1 a 5 de Julho | Modus

Introdução ao R

Fernando Meireles
Instituto de Estudos Sociais e Políticos (IESP-UERJ)
E-mail: fernando.meireles@iesp.uerj.br
Website: www.fmeireles.com

Apresentação

Este curso oferece uma introdução a uma das ferramentas mais potentes para análise de dados: o ambiente de programação R. Entre outros, o curso cobrirá desde a coleta e manipulação de dados até a produção de estatísticas descritivas e visualizações. As aula serão acompanhadas de atividades práticas que envolverão extrair e limpar dados de diversos meios, sintetizar informações e exportar resultados de forma automatizada para Word, LaTeX, PDF e HTML.

O curso está organizado em cinco aulas. Na primeira, trabalharemos com o básico da linguagem `R` necessário para criar e manipular vetores e bases de dados. Partindo disso, nas aulas 2 e 3 veremos como utilizar pacotes do `tidyverse` para limpar, transformar, sintetizar e visualizar dados. Por fim, além de aprendermos a usar funções para rodar testes e modelos comuns em pesquisas, nas aulas 4 e 5 aplicaremos todo o conteúdo visto ao longo do curso para criarmos um pequeno projeto de análise de dados inteiramente replicável.

Certamente não esgotaremos as possibilidades do mundo da análise de dados, ou *data science*, com o R. Na verdade, seria irrealista prometer isto. Ao fim do curso, entretanto, é esperado que as alunas estejam prontas para utilizar o `R` para realizar pesquisa social quantitativa.

Logística

Nossas aulas serão precedidas por breves exposições por meio de *slides* (que ficarão à disposição no repositório de materiais do curso). Apesar de aprender a programar

não demandar necessariamente leitura, realizaremos muitos exercícios em todas as aulas, e é extremamente importante que eles sejam realmente enfrentados, mesmo que com dificuldades. Não há atalho nesse ponto: escrever e reescrever código inúmeras vezes faz parte do processo de aprender a programar.

Como a parte principal do curso será orientada, auxílio e esclarecimento de dúvidas serão feitos principalmente em aula (outras questões podem ser discutidas antes ou depois dos nossos encontros, ou ainda por e-mail). Na pequena atividade final que teremos, trabalhos em duplas são permitidos – no melhor dos cenários, espero que isso estimule vocês a trabalhar em *papers* e projetos conjuntos.

Local e Horários

Local:	LMCS.
Horário das aulas:	18–21h.
Carga horária total:	15 horas.

Avaliação

Teremos dois tipos de atividades durante o curso. Primeiro, por meio de exercícios que deverão ser realizados durante e após as aulas. Essas atividades serão usadas para compor as notas finais, com peso de 50% da nota final (cada lista de exercícios de um dia de aula equivale, portanto, a 10% da nota final do curso). Não se preocupem em entregar respostas corretas, preocupem-se em entregar as melhores repostas que vocês puderem.

Segundo, trabalharemos em um pequeno projeto, de tema livre, inteiramente replicável. A ideia desse exercício é aplicar todo o conteúdo visto ao longo do curso para criar um relatório ou um pequeno *working paper* reportando os resultados de uma análise de dados em R. A título de exemplo, o projeto deve conter (1) uma brevíssima introdução (só o problema e questão); (2) um ou dois parágrafos sintetizando procedimentos e metodologia; e, (3), gráficos ou tabelas reportando resultados acompanhados de seus respectivos códigos em R. Os arquivos finais, códigos e dados servirão de base para a avaliação final.

GitHub

Neste curso, usaremos o GitHub para partilhar alguns arquivos e solucionar dúvidas. Para quem não o conhece, trata-se de uma plataforma colaborativa e aberta para desenvolvimento de projetos e *software*. Particularmente útil, o GitHub oferece um sistema de *track changes* que nos permite ver, etapa por etapa, como nosso código

mudou – e, se precisarmos, poderemos restaurá-lo a qualquer momento. Entre outras funções dele, também vamos usá-lo para postar dúvidas e fazer discussões com exemplos de código.

Para começar, vá ao *site* do GitHub e registre-se e, depois, tente dar *fork* no repositório deste curso, ele está disponível em: github.com/meirelesff/introducao_ao_r_modus. Caso não consiga, não se preocupe, cobriremos isso no primeiro dia de aula.

Informações Específicas

Pré-Requisitos

Embora este curso não tenha pré-requisitos – veremos apenas conteúdos introdutórios –, conhecimentos prévios sobre probabilidade, estatística e reprodutibilidade em pesquisa são úteis.

Política de Gênero

Em aulas de metodologia, homens frequentemente monopolizam a participação. Para evitar isso, seguiremos três protocolos muito simples neste curso: na ausência de computadores no laboratório para todas as pessoas, mulheres serão priorizadas; para intervir, é necessário estender a mão; quando mulheres falam, colegas não as interrompem.

Atendimento a Necessidades Especiais

Alunas com quaisquer necessidades ou solicitações individuais não devem exitar em procurar auxílio, tanto por e-mail quanto pessoalmente.

Usando o R: Instruções Pré-Curso

Para instalar o R

O curso será realizado em um laboratório com toda a estrutura de apoio necessária. Para instalar o `R` em outros computadores para praticar ou resolver exercícios, entretanto, basta ir ao site do CRAN (*Comprehensive R Archive Network*), que é a rede de fundadores e administradores do *core* da linguagem `R`, e baixar o *setup* indicado para o seu sistema operacional:

- cran.r-project.org

Feito isto, já é possível usar o `R` – mas só via *console*, o que não é tão fácil/útil. É por isso que usaremos uma IDE (i.e. Ambiente de Desenvolvimento Integrado) neste curso: especificamente, usaremos o *RStudio*. Para baixá-lo, basta entrar no seguinte site e escolher a opção mais adequada para o seu sistema operacional:

- www.rstudio.com

Para usar alguns recursos mais avançados no Windows, como o pandoc para criar documentos replicáveis em formatos como PDF ou HTML diretamente no R, também é útil ter instalado outro conjunto de ferramentas: o RTools. Assim como nos casos anteriores, ele pode ser baixado do seguinte *website*:

- cran.r-project.org/bin/windows/Rtools

Para quem tiver dúvidas ou problemas ao instalar o R e o RStudio, o seguinte tutorial pode ajudar:

- material.curso-r.com/instalacao

Materiais de Apoio

Para quem deseja complementar ou mesmo aprofundar o conteúdo do curso, recomendo três materiais, que cobrem conteúdos mais voltados para o uso do R em Ciências Sociais. Primeiro, o livro de Wickham & Grolemund (2016), que é a principal referência quando o assunto é R aplicado para análise de dados (ainda que esse seja, essencialmente, um livro sobre *data science*); especialmente útil, esse texto introduz o operador *pipe* e a lógica do `tidyverse`. Segundo, o já clássico texto de Aquino (2014), que é basicamente um manual completo sobre R aplicado às Ciências Sociais. Por fim, o *Hands-On Programming with R* de Grolemund (2014), mais focado na linguagem de programação que dá vida ao R.

Alguns materiais específicos também valem uma passada de olhos. Para um guia atual e bastante completo sobre visualização de dados, ver o livro de Healy (2018) – que tem versão digital e gratuita disponível na internet. Para pequenos tutoriais sobre diversos tópicos de interesse na pesquisa social, ver esse *website*. Para exemplos diversos de aplicação do R na academia e no mercado, recomendo seguir o R-Bloggers, que é uma espécie de repositório de *blogs* e *websites* sobre R. Finalmente, para um guia, ainda em desenvolvimento, sobre R com foco na Ciência Política, ver Meireles & Silva 2018.

Além de todos esses materiais, indico outros específicos, que cobrem os tópicos deste curso, no plano das aulas, a seguir.

Plano das aulas

Aula 1 *Básico*

1.1 Introdução ao `R`; *1.2* O básico sobre objetos e vetores; *1.3* Dados tabulares; *1.4* Manipulação de `data.frame`.

Materiais recomendados:

- Wickham & Grolemund (2016, Cap. 4);
- A (very) Short Introduction to R;
- Introduction to R, DataCamp (curso *online* gratuito);
- Meireles & Silva (2018, Cap. 1);
- R: Um Guia Prático.

Aula 2 *Manipulação I*

2.1 Pacotes; *2.2* Carregando diversos tipos de dados; *2.3* Manipulando `data.frame` e `tibble` com *tidyverse*.

Materiais recomendados:

- Wickham & Grolemund (2016, Cap. 5);
- Meireles & Silva (2018, Cap. 3 e 4);
- A Grammar of Data Manipulation;
- Introduction to dplyr.

Aula 3 *Manipulação II*

3.1 tidy data; *3.2* Síntese de informações; *3.3* Cruzamentos, *binds* e listas; *3.4* Noções de programação funcional.

Materiais recomendados:

- Wickham & Grolemund (2016, Cap. 5);
- Meireles & Silva (2018, Cap. 4);
- Grolemund (2014, Cap. 2, Sec. 2);
- Wickham (2014, Cap. 5);
- Joining Data with dplyr;
- dplyr Cheat Sheet.

Aula 4 *Modelos e Visualização*

4.1 Introdução ao `R Markdown`; *4.2* Criação de gráficos com `ggplot2`; *4.3* Estatística descritiva; *4.4* Modelos lineares e generalizados; *4.5* Exportação de gráficos e tabelas; *4.6* Criação de documentos.

Materiais recomendados:

- Wickham & Grolemund (2016, Cap. 2, 23, 24, 25 e 27);
- Gelman & Hill (2006, Cap. 3, 4, 5, e 6);
- A ggplot2 Tutorial for Beginners;
- ggplot2 (em português);
- Principles of Econometrics with R;
- Introduction to RMarkdown. .

Aula 5 *Projeto*

5.1 Gerenciando repositórios (RStudio + Git); *5.2* Gerenciamento de versões: uma rápida introdução ao `Docker` e ao `checkpoint`; *5.3* Um *workflow* para projetos inteiramente replicáveis; *5.4* Auxílio na tarefa final.

Materiais recomendados:

- Alvarez, Key & Núñez (2018);
- Lupia & Elman (2014);
- Data Access & Research Transparency;
- An Introduction to Docker for R Users.

Referências

Alvarez, R Michael, Ellen M Key & Lucas Núñez. 2018. "Research replication: Practical considerations." *PS: Political Science & Politics* 51(2):422–426.

Aquino, Jakson Alves de. 2014. "R para cientistas sociais.".

Gelman, Andrew & Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Grolemund, Garrett. 2014. *Hands-On Programming with R: Write Your Own Functions and Simulations*. "O'Reilly Media, Inc.".

Healy, Kieran. 2018. *Data visualization: a practical introduction*. Princeton University Press.

Lupia, Arthur & Colin Elman. 2014. "Openness in political science: Data access and research transparency: Introduction." *PS: Political Science & Politics* 47(1):19–42.

Meireles, Fernando & Denisson Silva. 2018. "Usando R: Um Guia para Cientistas Poíticos.".
URL: *http://electionsbr.com/livro/*

Wickham, Hadley. 2014. *Advanced R*. Chapman and Hall/CRC.

Wickham, Hadley & Garrett Grolemund. 2016. *R for data science: import, tidy, transform, visualize, and model data*. "O'Reilly Media, Inc.".