MAT02035 - Modelos para dados correlacionados

Visão geral de modelos lineares para dados longitudinais

Rodrigo Citton P. dos Reis

citton.padilha@ufrgs.br

Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

Porto Alegre, 2019

Introdução

Introdução

- Neste primeiro momento, o nosso foco será exclusivamente nos **modelos lineares** para **dados longitudinais** com **variáveis resposta contínuas** e **com distribuições aproximadamente simétricas**, **sem caudas excessivamente longas** (**ou assimetria**) **ou outliers**.
- Estes modelos fornecem as bases para modelos mais gerais para dados longitudinais quando a variável de resposta é discreta ou é uma contagem.
- Nesta aula apresentamos algumas notações de vetores e matrizes e apresentamos um modelo de regressão linear geral para dados longitudinais.

Introdução

- Nas próximas duas aulas:
  1. Apresentamos uma ampla visão geral de diferentes abordagens para modelar a resposta média ao longo do tempo e para contabilizar a correlação entre medidas repetidas no mesmo indivíduo.
  2. Consideramos alguns métodos descritivos elementares para explorar dados longitudinais, especialmente tendências na resposta média ao longo do tempo.
  3. Concluímos nossa discussão com uma pesquisa histórica de alguns dos primeiros desenvolvimentos em métodos para analisar dados de medidas longitudinais e repetidas.

Introdução

- Devemos enfatizar desde o início que os métodos estatísticos apresentados nesta primeira parte usam a suposição de que as respostas longitudinais têm uma **distribuição normal multivariada** *aproximada* para derivar estimativas e testes estatísticos, mas não exigem isso.

Normalidade $\rightsquigarrow$ Máxima verossimilhança $\rightsquigarrow$ Estimação intervalar $\rightsquigarrow$ Testes de hipóteses $\rightsquigarrow$ Avaliação da adequabilidade dos modelos.

Notação e suposições distribucionais

Notação

- Assumimos uma amostra de N indivíduos são medidos repetidamente ao longo do tempo.
- Denotamos Y_{ij} a variável resposta do i-ésimo indivíduo na j-ésima ocasião de medição.
- Como mencionado anteriormente, os indivíduos podem não ter o mesmo número de medidas e podem não ser medidos nas mesmas ocasiões.
  - Para tal, utilizamos n_i para representar o número de medidas repetidas e t_{ij} os tempos de medida do i-ésimo indivíduo.
    - Se n é o número de **ocasiões planejadas** do estudo, então $n_i \leq n$.

Notação

- É conveniente **agrupar** n_i medidas repetidas da variável resposta do i-ésimo indivíduo em um vetor $n_i \times 1$

$$Y_i = \begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix}, \ i = 1, \ldots, N.$$

Notação

- Presume-se que os vetores de respostas Y_i, para os N indivíduos, sejam **independentes** um do outro.
- Observe, no entanto, que embora os vetores de respostas obtidas em diferentes indivíduos possam geralmente ser considerados independentes uns dos outros (por exemplo, não se espera que medidas repetidas de um resultado de saúde para um paciente em um estudo clínico prevejam ou influenciem os resultados de saúde para outro paciente no mesmo estudo), as medidas repetidas sobre o mesmo indivíduo não são enfaticamente consideradas observações independentes.

Notação

- Quando o número de medidas repetidas é o mesmo para todos os indivíduos do estudo (e não há dados ausentes), não é necessário incluir o índice i em n_i (já que $n_i = n$ para $i = 1, \ldots, N$).
- Da mesma forma, se as medidas repetidas forem observadas no mesmo conjunto de ocasiões, não é necessário incluir o índice i em t_{ij} (já que $t_{ij} = t_j$ para $i = 1, \ldots, N$).

Notação

- Associado a cada resposta, Y_{ij}, há um vetor $p \times 1$ de covariáveis

$$X_{ij} = \begin{pmatrix} X_{ij1} \\ X_{ij2} \\ \vdots \\ X_{ijp} \end{pmatrix}, \; i = 1, \ldots, N, j = 1, \ldots, n_i.$$

- Observe que X_{ij} é um vetor de covariáveis associadas a Y_{ij}, a variável de resposta para o i-ésimo indivíduo na j-ésima ocasião.
- As p linhas de X_{ij} correspondem a **diferentes covariáveis**.

Notação

- Existe um vetor correspondente de covariáveis associado a cada uma das n_i medidas repetidas no i-ésimo indivíduo.
  - X_{i1} é o vetor $p \times 1$ cujos elementos são os valores das covariáveis associadas à variável de resposta do i-ésimo indivíduo na 1ª ocasião de medição;
  - X_{i2} é o vetor $p \times 1$ cujos elementos são os valores das covariáveis associadas à variável de resposta do i-ésimo indivíduo na 1ª ocasião de medição e assim por diante.

Notação

- O vetor X_{ij} pode incluir dois tipos principais de covariáveis: covariáveis cujos valores não mudam ao longo da duração do estudo e covariáveis cujos valores mudam ao longo do tempo.
- Exemplos do primeiro incluem tratamentos experimentais fixos.
- Exemplos deste último incluem o tempo desde a linha de base, o status atual do tabagismo e as exposições ambientais.
- No primeiro caso, os mesmos valores das covariáveis são replicados nas linhas correspondentes de X_{ij} para $j = 1, \ldots, n_i$.
- Neste último caso, os valores obtidos pelas covariáveis podem variar ao longo do tempo (para pelo menos alguns indivíduos) e os valores nas linhas correspondentes de X_{ij} podem ser diferentes a cada ocasião da medição.

Notação

- Podemos **agrupar** os vetores de covariáveis em matrizes $n_i \times p$ de covariáveis:

$$X_i = \begin{pmatrix} X'_{i1} \\ X'_{i2} \\ \vdots \\ X'_{in_i} \end{pmatrix} = \begin{pmatrix} X_{i11} & X_{i12} & \cdots & X_{i1p} \\ X_{i21} & X_{i22} & \cdots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{in_i1} & X_{in_i2} & \cdots & X_{in_ip} \end{pmatrix}, \ i = 1, \ldots, N.$$

- As linhas de X_i correspondem às covariáveis associadas às respostas nas n_i diferentes ocasiões de medição;
- As colunas de X_i correspondem às p covariáveis distintas.

Notação

- Consideramos um modelo de regressão linear para alterações na resposta média ao longo do tempo e para relacionar as alterações às covariáveis,

$$Y_{ij} = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \ldots + \beta_p X_{ijp} + e_{ij}, j = 1, \ldots, n_i; \tag{1}$$

- $\beta_1, \ldots, \beta_p$ são **coeficientes de regressão desconhecidos** relacionando a média de Y_{ij} às suas correspondentes covariáveis.
- Este modelo descreve como as respostas em cada ocasião são relacionadas com as covariáveis.

$$\begin{array}{ccccc} Y_{i1} & = & \beta_1 X_{i11} + \beta_2 X_{i12} + \ldots + \beta_p X_{i1p} + e_{i1} & = & X'_{i1}\beta + e_{i1}, \\ Y_{i2} & = & \beta_1 X_{i21} + \beta_2 X_{i22} + \ldots + \beta_p X_{i2p} + e_{i2} & = & X'_{i2}\beta + e_{i2}, \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Y_{in_i} & = & \beta_1 X_{in_i1} + \beta_2 X_{in_i2} + \ldots + \beta_p X_{in_ip} + e_{in_i} & = & X'_{in_i}\beta + e_{in_i}, \end{array} \tag{2}$$

Notação

- No modelo (1) os e_{ij} são erros aleatórios, com **média zero**, representando desvios das respostas a partir de suas respectivas médias preditas

$$\mathrm{E}\left(Y_{ij}|X_{ij}\right) = \beta_1 X_{ij1} + \beta_2 X_{ij2} + \ldots + \beta_p X_{ijp}.$$

- Tipicamente, mas não sempre, $X_{ij1} = 1$ para todo i e j, e então β_1 é o termo de intercepto do modelo.
  - Não utilizaremos β_0 nem α.

Notação

- Por fim, usando notação de vetor e matriz, o modelo de regressão dado por (1) ou (2) pode ser expresso de uma forma ainda mais compacta,

$$Y_i = X_i\beta + e_i, \tag{3}$$

em que $e_i = (e_{i1}, e_{i2}, \ldots, e_{in_i})'$ é um vetor $n_i \times 1$ de erros aleatórios.

- O modelo de regressão dado por (3) é simplesmente uma representação abreviada para

$$\begin{pmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{pmatrix} = \begin{pmatrix} X_{i11} & X_{i12} & \cdots & X_{i1p} \\ X_{i21} & X_{i22} & \cdots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{in_i1} & X_{in_i2} & \cdots & X_{in_ip} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_{i1} \\ e_{i2} \\ \vdots \\ e_{in_i} \end{pmatrix}.$$

Ilustração: Tratamento de crianças expostas ao chumbo

- Lembre-se de que no **estudo sobre tratamento de crianças expostas ao chumbo**, há 100 participantes do estudo que têm níveis de chumbo no sangue medidos no mesmo conjunto de quatro ocasiões: linha de base (ou semana 0), semana I, semana 4 e semana 6.
- Como todos os indivíduos tem o mesmo número de medidas repetidas observadas no mesmo conjunto de ocasiões, o índice i pode ser retirado de n_i e t_{ij}.
  - Ou seja, $n_1 = n_2 = \ldots = n_N = n$ e da mesma forma $t_{1j} = t_{2j} = \ldots = t_{Nj} = t_j$ para $j = 1, \ldots, 4$.
- No estudo TLC, o vetor de resposta tem comprimento $4 (n = 4)$ e todos os indivíduos são medidos no mesmo conjunto de ocasiões: $t_1 = 0, t_2 = 1, t_3 = 4$ e $t_4 = 6$.

Ilustração: Tratamento de crianças expostas ao chumbo

- Suponha que seja interessante ajustar um modelo à resposta média que pressupõe que o nível médio de chumbo no sangue mude linearmente ao longo do tempo, mas a uma taxa que pode ser diferente para os dois grupos de tratamento.
- Em particular, podemos querer ajustar um modelo em que os dois grupos de tratamento tenham o mesmo intercepto (ou resposta média na linha de base), mas inclinações diferentes.
    - Isso pode ser representado no seguinte modelo de regressão

$$
\begin{aligned}
Y_{ij} &= \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + e_{ij} \\
&= X_{ij}'\beta + e_{ij},
\end{aligned}
$$

em que $X_{ij1} = 1$ para todo i e j (β_1 é um termo de intercepto).

Ilustração: Tratamento de crianças expostas ao chumbo

- A segunda covariável, $X_{ij2} = t_j$, representa a semana em que o nível de chumbo no sangue foi obtido.
- Por fim, $X_{ij3} = t_j \times$ Grupo, em que $\text{Grupo}_i = 1$ se o i-ésimo indivíduo é designado ao grupo succimer e $\text{Grupo}_i = 0$ se o i-ésimo indivíduo é designado ao grupo placebo.

Ilustração: Tratamento de crianças expostas ao chumbo

- Essa codificação de X_{ij2} e X_{ij3} permite que as inclinações do tempo sejam diferentes para os dois grupos de tratamento.
- As três covariáveis podem ser agrupadas em um vetor 3×1 das covariáveis X_{ij}.
- Assim, para crianças do grupo placebo

$$\mathrm{E}\left(Y_{ij}|X_{ij}\right) = \beta_1 + \beta_2 t_j.$$

- β_1 representa o nível de chumbo no sangue médio na linha de base (semana 0);
- β_2 tem interpretação como uma mudança no nível médio de chumbo no sangue (em $\mu g/dL$) por semana.

Ilustração: Tratamento de crianças expostas ao chumbo

- Similarmente para as crianças no grupo succimer

$$\mathrm{E}\left(Y_{ij}|X_{ij}\right) = \beta_1 + (\beta_2 + \beta_3)t_j.$$

- β_1 representa o nível de chumbo no sangue médio na linha de base (assumido ser o mesmo como no grupo placebo, pois o ensaio aleatorizou indivíduos para dois grupos);
- $\beta_2 + \beta_3$ tem interpretação como uma mudança no nível médio de chumbo no sangue (em $\mu g/dL$) por semana.
- Assim, se os dois grupos de tratamentos diferem em suas taxas de declínio nos níveis de chumbo no sangue, então $\beta_3 \neq 0$.

Ilustração: Tratamento de crianças expostas ao chumbo

- Os parâmetros de regressão têm interpretações úteis que se relacionam diretamente com questões de interesse científico.
- Além disso, hipóteses de interesse podem ser expressas em termos da ausência de certos parâmetros de regressão.
- Por exemplo, a hipótese de que os dois tratamentos são igualmente eficazes na redução dos níveis de chumbo no sangue corresponde a uma hipótese que $\beta_3 = 0$.

Ilustração: Tratamento de crianças expostas ao chumbo

- Os valores das respostas para os indivíduos 79 e 8 são apresentados

$$Y_{79} = \begin{pmatrix} 30.8 \\ 26.9 \\ 25.8 \\ 23.8 \end{pmatrix} \text{ e } Y_8 = \begin{pmatrix} 26.5 \\ 14.8 \\ 19.5 \\ 21.0 \end{pmatrix}.$$

Ilustração: Tratamento de crianças expostas ao chumbo

- Associados aos vetores de repostas, temos as matrizes de covariáveis

$$X_{79} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 4 & 0 \\ 1 & 6 & 0 \end{pmatrix} \text{ e } X_8 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 4 & 4 \\ 1 & 6 & 6 \end{pmatrix}.$$

Ilustração: Tratamento de crianças expostas ao chumbo

- O modelo para a média dos níveis de chumbo no sangue pode ser representado

$$E(Y_i|X_i) = X_i\beta,$$

$$E(Y_i|X_i) = \begin{pmatrix} E(Y_{i1}|X_{i1}) \\ E(Y_{i2}|X_{i2}) \\ E(Y_{i3}|X_{i3}) \\ E(Y_{i4}|X_{i4}) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 4 & 0 \\ 1 & 6 & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_1 + \beta_2 \\ \beta_1 + 4\beta_2 \\ \beta_1 + 6\beta_2 \end{pmatrix}$$

para crianças no grupo placebo, e

$$E(Y_i|X_i) = \begin{pmatrix} E(Y_{i1}|X_{i1}) \\ E(Y_{i2}|X_{i2}) \\ E(Y_{i3}|X_{i3}) \\ E(Y_{i4}|X_{i4}) \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 4 & 4 \\ 1 & 6 & 6 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \beta_1 + (\beta_2 + \beta_3) \\ \beta_1 + 4(\beta_2 + \beta_3) \\ \beta_1 + 6(\beta_2 + \beta_3) \end{pmatrix}$$

para crianças no grupo succimer.

Suposições distribucionais

- Até agora, as únicas suposições feitas diziam respeito a padrões de mudança na resposta média ao longo do tempo e sua relação com covariáveis.
- Especificamente, dado que o vetor de erros aleatórios, e_i, é assumido como tendo média zero, o modelo de regressão dado por (3) implica que

$$\mathrm{E}\,(Y_i|X_i) = \mu_i = X_i\beta, \tag{4}$$

em que $\mu_i = (\mu_{i1}, \ldots, \mu_{in_i})'$ é o vetor $n_i \times 1$ de médias condicionais para o i-ésimo indivíduo, com $\mu_{ij} = \mathrm{E}\,(Y_{ij}|X_i) = \mathrm{E}\,(Y_{ij}|X_{ij})$.

Suposições distribucionais

- Em seguida, consideramos as suposições distribucionais relativas ao vetor de erros aleatórios, e_i.
- O vetor de resposta Y_i em (3) é assumido como sendo composto por dois componentes:
  1. um "componente sistemático", $X_i\beta$
  2. um "componente aleatório", e_i.
- A variabilidade aleatória de Y_i decorre da adição de e_i.
  - Isso implica que suposições feitas sobre a forma da distribuição dos erros aleatórios se traduzem em suposições sobre a forma da **distribuição condicional** de Y_i dado X_i.

Suposições distribucionais

- Em seguida, Y_i, o vetor de respostas contínuas, é assumido como tendo uma distribuição condicional que é **normal multivariada**, com vetor de resposta médio

$$\mathrm{E}\left(Y_i | X_i\right) = \mu_i = X_i \beta$$

e matriz de covariância

$$\Sigma_i = \mathrm{Cov}\left(Y_i | X_i\right).$$

- Lembre-se de que, embora as observações de indivíduos diferentes sejam consideradas independentes umas das outras, as medidas repetidas do mesmo indivíduo não são consideradas independentes.
  - Essa falta de independência é capturada pelos elementos fora da diagonal da matriz de covariância Σ_i.

A distribuição normal multivariada

- A base para grande parte da estatística é a teoria das probabilidades.
- De fato, a base formal para muitos métodos estatísticos é uma distribuição de probabilidade assumida para a variável resposta.
- Em termos gerais, uma distribuição de probabilidade descreve a frequência relativa de ocorrência de valores particulares da variável resposta.
- Em particular, a função de densidade de probabilidade para Y, descrita por $f(y)$, descreve a frequência relativa de ocorrência de valores particulares de Y.

A distribuição normal multivariada

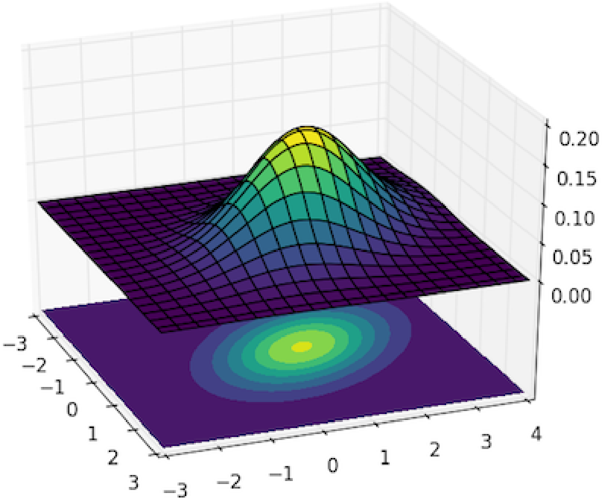
- A distribuição normal (gaussiana) multivariada é uma generalização natural da distribuição normal univariada.
- A função densidade de probabilidade conjunta normal multivariada para Y_i dado X_i pode ser expresso como

$$\begin{aligned} f(y_i) &= f(y_{i1}, y_{i2}, \ldots, y_{in_i}) \\ &= (2\pi)^{-n_i/2} |\Sigma_i|^{-1/2} \exp\left\{-\frac{1}{2}(y_i - \mu_i)' \Sigma_i^{-1} (y_i - \mu_i)\right\}, \end{aligned}$$

em que

- $-\infty < y_{ij} < \infty$ para $j = 1, \ldots, n_i$,
- $\mu_i = \mathrm{E}(Y_i | X_i) = (\mu_{i1}, \ldots, \mu_{in_i})'$,
- $\Sigma_i = \mathrm{Cov}(Y_i | X_i)$,
- e $|\Sigma_i|$ denota o **determinante** de Σ_i.

A distribuição normal multivariada



A distribuição normal multivariada

- A suposição de normalidade multivariada não é crucial para **estimação** e validade das inferências com respeito a β quando os dados são completos (sem ausência de dados).
- Os desvios da normalidade, a menos que sejam muito extremos (por exemplo, dados de resposta altamente assimétricos), não são tão críticos.
- No cenário de dados longitudinais, existem resultados muito semelhantes, o que sugere que são as suposições sobre a dependência entre os erros e as suposições sobre as variações e covariâncias que têm maior impacto na inferência estatística.
  - Desvios da normalidade multivariada, a menos que sejam muito extremas, não são tão críticas.

Breve introdução ao R

O que é o R?

- O R é uma linguagem de programação desenvolvida para:
  - Manipulação de dados;
  - Análise estatística;
  - Visualização de dados.
- O que diferencia o R de outras ferramentas de análise de dados?
  - Desenvolvido por estatísticos;
  - É um software livre;
  - É extensível através de pacotes.

Breve histórico

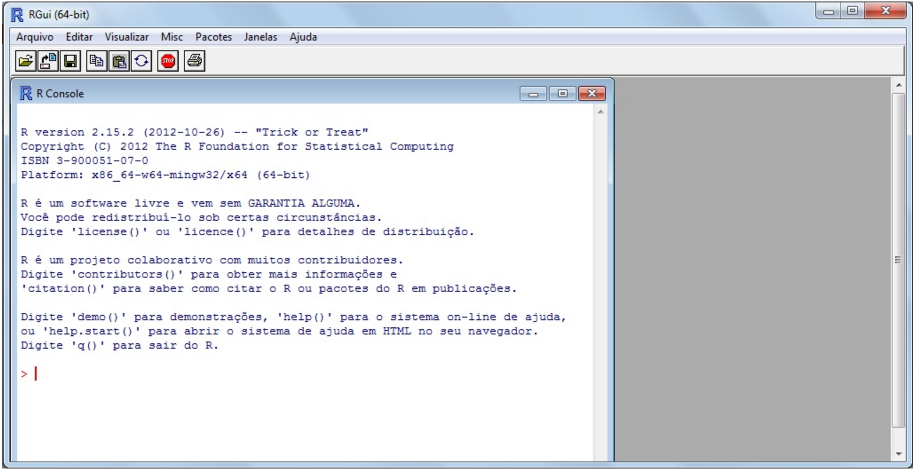
- R é a versão livre, de código aberto, e gratuita do **S**.
  - Nos anos 1980 o **S** foi desenvolvido nos **Laboratórios Bell**, por **John Chambers**, para análise de dados e geração de gráficos.

Breve histórico

- O **R** foi inicialmente escrito no começo dos anos 1990.
  - **Robert Gentleman** e **Ross Ihaka** no Dep. de Estatística da Universidade de Auckland.
  - O nome **R** se dá em parte por reconhecer a influência do **S** e por ser a inicial dos nomes **Robert** e **Ross**.

- Desde 1997 possui um grupo de 20 desenvolvedores.
  - A cada 6 meses uma nova versão é disponibilizada contendo atualizações.

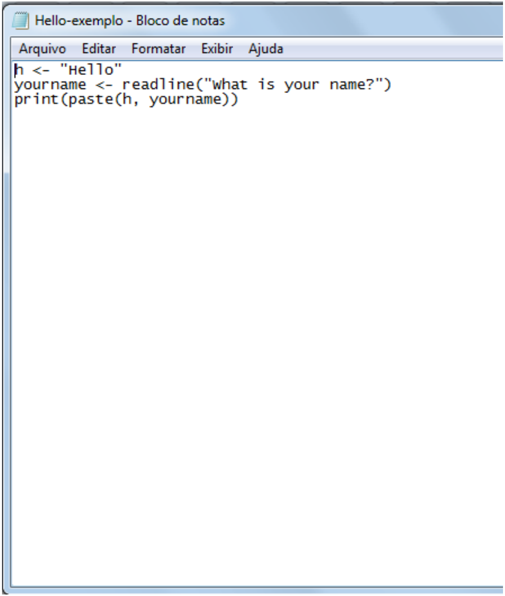
Interface do R



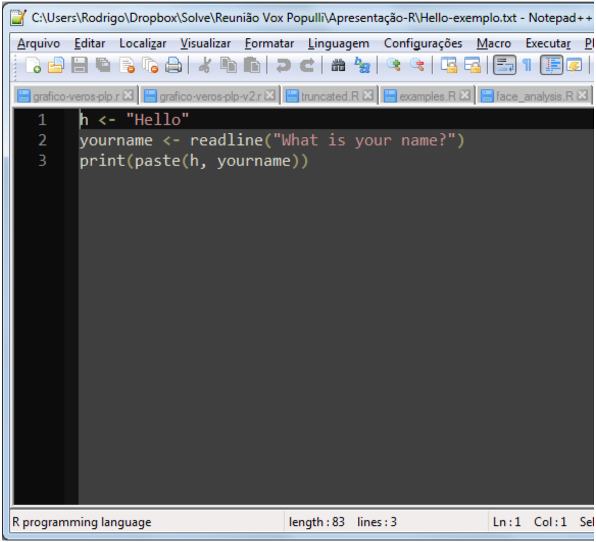
Como trabalhar com o R?

- Por ser uma linguagem de programação, o R realiza suas tarefas através de **funções** e **operadores**.
  - A criação de ***scripts*** (rotinas) é **a melhor prática para se trabalhar** com o R.
    - **OBSERVAÇÃO:** sempre salve seus scripts (em um pen drive, dropbox ou e-mail); você pode querer utilizá-los novamente no futuro.
  - Utilização de editores de texto: **bloco de notas**, **Notepad ++**, **Tinn-R**, etc.
  - Interfaces de R para usuários: **RStudio**.

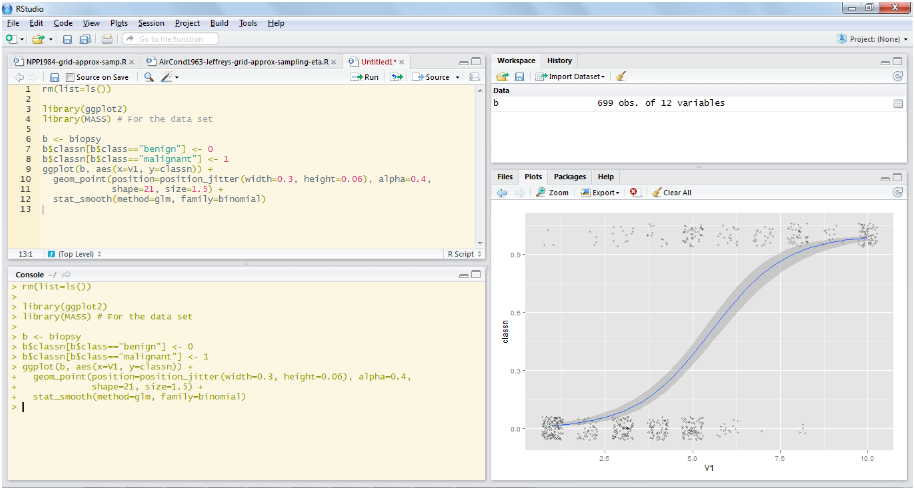
Editores de texto



Editores de texto



Interface do RStudio



Analisando dados

Fases de análise

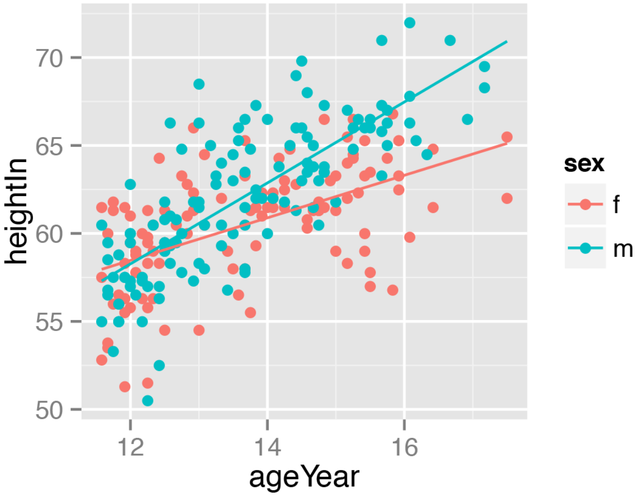
1. **Manipulação inicial** dos dados.
   - Limpeza dos dados.
   - Criação, transformação e recodificação de variáveis.
2. **Análise preliminar**.
   - Conhecimento dos dados, identificação de outliers, investigação preliminar.
3. **Análise definitiva**.
   - Disponibiliza a base para as conclusões.
4. **Apresentação das conclusões** de forma precisa, concisa e lúcida.

Você pode usar o R para

- **Importação e exportação de dados**
- **Manipulação de dados:** Transformação e recodificação de variáveis; Aplicação de filtros
- **Visualização de dados:** Diversos gráficos; Mapas; Gráficos e mapas interativos
- **Análise de dados:** Análise descritiva; Ajuste de modelos; Técnicas multivariadas; Análise de amostras complexas
- **Geração de relatórios:** Relatórios nos formatos pdf, HTML, Word, Power Point

Resumindo: você pode usar o R em todas as etapas de uma análise de dados!

Gráficos do R



Comunicação de resultados através do R: R Markdown

1. Produz **documentos dinâmicos** em R.
2. Documentos R Markdown são completamente **reproduzíveis**.
3. R Markdown suporta dezenas de formatos de saída, incluindo **HTML**, **PDF**, **MS Word**, **Beamer**, ***dashboards***, **aplicações shiny**, **artigos científicos** e muito mais.

Comunicação de resultados através do R: CompareGroups

Características dos grupos do estudo

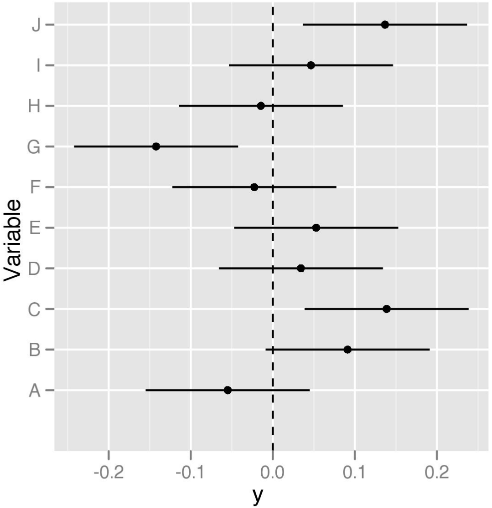
	Total N=6324	Control N=2042	MDN N=2100	MDV N=2182	p-valor
Age	67.0 (6.17)	67.3 (6.28)	66.7 (6.02)	67.0 (6.21)	0.003
Sex: Female	3645 (57.6%)	1230 (60.2%)	1132 (53.9%)	1283 (58.8%)	<0.001
Smoking:					0.444
Never	3892 (61.5%)	1282 (62.8%)	1259 (60.0%)	1351 (61.9%)	
Current	858 (13.6%)	270 (13.2%)	296 (14.1%)	292 (13.4%)	
Former	1574 (24.9%)	490 (24.0%)	545 (26.0%)	539 (24.7%)	
Waist circumference	100 [93.0;107]	101 [94.0;108]	100 [93.0;107]	100 [93.0;107]	0.085
Hormone-replacement therapy	97 (2.80%)	31 (2.64%)	30 (2.81%)	36 (2.95%)	0.898

Comunicação de resultados através do R: `stargazer`

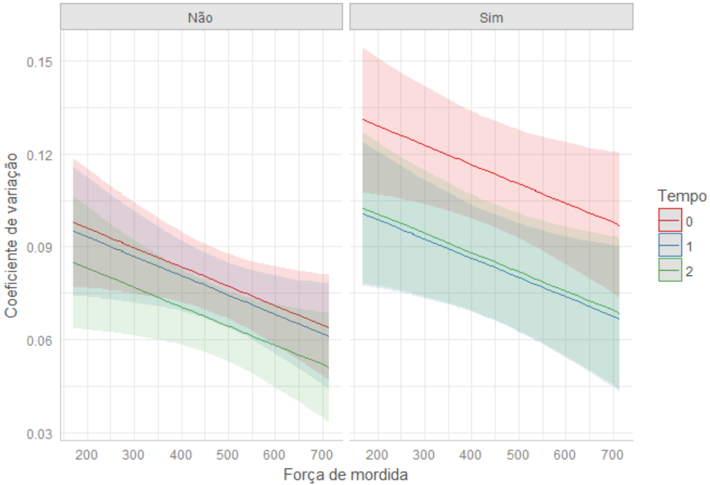
Estimativas dos efeitos fixos dos modelos simples.

	Variável resposta				
	Média de cinza				
	(1)	(2)	(3)	(4)	(5)
time1	4.190**	4.183**	4.190**	4.199**	4.191**
	(0.364, 8.016)	(0.355, 8.011)	(0.363, 8.017)	(0.372, 8.026)	(0.364, 8.019)
time2	9.155***	9.138***	9.161***	9.081***	9.178***
	(4.789, 13.521)	(4.768, 13.508)	(4.791, 13.532)	(4.712, 13.450)	(4.808, 13.549)
forca.de.mordida	0.096***				
	(0.041, 0.150)				
idade		-1.241**			
		(-2.376, -0.105)			
sexoFeminino			-6.492		
			(-27.707, 14.722)		
provisorioSim				16.420*	
				(-0.556, 33.396)	
archMandíbula					9.322
					(-6.396, 25.040)
Constant	51.023***	172.271***	100.214***	90.139***	90.109***
	(24.326, 77.721)	(101.403, 243.139)	(81.940, 118.489)	(79.631, 100.646)	(76.930, 103.287)
Observations	319	319	319	319	319
Note:					*p<0.1;* **p<0.05;** p<0.01

Comunicação de resultados através do R



Comunicação de resultados através do R

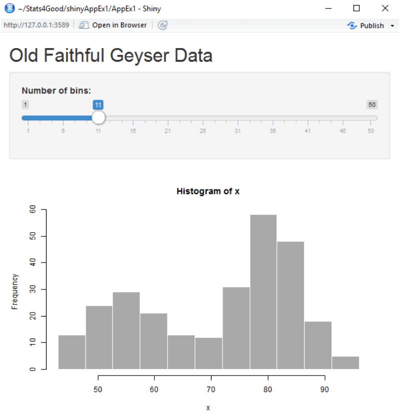


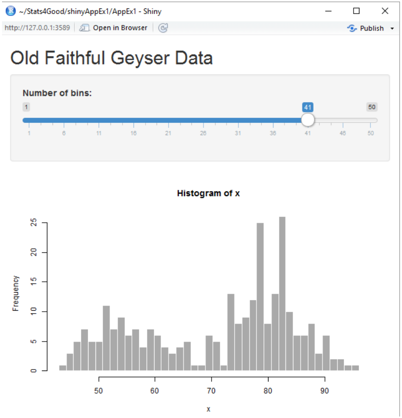
Comunicação de resultados através do R: `Shiny`

- `Shiny` é um pacote do R que torna mais fácil a construção de **aplicações web interativas** (apps) diretamente do R.
  - Permite a criação e compartilhamento de aplicativos.
  - Espera **nenhum conhecimento** de tecnologias web como HTML, CSS ou JavaScript (mas você pode aproveitá-las, caso as conheça)
  - Um aplicativo `Shiny` consiste em duas partes: uma **interface de usuário** (UI) e um **servidor**.

Shiny

```r
# Run the application
shinyApp(ui = ui, server = server)
```





Baixando e instalando o R

Para instalação do R acesse o site https://www.r-project.org/:

1. Em **Download** clique em CRAN.
    - O **CRAN** (*The Comprehensive R Archive Network*) é uma rede de servidores ftp e web em todo o mundo que armazena versões de código e documentação idênticas e atualizadas para o R.
2. Escolha um repositório de sua preferência, por exemplo, Universidade Federal do Paraná (http://cran-r.c3sl.ufpr.br/).
3. Em **Download and Install R** clique no link adequado para o seu sistema operacional (no caso de Windows, clique no link **Download R for Windows**).
4. Clique no link **base** (no caso do sistema operacional ser Windows).
5. Finalmente clique no link para baixar o arquivo executável (a versão mais atual **Download R 3.5.1 for Windows**).

Após baixar o arquivo executável, abra-o e siga as etapas de instalação conforme as configurações padrões.

Baixando e instalando o RStudio

Para instalação do RStudio acesse o site
https://www.rstudio.com/products/rstudio/download/.

- Em **Installers for Supported Platforms** baixe a versão mais recente do instalador do RStudio de acordo com o seu sistema operacional (no caso de Windows clique no link **RStudio 1.1.456 - Windows Vista/7/8/10**).

Pacotes

- Assim como a maioria dos softwares estatísticos, o R possui os seus "módulos", mais conhecidos como **pacotes** do R.
- **Pacote:** é uma coleção de funções do R; os pacotes também são gratuitos e disponibilizados no **CRAN**.
- Um pacote inclui: **funções** do R, **conjuntos de dados** (utilizados em exemplos das funções), arquivo com **ajuda (*help*)**, e uma **descrição** do pacote.
- Atualmente, o repositório oficial do R possui mais de 12.000 pacotes disponíveis.
- As funcionalidades do R, podem ser ampliadas carregando estes pacotes, tornando-o um software muito poderoso, capaz de realizar inúmeras tarefas.

Pacotes

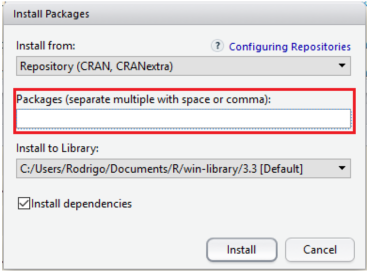
- Alguns exemplos destas tarefas e alguns destes pacotes são listados abaixo:
  - **Importação e exportação de dados**
    - `foreign`, `readr`, `haven`
  - **Manipulação de dados**
    - Transformação e recodificação de variáveis: `reshape2`, `stringr`
  - **Visualização de dados**
    - Diversos gráficos: `graphics`, `ggplot2`, `ggthemes`
    - Mapas: `ggmap`
    - Gráficos e mapas interativos: `plotly`
  - **Análise de dados**
    - Análise descritiva: `compareGroups`
    - Ajuste de modelos: `stats`, `survival`
    - Análise de amostras complexas: `survey`
  - **Geração de relatórios**
    - Relatórios nos formatos pdf, HTML, Word, Power Point: `knitr`, `rmarkdow`, `officer`

Instalando pacotes

- Para **instalação de um pacote**, basta um simples comando.

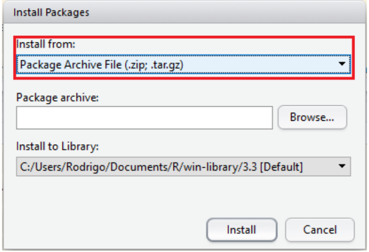
```
install.packages("survey")
```

- Além da opção de comando, também podemos instalar pacotes utilizando o menu **Tools** do `RStudio`, opção **Install packages ...** e preenchendo com o(s) nome(s) do(s) pacote(s):



Instalando pacotes

- Outra opção é instalar o pacote a partir de seu arquivos fonte (**.zip** ou **.tar.gz**):
  - Para isso, obtenha o arquivo fonte do pacote (geralmente através do **CRAN**) e no menu **Tools** do `RStudio`, opção **Install packages ...** em **Install from** escolha a seguinte opção:



Instalando pacotes

Após a instalação do pacote, temos que **carregar o pacote** para nossa área de trabalho para podermos usufruir de suas funções.

```
library("survey")
require("survey")
```

Obtendo ajuda no R

- Para conhecer quais as funções disponíveis no pacote, faça:

```
help(package = "survey")
```

- Para pedir ajuda de uma determinada função:

```
?glm
help("glm")
```

- Obtendo ajuda na internet:

```
help.search("t.test")
```

Obtendo ajuda no R

- Procurando por alguma função, mas esqueci o nome:

```
apropos("lm")
```

- Para todas as outras dúvidas existe o **Google!**
- Ver também http://www.r-bloggers.com/ e https://rstudio.cloud/
- Para algumas demonstrações da capacidade gráfica do R:

```
demo(graphics)
demo(persp)
demo(Hershey)
demo(plotmath)
```

Exercícios

Exercícios

- Com o auxílio do computador, faça os exercícios do Capítulo 2 do livro "**Applied Longitudinal Analysis**" (páginas 44 e 45).
- Enviar soluções pelo Moodle através do fórum.

Avisos

- **Para casa:** ler o Capítulo 3 do livro "**Applied Longitudinal Analysis**". Caso ainda não tenha lido, leia também os Caps. 1 e 2.
  - Ver https://datathon-ufrgs.github.io/Pintando_e_Bordando_no_R/
- **Próxima aula:** Métodos de análise descritiva para dados longitudinais.

Bons estudos!