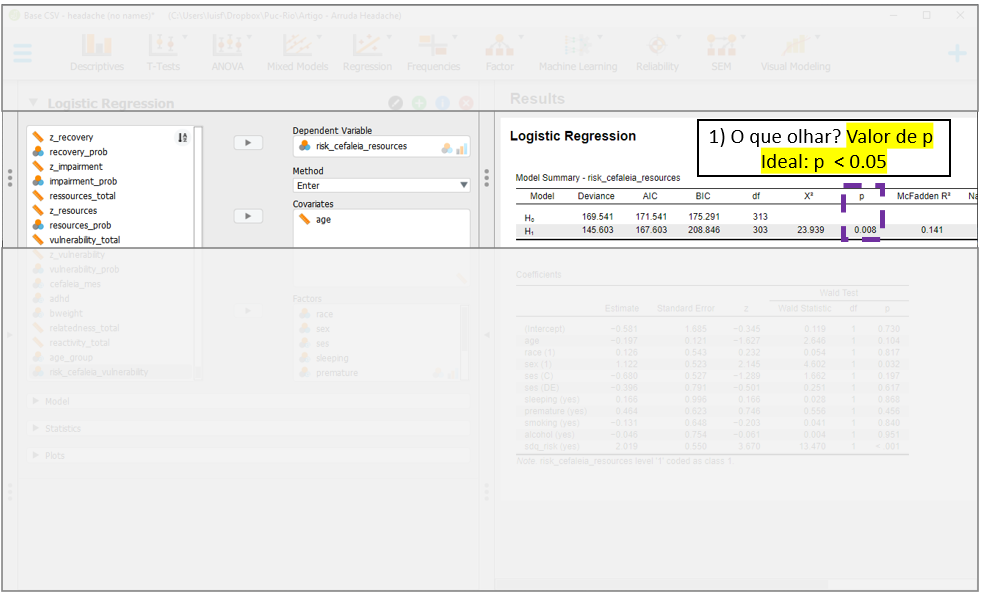
Para realizar o modelo, será necessário levar a risk\_cefaleia\_resources para Dependent Variable. Em seguida, age para seção Covariates e race, sex, ses, sleeping, premature, smoking, alcohol e sdq\_risk para Factors. Ao fazer isso, o JASP irá fazer as principais análises e apresentar os resultados em uma tabela específica, ao lado direito da tela.



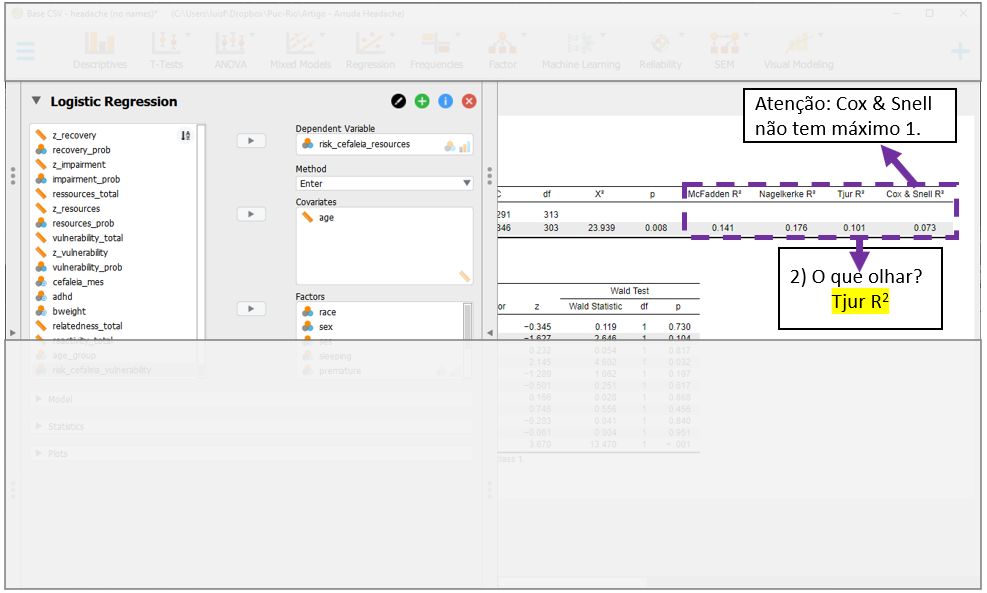
É fácil notar que os resultados apresentados são muitos. É recomendado uma ordem específica para interpretá-los.

Em primeiro momento, é importante verificar o ajuste do modelo na seção Model summary, na parte superior dos resultados. O JASP compara o modelo ajustado H1 contra um modelo nulo H0 em uma métrica de desvio (*Deviance*) e pelos critérios de informação de Akaike (AIC) e Bayesiano (BIC). De maneira similar à regressão linear, H0 representa o modelo mais simples possível (sem preditores) e H1 representa o modelo com os preditores que estão sendo testados. Valores de P e resultados baixos dos critérios AIC e BIC são utilizados como guias de decisão sobre o melhor modelo.

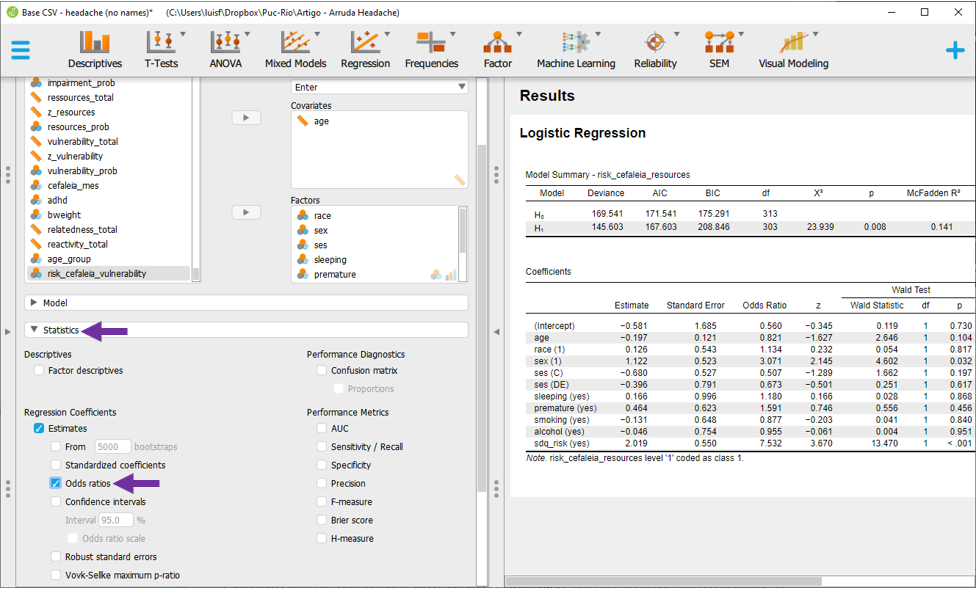
A síntese desses resultados indica que o modelo testado é mais informativo que o modelo nulo, uma vez que o valor de P foi de 0.008 e o AIC do modelo testado foi inferior, apesar do BIC ter sido superior.



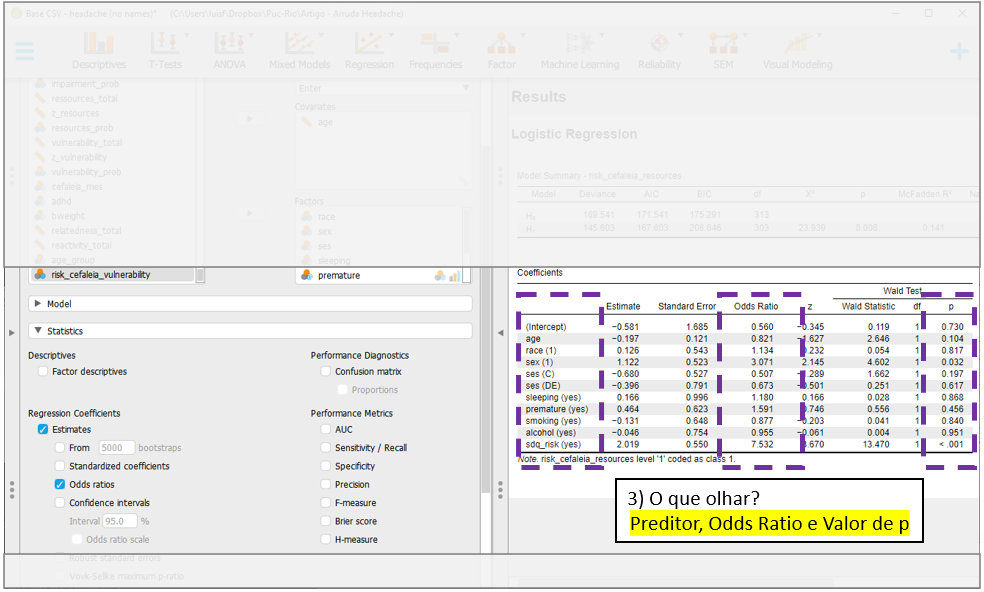
O segundo momento consiste na interpretação de uma métrica análoga ao de modelos de regressão linear. Na seção Model summary, o JASP apresenta algumas opções de uso. Existem debates na literatura sobre qual resultado é o mais indicado e, cada vez mais, o tem sido escolhido. Essa é uma métrica 0-1 que é computada verificando as diferenças médias dos resultados obtidos pelo modelo nulo e pelo modelo testado. Independentemente do escolhido, é importante ter claro que os valores do não estão contido no intervalo 0-1, o que gera dificuldade em sua interpretação. No modelo testado, o foi de 0.101. Não há uma regra geral para interpretar o . Em Psicologia, por exemplo, é bastante difícil que ele chegue a 0.5.



O terceiro momento é a análise das estimativas obtidas em cada um dos preditores listados na seção Coefficients. No entanto, o JASP apresenta o log(odds) dos resultados, cuja interpretação é pouco intuitiva. Desta maneira, é necessário adicionar o *Odds Ratio* à tabela. Para fazer isso, ao lado esquerdo da interface, é necessário clicar em Statistics e, em seguida Odds Ratio.



O *Odds Ratio* será adicionado à tabela. Agora, cada um dos preditores poderá ser analisado individualmente, com interpretação levando em consideração o OR e os valores de P.



A interpretação do *Odds Ratio* (OR) precisa ser feita com atenção: ele indica a chance de ocorrência de um desfecho caso o preditor analisado tenha ocorrido, em comparação com sua não ocorrência. Esses valores não indicam diretamente sobre a probabilidade de um desfecho, apesar disso poder ser feito.

Os preditores sex[1 - Mulheres] e sqd\_risk foram significativos. A interpretação pode ser feita da seguinte maneira:

* A chance do participante vir a apresentar baixo recurso psicológico é 3.07 maior em meninas do que em meninos.
* A chance do participante vir a apresentar baixo recursos psicológico é 7.53 maior em participantes com TDAH do que em participantes sem TDAH..

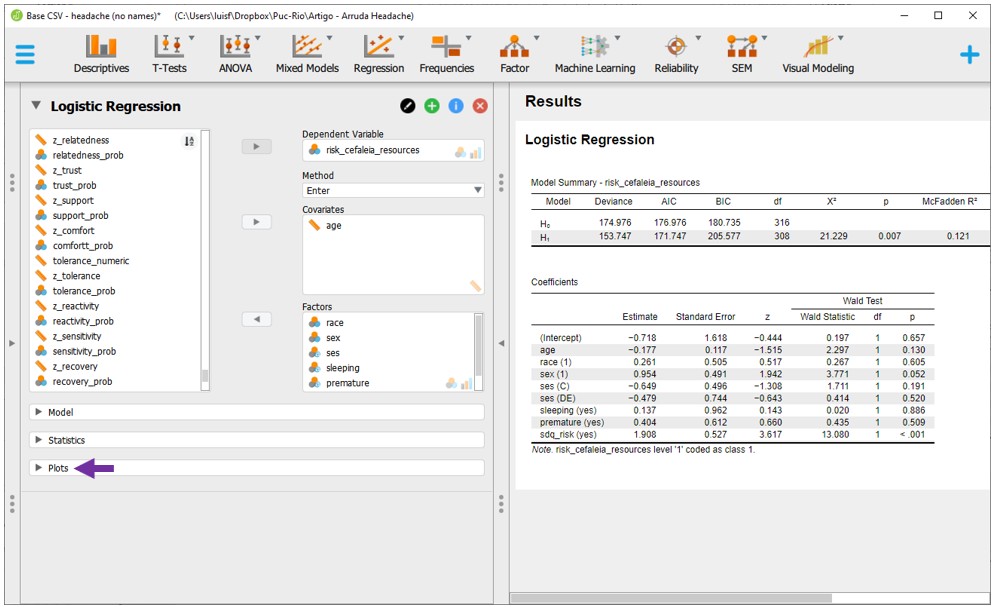
Eventualmente, variáveis contínuas também podem ser significativas. Caso a variável age tivesse sido significativa, sua interpretação poderia ser:

* O aumento de 1 ano de idade impacta a chance de vir a apresentar baixos recursos psicológicos em 0.82, indicando uma característica protetiva.

**Atenção**: A validade das inferências dos resultados depende da adequação ou não dos pressupostos dos testes estatísticos. A avaliação destas condições é parte de um procedimento diagnóstico que deve ser sempre feito.

Uma vez que o modelo já foi realizado, a interpretação dos resultados depende da adequação de seus pressupostos. A violação destes pressupostos distorce, limita ou invalida as interpretações teóricas propostas, uma vez que tanto o aumento do erro do tipo 1 (falso positivo), como do tipo 2 (falso negativo) podem ocorrer (Barker & Shaw, [2015](#ref-Barker2015); Ernst & Albers, [2017](#ref-Ernst2017); Lix et al., [1996](#ref-Lix1996)). Corriqueiramente, testar os pressupostos é uma etapa anterior à própria realização do teste inferencial. Entretanto, pedagogicamente a apresentação deles após a execução do teste parece mais adequada.

Nesta versão do JASP, não há testes formais para verificar a multicolinearidade, mas há para verificar a relação linear entre o logit e preditores contínuos. Para realizar esta análise, é necessário clicar em Plots, na parte esquerda da interface.



Um conjunto de opção irá ser exibida. É necessário marcar Inferential Plots e Display conditional estimates plots. O JASP irá fazer gráficos individuais para cada preditor. No entanto, é importante olhar apenas o relacionamento entre os preditores contínuos (por exemplo, age) e o desfecho. O ideal é que haja um relacionamento linear entre ambos, tal como ocorre neste caso.



Dentre as análises possíveis em regressão logística, o JASP também oferece o cálculo de algumas medidas de performance, tal como acurácia e curva ROC. Apesar delas não serem abordadas neste momento, elas são especialmente importantes quando se deseja comparar modelos de regressão.

Após essas análises, é possível interpretar os resultados.

## 3.4 Escrita dos resultados

De uma forma geral, o principal achado do modelo de regressão logística é que a chance de mulheres e participantes com TDAH apresentarem baixos recursos psicológicos é significativamente maior do que seus pares homens e pessoas sem TDAH. Essas condições possuem uma grande importância em áreas de saúde, uma vez que elas permitem que clínicos não apenas entendam melhor o fenômeno, como também possam mapear as possíveis condições que o mantém. Abaixo uma sugestão de escrita baseada nas recomendações da American Psychological Association (APA).

**Como escrever os resultados**

Um modelo de regressão logística foi conduzido para verificar os efeitos de diferentes condições sociais, ambientas e psicológicas na chance dos participantes apresentarem baixos recursos psicológicos. O modelo foi estatisticamente significativo (X2(308) = 21.229, p = 0.007, = 0.10) e indicou que mulheres - em comparação com os homens - têm uma chance 3.07 maior de apresentarem tal condição (OR = 3.07, p = 0.032). Participantes com TDAH apresentam uma chance 7.53 maior de apresentarem tal desfecho (OR = 7.53, p < 0.001), quando comparados àqueles sem TDAH. Dessa forma, ambas as características são fatores de risco à apresentação de baixos recursos psicológicos.

## 3.5 Regressão logística, OR e Qui-quadrado

Conforme apresentado em capítulos anteriores, o teste Qui-quadrado é um caso particular do modelo de regressão logística em que há uma única variável independente categórica. Além disso, o *Odds Ratio* é uma medida de efeito que costuma ser obtida pela mesma tabela de contingência em que o teste Qui-quadrado é calculado.

Para demonstrar este conceito, a relação entre o sexo do participante (Masculino ou Feminino) e estar em risco de baixo recurso psicológico (sim ou não) será modelada.

Inicialmente, a tabela de contingência descreve o relacionamento entre as variáveis.

|  |  |  |  |
| --- | --- | --- | --- |
| sex | 0 | 1 | Total |
| Male | 95% (150) | 5% (8) | 100% (158) |
| Female | 90% (162) | 10% (19) | 100% (181) |
| Total | 92% (312) | 8% (27) | 100% (339) |

Essa tabela é particularmente útil e permite explorar possíveis associações entre as variáveis utilizando o teste Qui-quadrado de independência. Os resultados foram significativos, , indicando pela associação entre sexo e risco. No entanto, como o Qui-quadrado não trabalha com o conceito de variável independente e dependente, seus resultados não indicam em quais células os valores esperados e observados são significativamente diferentes.

Por esta mesma tabela, é possível calcular a chance (*Odds*) de um participante do sexo masculino estar no grupo de risco:

Bem como a chance (*Odds*) de um participante do sexo feminino estar estar no grupo de risco:

Os resultados obtidos às mulheres são superiores ao dos homens. Isso sugere que elas apresentam uma chance consideravelmente maior de estarem nesta esta condição negativa quando comparadas aos homens. O cálculo do *Odds Ratio*, neste sentido, é fundamental para indicar o quão maior é a chance. Como o nome indica, a conta é feita pela razão entre as chances, que indica OR = 2.2.

Esses dois resultados são importantes, mas seriam interpretados separadamente. Eles concluiriam que existe uma associação significativa (com ) entre o sexo do participante e apresentar baixos recursos psicológicos, bem como uma chance aumentada das mulheres apresentarem baixos recursos psicológicos quando comparados aos homens.

Ao modelar esses dados por uma regressão logística, essas duas análises são feitas simultaneamente Neste sentido, a regressão (1) realiza um teste Qui-quadrado de independência, com a particular vantagem de definir claramente uma variável independente e outra dependente, bem como (2) computa os OR dos preditores e testa sua significância estatística. Conceitualmente:

base\_uso %>%   
 mutate(risk = if\_else(risk\_cefaleia\_resources == "0",0,1)) %>% #nuneric for poisson   
 count(risk, sex) %>%   
 {glm(n ~ risk \* sex, data = .,family = poisson())} %>%   
 anova(., test = 'Rao') %>%   
 pander()

Analysis of Deviance Table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Df | Deviance | Resid. Df | Resid. Dev | Rao | Pr(>Chi) |
| **NULL** | NA | NA | 3 | 286.6 | NA | NA |
| **risk** | 1 | 281.5 | 2 | 5.076 | 239.6 | 4.801e-54 |
| **sex** | 1 | 1.562 | 1 | 3.514 | 1.56 | 0.2116 |
| **risk:sex** | 1 | 3.514 | 0 | -3.997e-15 | 3.398 | 0.06527 |

A tabela a seguir apresenta o ajuste deste modelo de regressão (risk\*sex: 3.398), que foi o mesmo valor encontrado na estatística de teste do Qui-quadrado. O valor de P é também o mesmo ().

Por sua vez, o *Odds Ratio* calculado previamente pela tabela de contingência também é encontrado na tabela de resultados da regressão logística ao se fazer . O exponencial, eventualmente, é também chamado de antilogaritmo.

|  |  |  |
| --- | --- | --- |
| Preditor | Estimate | OR |
| (Intercept) | -2.9 | 0.1 |
| sexFemale | 0.8 | 2.2 |

## 3.6 Resumo

1. A regressão logística permite modelar variáveis categóricas, sejam elas binárias ou ordinais
2. A variável dependente desta regressão é uma função (logit) e não uma probabilidade
3. A interpretação desta regressão costuma ser feita pelo conceito de Razão de chances (*Odds Ratio*)
4. O Qui-quadrado é um caso particular de uma regressão logística com uma única VI categórica

## 3.7 Pesquisas adicionais

1. Cardiovascular Disease in Patients With Schizophrenia in Saskatchewan, Canada (DOI: 10.4088/JCP.v65n0519)  
   Esta é uma pesquisa de larga escala, feita para verificar a prevalência de problemas cardiológicos em pacientes portadores de esquizofrenia. O estudo utilizou pesadamente modelos de regressão logística e *Odds ratio* para verificar as diferentes características que impactam em problemas cardiológicos, bem como estimar o tamanho do efeito que estas características possuem.
2. Predictive model and determinants of under-five child mortality: evidence from the 2014 Ghana demographic and health survey (DOI: 10.1186/s12889-019-6390-4)  
   Este estudo visou identificar os preditores relacionados à sobrevivência/mortalidade infantil em Gana, um país africano. O modelo da regressão logística foi utilizado e entre os preditores relacionados à maior chance de sobrevivência estão o sexo da criança e a escolaridade da mãe. Meninas e filhos(as) de mães não-analfabeta tem chance maior de sobrevivência.

## 3.8 Questões

1. (Retirado de Andy Field - Dicovering statistics) A regressão logística assume entre seus pressupostosa) Relacionamento linear entre os preditores contínuos e a variável resposta.b) Relacionamento linear entre os preditores contínuos e o logit da variável resposta.c) Relacionamento linear entre os preditores contínuos.d) Relacionamento linear entre todas as observações. e) Todas as opções são incorretas.
2. (Retirado de Analista de Tecnologia da Informação Análise de Informações, DATAPREV, COSEAC, 2008). Em uma regressão logística múltipla, se uma variável explicativa Xi é dicotômica, seu coeficiente pode ter uma interpretação especial: a razão de chance (*odds ratio*) estimada da resposta para dois níveis possíveis de Xi. Pode- se definir a razão de chance como: a) antilogaritmo de Xi.b) logaritmo de Xi.c) razão entre X1/X2.d) soma dos quadrados de Xi.e) razão entre Xi/ exp(Xi)
3. (Retirado de FUB, CESPE, 2010) Em um estudo clínico utilizou-se um modelo de regressão logística em que y é a variável resposta, como preditor linear, a expressão a + bx + cz, em que x = 0 para o grupo placebo e x = 1 para o grupo de tratamento; z é uma medida de colesterol (em escala de 0 a 5) antes do início do tratamento. Com base nessas informações, julgue os itens subsequentes). A variável resposta y é binária.a) Certo. b)Errado.

Gabarito: 1-b;2-a;3-a

Barker, L. E., & Shaw, K. M. (2015). Best (but oft-forgotten) practices: Checking assumptions concerning regression residuals. *The American Journal of Clinical Nutrition*, *102*(3), 533–539. <https://doi.org/10.3945/ajcn.115.113498>

Ernst, A. F., & Albers, C. J. (2017). Regression assumptions in clinical psychology research practicea systematic review of common misconceptions. *PeerJ*, *5*, e3323. <https://doi.org/10.7717/peerj.3323>

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance "f" test. *Review of Educational Research*, *66*(4), 579. <https://doi.org/10.2307/1170654>

Portugues, E. G. (2020). Lab notes for statistics for social sciences ii: Multivariate techniques. In *Statistics for Social Sciences II: Multivariate Techniques*. <https://bookdown.org/egarpor/SSS2-UC3M/>