Modelando dados amostrais complexos

Leo Bastos (PROCC/Fiocruz)

MGEST, 5 e 6 de outubro de 2017

Conteúdo

1. Inferência estatística
   - Exemplo: Tabagismo em BH
   - População ou Processo?
   - Pesos amostrais
   - Vigitel 2016
2. Software, dados, e scripts
   - Por quê R?
   - O pacote *survey*
   - Descritivas para tabagismo em BH
3. Modelos estatísticos para amostras complexas usando o *survey*
   - Modelos lineares generalizados (MLG)
   - MLG incorporando o desenho amostral
   - Aplicando aos dados de tabagismo em BH
   - Bayes (?)

Referência principal

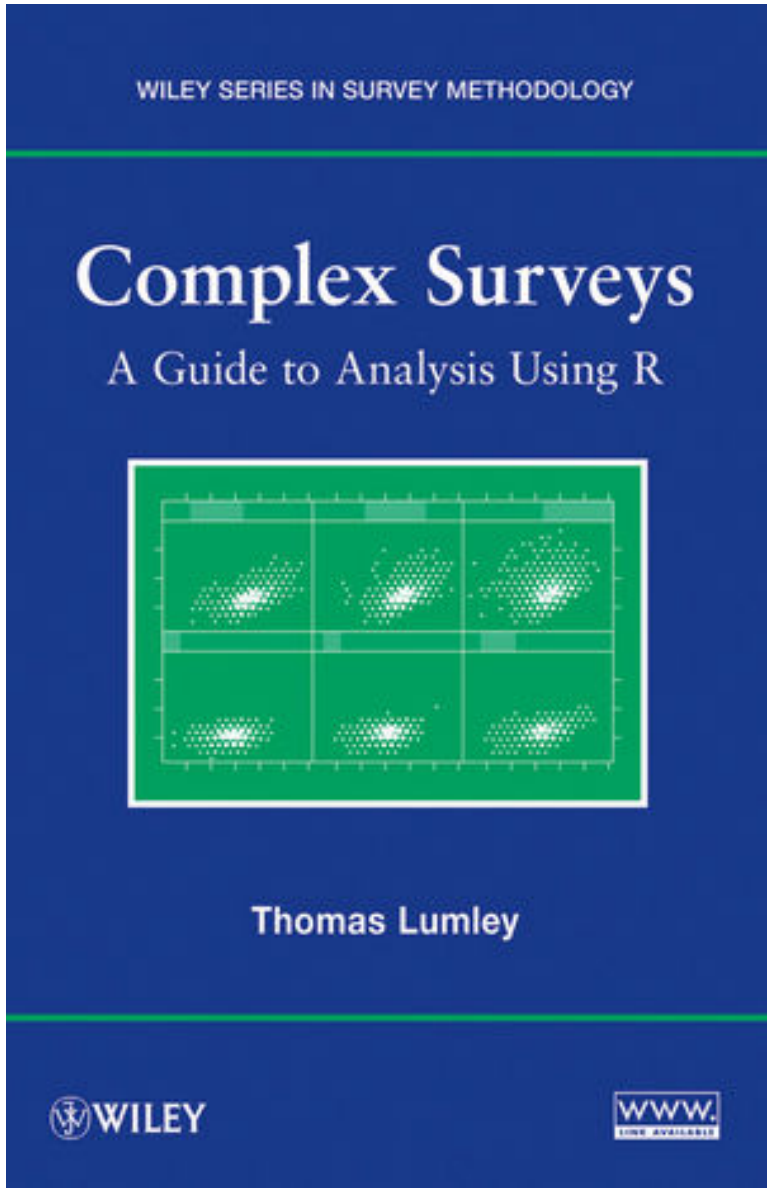


Figure 1: Livro texto

Inferência estatística

- Exemplo: Tabagismo em BH
- População ou Processo?
- Pesos amostrais
- Vigitel 2016

Exemplo

- Suponha que nosso objetivo seja estimar a prevalência de fumantes em Belo Horizonte e avaliar fatores associados ao tabagismo.

Figure 2: Região Metropolitana de Belo Horizonte à noite a partir da Estação Espacial Internacional. (Wikipedia)

Tabagismo

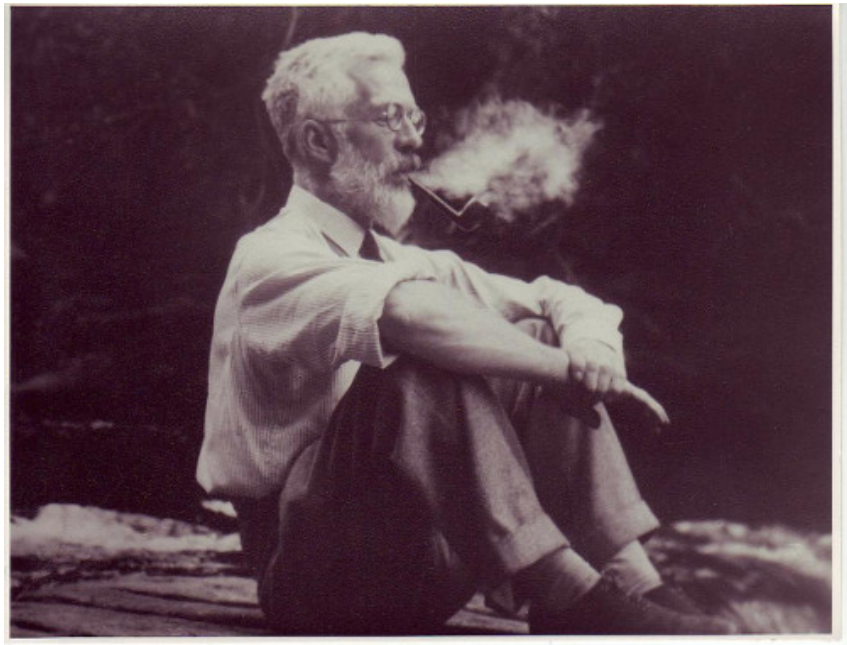


Figure 3: Estatístico famoso em 1946.

Tabagismo no Mundo

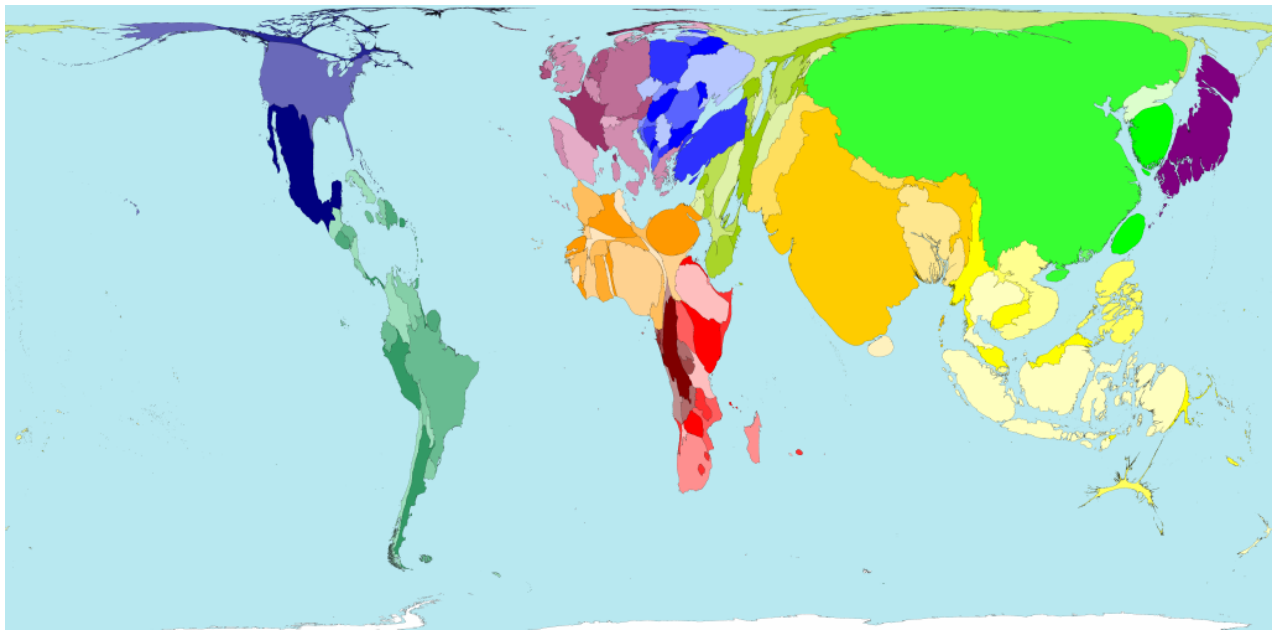


Figure 4: Cartograma da prevalência de tabagismo em homens no mundo.

Tabagismo no Mundo

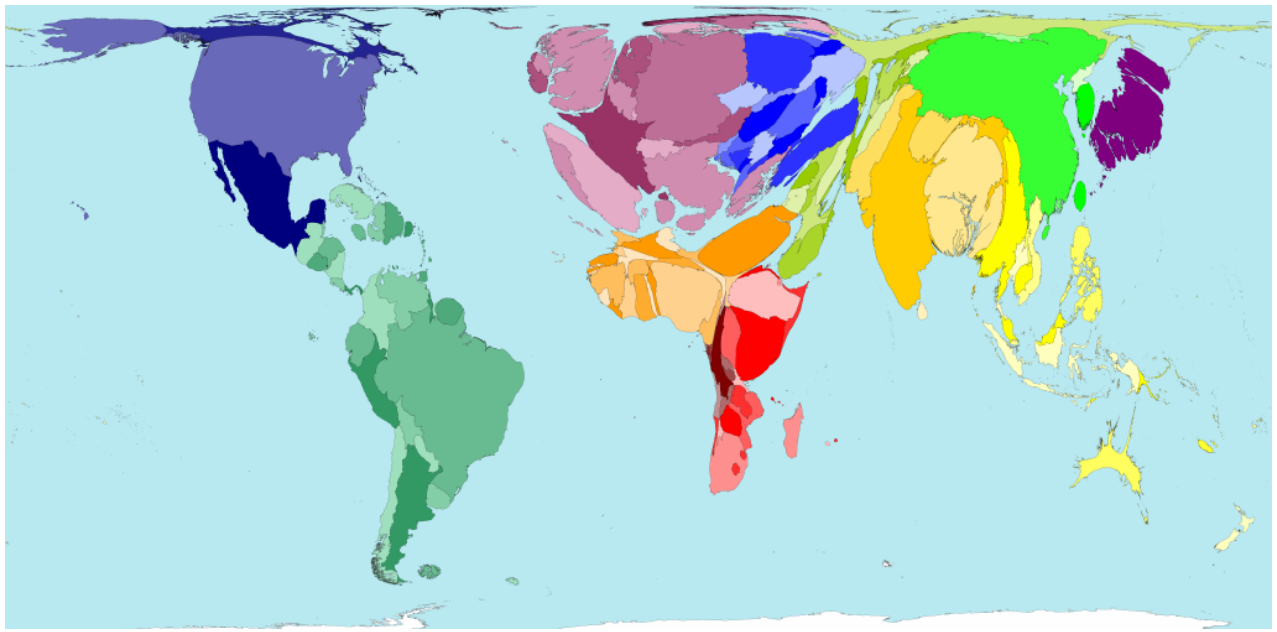


Figure 5: Cartograma da prevalência de tabagismo em mulheres no mundo.

Como estimar o tabagismo em BH?

- Vamos coletar uma amostra **representativa** da população de belorizontinos
- Avaliar o status de tabagismo, i.e.

$$Y_i = \begin{cases} 1, & \text{se a pessoa i é fumante,} \\ 0, & \text{caso contrário.} \end{cases} \qquad i = 1, 2, \ldots, n.$$

- n é o tamanho da amostra
- Como coletar a amostra? O que garante uma amostra representativa?
- Se n fosse a população de BH, saberíamos exatamente a prevalência de fumantes em BH.

População ou processo?

- Na **amostragem**, a análise de dados é baseada no **desenho**, *design-based*.
- Assume-se que os dados da população são **desconhecidos**, porém **fixos**.
- Sob essa ótica, não há distribuição de probabilidade para uma variável.
- A aleatoriedade se dá através do desenho, **a amostra é aleatória**.
- No nosso exemplo, o status de fumar ou não é fixo, a única fonte de aleatoriedade por aqui é o sorteio da amostra.

População ou processo?

- Em contraste, a inferência estatística usual, é baseada em **processos**.
- Assume-se que os dados são realizações de um processo aleatório.
- Existe uma distribuição de probabilidades a esse processo.
- Abordagem baseada em modelos estatísticos, *model-based*
- No nosso exemplo, uma pessoa de BH aleatoriamente selecionada fuma com probabilidade θ.
- Essa probabilidade é a prevalência que estamos interessados.

Abordagem baseada em modelos com o desenho

- Rubin (1976) define o conceito de ignorabilidade.
- Dizemos que um desenho é ignorável, se o processo estocástico que estamos estudando é independente do desenho.
- Exemplo de desenho amostral ignorável: **amostra aleatória simples com reposição**
- Será possível sortear uma amostra aleatória simples para a população de BH?
- A ignorabilidade é uma suposição razoável em uma pesquisa amostral?
- Como podemos sortear uma amostra **representativa** de belorizontinos?

Uma amostra representativa

- Propriedades
  1. Cada indivíduo da população DEVE ter uma probabilidade $\pi_i > 0$
  2. A probabilidade π_i deve ser conhecida para todo indivíduo que cair na amostra.
  3. Todo par (i, j) de indivíduos da população DEVE ter probabilidade $\pi_{ij} > 0$
  4. A probabilidade π_{ij} deve ser conhecida para todo par de indivíduos da amostra.
- 1 e 2 são necessários, enquanto 3 e 4 dependem de π_{ij} que pode ser calculado segundo o desenho.

Peso amostral

- A ideia fundamental por trás da inferência baseada no desenho é que uma pessoa amostrada com probabilidade π_i representa $1/\pi_i$ pessoas daquela população.
- $\omega_i = 1/\pi_i$ é chamado de **peso amostral**
- Suponha que estamos interessado em um total de uma variável X (e.g. renda) da população
- A contribuição de cada indivíduo amostrada para o total é dada por

$$X_i \omega_i = \frac{X_i}{\pi_i}$$

Estimador Horvitz-Thompson

- O estimador Horvitz-Thompson para o total (Horvitz and Thompson, 1952, JASA)

$$\hat{T}_{HT} = \sum_{i=1}^{n} \frac{X_i}{\pi_i}$$

- O estimador da variância de $\hat{T}_{HT}$

$$\widehat{\mathbb{V}ar}[\hat{T}] = \sum_{i,j} \left(\frac{X_i X_j}{\pi_{ij}} - \frac{X_i}{\pi_i} \frac{X_j}{\pi_j} \right)$$

Pesos

O GLOBO

oglobo.com.br

SEGUNDO CADERNO

David Cronenberg
UM ESTREANTE DE PESO NA LITERATURA

ela

Downton Abbey
DAMAS DE HOJE

PANTALONA CURTA: A CALÇA DO VERÃO

PROSA
Revista
UMA REFERÊNCIA CINQUENTENÁRIA

NUNCA ANTES

Erro em 'numerinho' obriga IBGE a corrigir toda a Pnad

Com mudança, desigualdade cai em vez de subir e ganho de renda é menor

Instituto superestimou peso das estatísticas de sete estados. Problema foi identificado por Ipea e consultorias

ELEIÇÕES 2014

Figure 6: Capa do O Globo publicada em 20 de setembro de 2014.

Peso amostral

- Se pegarmos uma amostra aleatória de tamanho 2500 de BH (Pop. 2,5 milhões hab.), então cada pessoa teria chance de 1 em mil de ser selecionada.
- Cada pessoa amostrada em BH representaria 1000 belorizontinos.
- Se por acaso, 500 pessoas sorteadas se declararem fumantes, significa então que BH tem 500.000 fumantes (500 * 1000).

Peso amostral

- Se pegarmos uma amostra de tamanho 2500 de Diamantina (Pop. 50 mil hab.), então cada pessoa teria probabilidade de 20%, ou 200 em mil, de ser selecionada.
- Cada pessoa amostrada em Diamantina representaria 5 diamantinenses.
- Se por acaso, 500 pessoas amostradas tiverem hipertensão, significa que Diamantina tem 10.000 hipertensos (200 * 500).

Poś-estratificação (ou calibração)

- Outra característica comum em pesquisas amostrais (*surveys*) é a pós-estratificação
- Depende de informações externas como censos e projeções populacionais
- Os pesos podem ser recalculados de modo que estimativas da amostra coincidam com estimativas populacionais conhecidas:
  - Total populacional
  - Distribuições sócio-demográficas (sexo, faixa etária, escolaridade, etc.)
  - Outros fatores que possam estar associados a probabilidade de seleção

Outros pesos

- **Sampling weights** um peso de 1000 significa que a observação representa 1000 indivíduos daquela população.
- **precision weights** um peso de 1000 significa que a observação tem uma variância 1000 vezes menor que uma observação de peso 1.
- **frequency weights** um peso de 1000 significa que a amostra contêm 1000 observações identicas e espaço (computacional) está sendo salvo representando-os com apenas uma observação.

VIGITEL

PS: Microdado disponível para download no site do Datasus.

VIGITEL

No Vigitel tem-se a pergunta:

Q60. Atualmente, o(a) sr(a) fuma?

1. ( ) sim, diariamente
2. ( ) sim, mas não diariamente
3. ( ) não

Vigitel

- Pesquisa anual, realizada desde de 2006 nas 26 capitais estaduais + DF.
- Os pesos são atribuídos primeiramente considerando dois fatores:
  - número de linhas telefônicas no domicílio entrevistado
  - número de adultos no domicílio entrevistado
- O peso final é atribuído usando pós-estratificação equiparando a distribuição sócio-demográfica da amostra Vigitel (população com telefone fixo) com a da população geral da capital em questão.
- São usadas projeções oficiais para as características:
  - sexo
  - faixa etária
  - nível de instrução

Software, dados, e scripts

- Por quê R?
- O pacote *survey*
- Descritivas para tabagismo em BH

Softwares

- Existem alguns softwares que fazem inferência baseada em desenho.
  - SUDAAN [http://www.rti.org/sudaan/]
  - SAS
  - SPSS
  - STATA
  - Epi Info
  - R

Por que R?

- Por que usar o R?
  - Software livre
  - Alto número de pacotes desenvolvidos e em desenvolvimento
  - Elevado número de usuários e desenvolvedores

Pré requisitos

- Pré-requisitos de sistema
  - Praticamente não há;
  - Eu recomendo: GNU/Linux 64 bits
  - Quanto mais memória RAM melhor
  - Permite:
    - computação paralela
    - CUDA
    - uso de softwares externos
    - etc

Baixando o R e instalando pacotes

- Site do R [http://www.r-project.org]
  - R version 3.4.1 (2017-06-30)
- RStudio (editor de texto integrado com R) [http://www.rstudio.com]
- Instalando pacotes no R

```
install.packages("nome_do_pacote")
```

O pacote *survey*

- A grande maioria dos pacotes desenvolvidos no R são para métodos de inferência baseada em modelos.
- Existem alguns pacotes (poucos) desenvolvidos para inferência baseada no desenho.
- O pacote survey, é um pacote dedicado a análise de dados amostrais complexos
- Desenvolvido por Prof. Thomas Lumley (The University of Auckland)
- [`http://r-survey.r-forge.r-project.org/survey/`]

```
# Instalando o pacote
  install.packages("survey")
```

Comandos básicos

```
# Chamando a biblioteca survey
require(survey)

# Lendo os microdados do Vigitel 2016 de BH
BH <- read.csv("Data/VIGITEL_MicroDado_2016_BH.csv")

# Tamanho da amostra
nrow(BH)
```

```
## [1] 2004
```

Comandos básicos

```
# Olhar o dicionário de variáveis!
# Variável sexo - q7
BH$sexo <- factor(x = BH$q7, levels = 1:2,
                  labels = c("Masculino", "Feminino"))

table(BH$sexo)
```

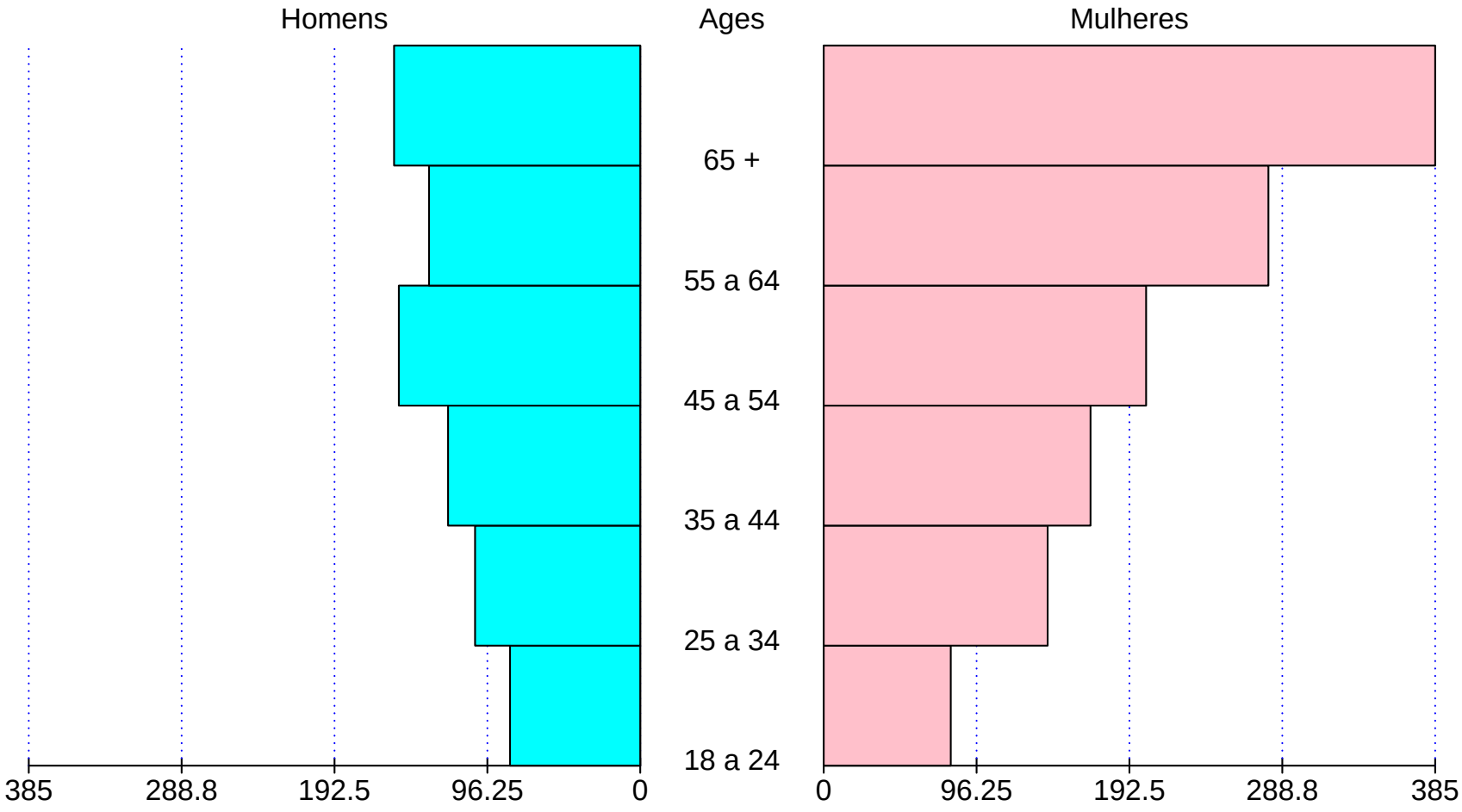
```
##
## Masculino  Feminino
##       747      1257
```

Estimativas de tabagismo em BH, 2016 (por sexo)

```r
# Tabagismo por sexo (ignorando o desenho)
by(data = BH$fumante, INDICES = BH$sexo, FUN = function(x)
      c(mean = mean(x),
        sd = sqrt( mean(x) * (1-mean(x)) / length(x) )
       )  )
```

```
## BH$sexo: Masculino
##       mean         sd
## 0.13654618 0.01256316
## ------------------------------------------------------------
## BH$sexo: Feminino
##       mean         sd
## 0.093078759 0.008194877
```

Faixa etaria amostra Vigitel



Análise de amostras complexas usando o *survey*

- O primeiro passo é descrever para o R qual o desenho da amostra, para isso usa-se a função *svydesign*

```
# Definindo o desenho
BH.svy <- svydesign( id=~1, strata =NULL, fpc=NULL,
                     weights = ~pesorake, data=BH)
# id -- variavel que define os clusters
#      ~1 significa que que não tem clusters
# strata -- variável que define os estratos
# fpc -- correção de população finita, aponta para a
#       variável do banco com o tamanho da população
# weights -- pesos amostrais
# data -- data frame com os dados gerados
```

Descritivas para o Vigitel, BH, 2016

```
# Estimando o total de fumantes de BH em 2016
svytotal(~fumante, BH.svy)
```

```
##             total    SE
## fumante 202808 15693
```

```
# Estimando prevalência de tabagismo na capital
svymean(~fumante, BH.svy)
```

```
##               mean     SE
## fumante 0.10868 0.0083
```

Estimativas de tabagismo em BH, 2016

```
# Tabagismo por sexo
svyby(formula = ~fumante, by = ~sexo, design = BH.svy,
      FUN = svymean)
```

```
##                   sexo    fumante          se
## Masculino Masculino 0.13460369 0.014420938
## Feminino   Feminino 0.08691698 0.009199766
```

```
# Tabagismo por escolaridade
svyby(formula = ~fumante, by = ~fesc, design = BH.svy,
      FUN = svymean)
```

```
##                                fesc    fumante          se
## 0 a 8 anos           0 a 8 anos 0.14518127 0.01635608
## 9 a 11 anos         9 a 11 anos 0.10235845 0.01337693
## 12 anos e mais 12 anos e mais 0.08119574 0.01325516
```

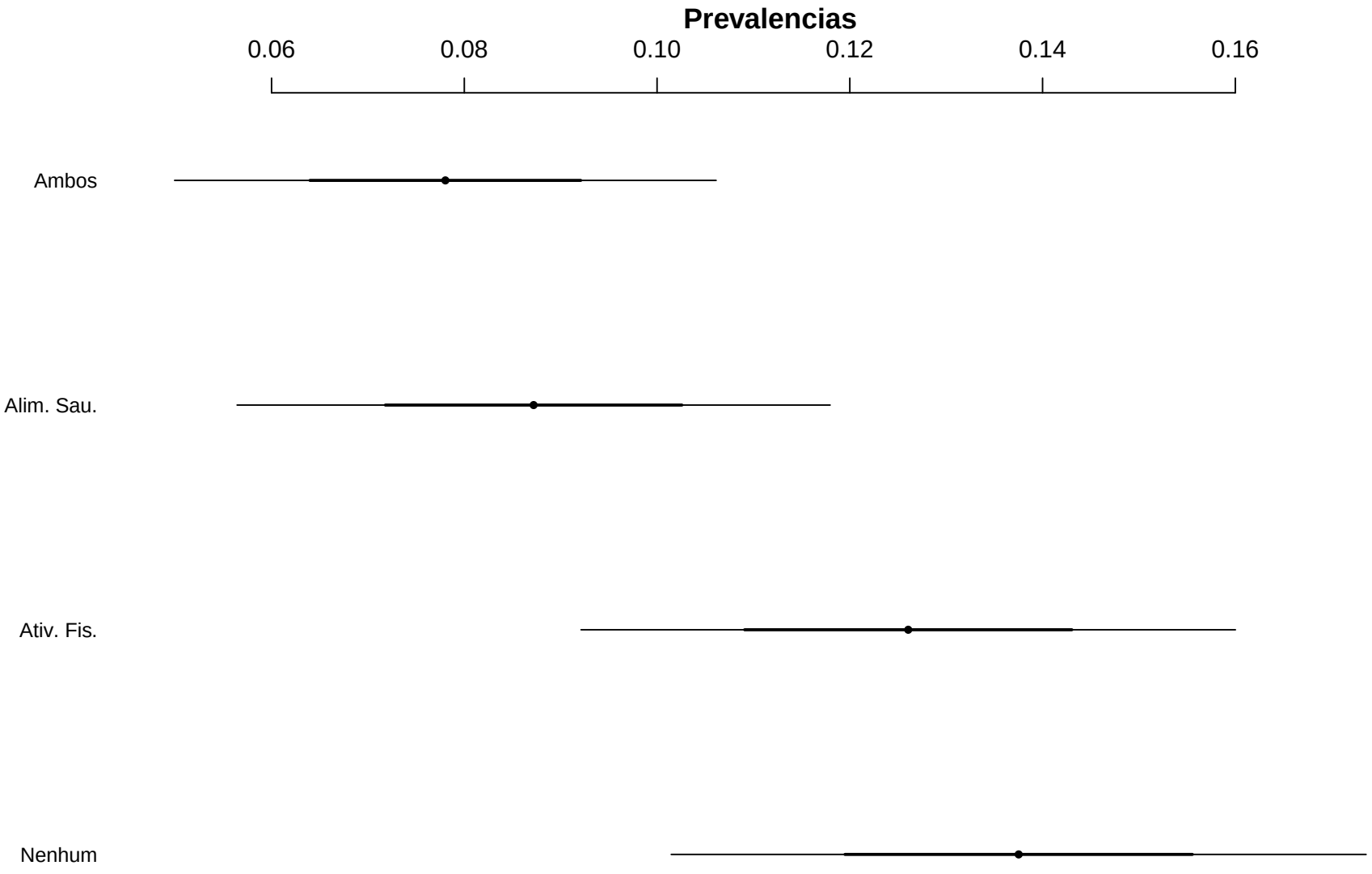
Estimativas de tabagismo em BH, 2016 (por faixa etaria)

```
##                fumante          se
## 18 a 24 0.05178085 0.01880055
## 25 a 34 0.07167517 0.01737145
## 35 a 44 0.11821970 0.02157904
## 45 a 54 0.16477135 0.02276484
## 55 a 64 0.18072785 0.02137502
## 65 +    0.07166733 0.01313629
```

Comportamento saudável

- Uma proxy para comportamento saudável é a composição de duas variáveis do Vigitel
    1. Consumo regular (5x ou mais por semana) de Frutas, Legumes e Hortaliças
    2. Atividade física $>$ 150 min por semanas

Prevalência de tabagismo segundo comportamento de saúde



Modelos estatísticos para amostras complexas usando o *survey*

- Modelos lineares generalizados (MLG)
- MLG incorporando o desenho amostral
- Aplicando aos dados de tabagismo em BH
- Bayes (?)

Prevalência via modelo de regressão logística

- Podemos estimar a prevalência via regressão logística
- Seja

$$Y_i \sim Bernoulli(\theta_i)$$

onde

$$g(\theta_i) = \alpha + \mathbf{x_i}^T \beta$$

- A prevalência para o grupo $\mathbf{x}^*$ é dada por

$$\theta^* = g^{-1}(\alpha + \mathbf{x}^{*T} \beta)$$

Modeos lineares generalizados

- Os modelos lineares generalizados tem 3 componentes:
  1. Componente aleatorio na família exponencial

$$Y \sim FE(\theta)$$

  2. Componente determinístico

$$\eta = \mathbf{x}^T \beta$$

  3. Funcao de ligação

$$g(\mathbb{E}[Y_i]) = \eta_i, \qquad i = 1, 2, \ldots, n.$$

- Os coeficientes β sao estimados maximizando a função de verossimilhança.

$$L(\beta) = \prod_{i=1}^{n} p(y_i \mid \beta, \mathbf{x}_i)$$

Modeos lineares generalizados com peso amostral

- Continuamos com os mesmos 3 componentes:
  1. Componente aleatorio na família exponencial

$$Y \sim FE(\theta)$$

  2. Componente determinístico

$$\eta = \mathbf{x}^T \beta$$

  3. Funcao de ligação

$$g(\mathbb{E}[Y_i]) = \eta_i, \qquad i = 1, 2, \ldots, n.$$

- Os coeficientes β sao estimados maximizando a funcao de **pseudo-verossimilhanca** (Lumley and Scott, 2017)

$$L(\beta) = \prod_{i=1}^{n} p(y_i \mid \beta, \mathbf{x}_i)^{w_i}$$

Tabagismo em BH, 2016

Modelo para estimar o efeito do comportamento saudável no tabagismo

```
# Modelo
modelo <- fumante ~ compsaude

# Ajuste
output <- svyglm(formula = modelo,
                 family = binomial,
                 design = BH.svy)
```

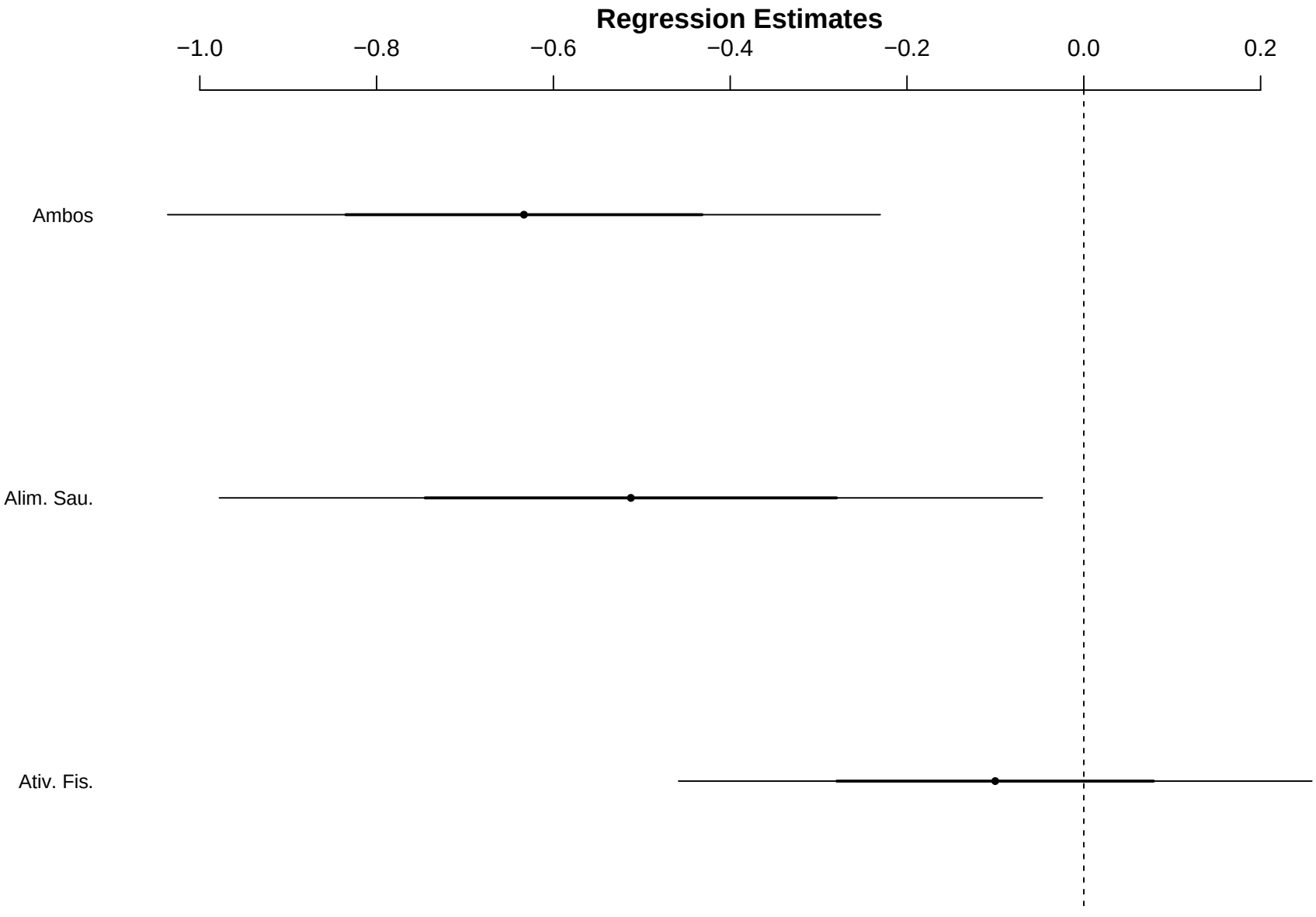
```
## Warning: non-integer #successes in a binomial glm!
```

Tabagismo em BH, 2016

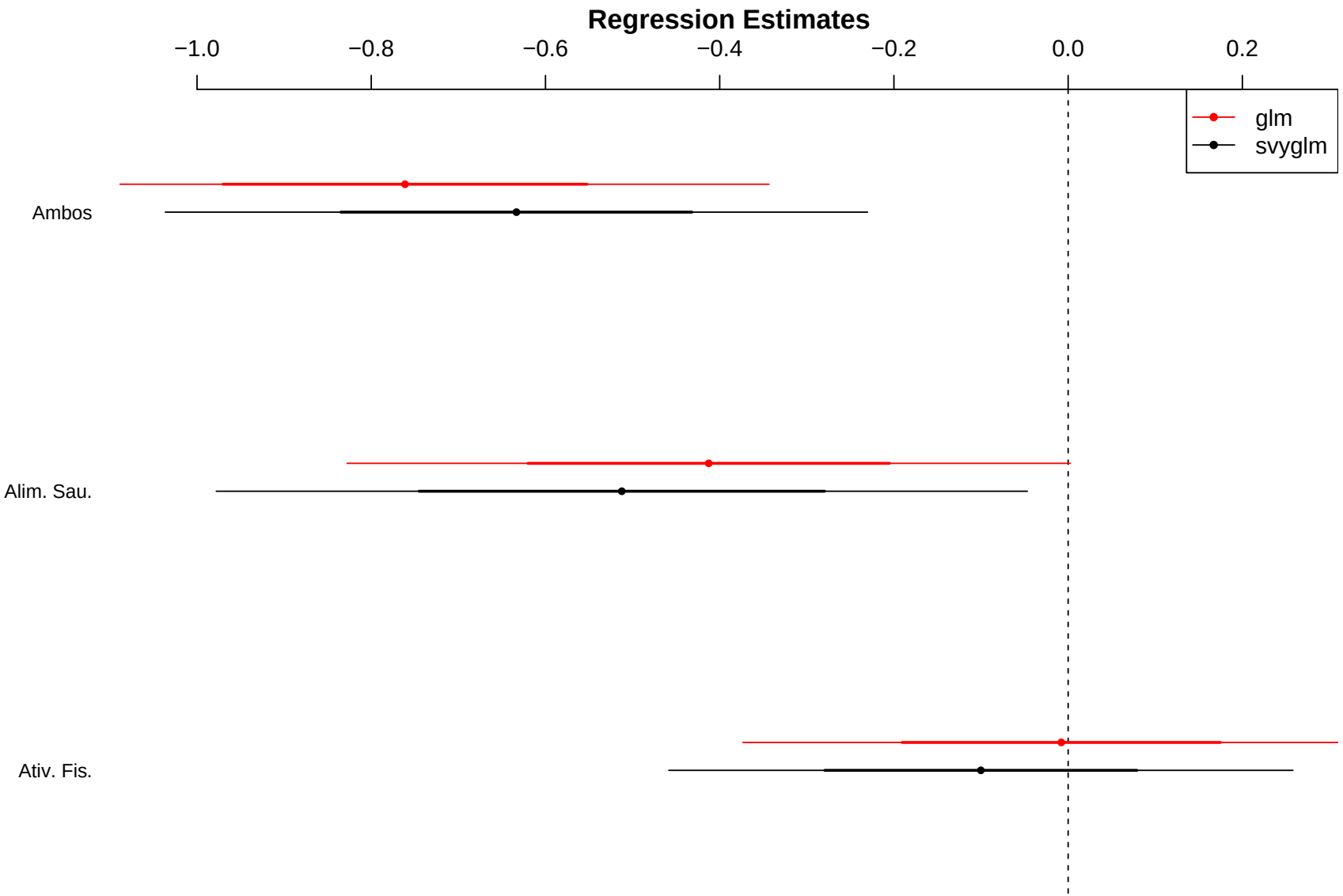
```
summary(output)
```

```
##
## Call:
## svyglm(formula = modelo, family = binomial, design = BH
##
## Survey design:
## svydesign(id = ~1, strata = NULL, fpc = NULL, weights =
##     data = BH)
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -1.8360     0.1520 -12.081   <2e-16
## compsaudeAtiv. Fis.       -0.1002     0.2164  -0.463   0.6434
## compsaudeAlim. Sau.       -0.5124     0.2459  -2.084   0.0373
## compsaudeAmbos            -0.6333     0.2475  -2.559   0.0106
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

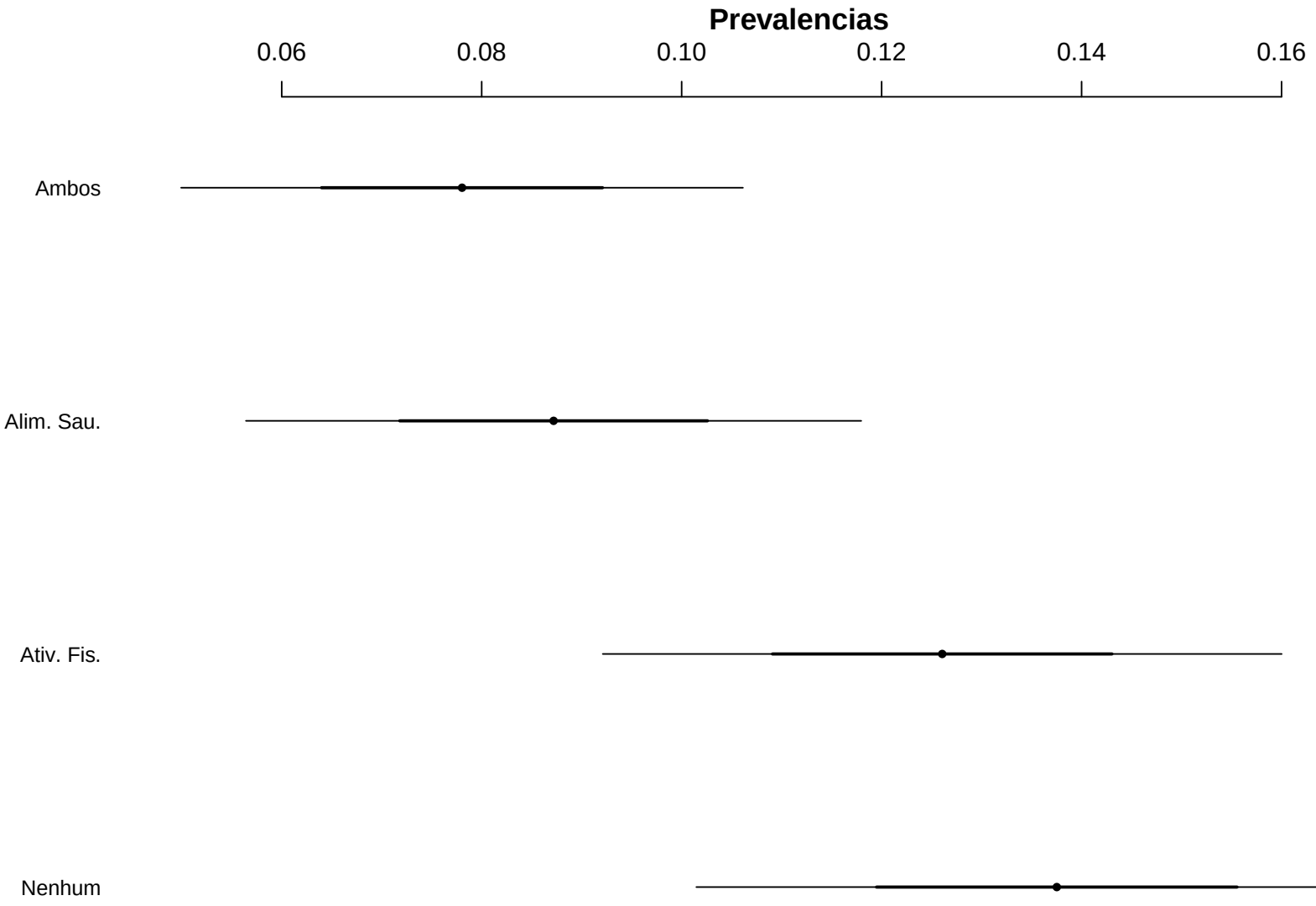
Coefplot


Regression Estimates
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
Ambos
Alim. Sau.
Ativ. Fis.

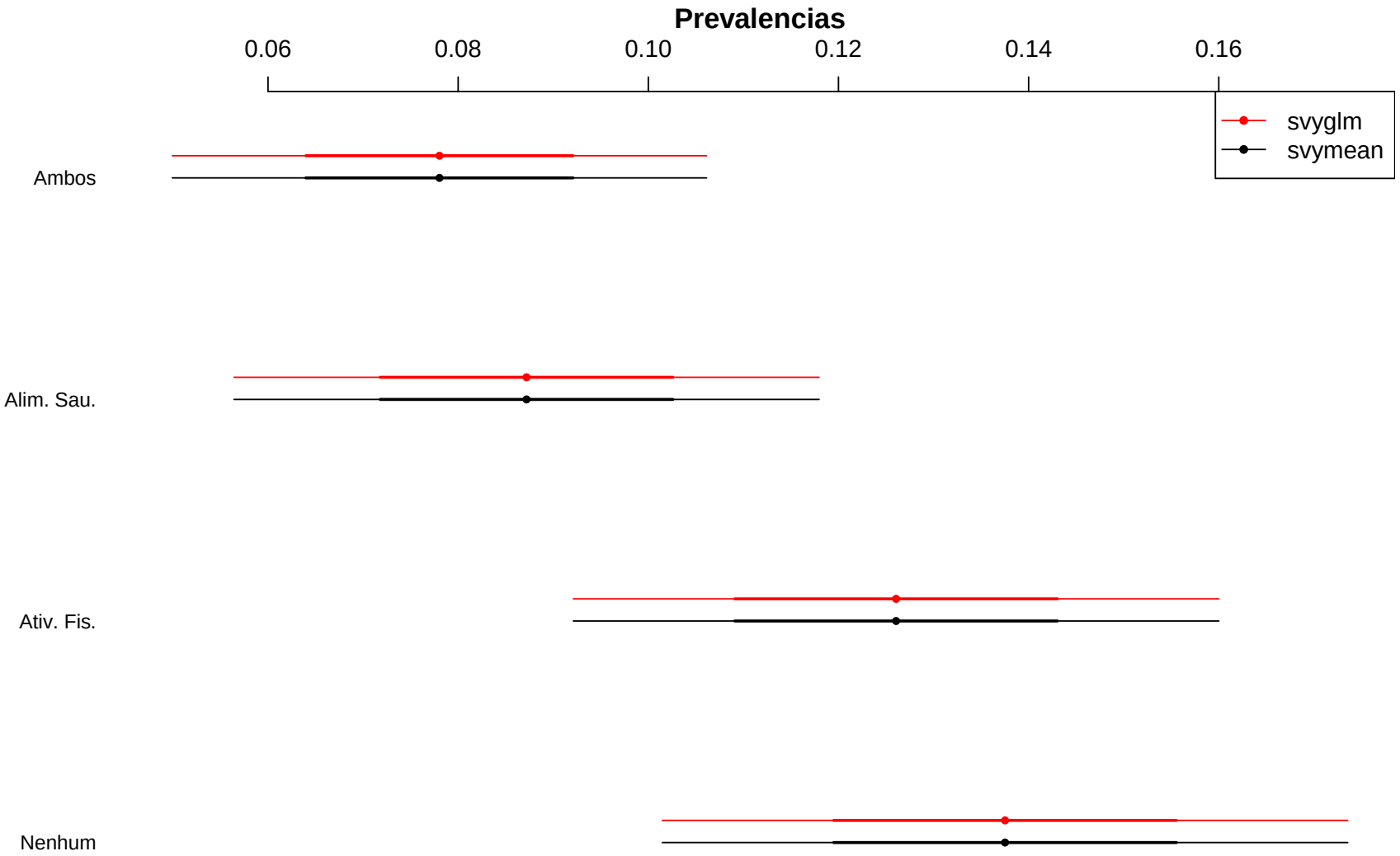
Coefplot


Regression Estimates
−1.0
−0.8
−0.6
−0.4
−0.2
0.0
0.2
glm
svyglm
Ambos
Alim. Sau.
Ativ. Fis.

Prevalências de tabagismo em BH, 2016



Prevalências de tabagismo em BH, 2016



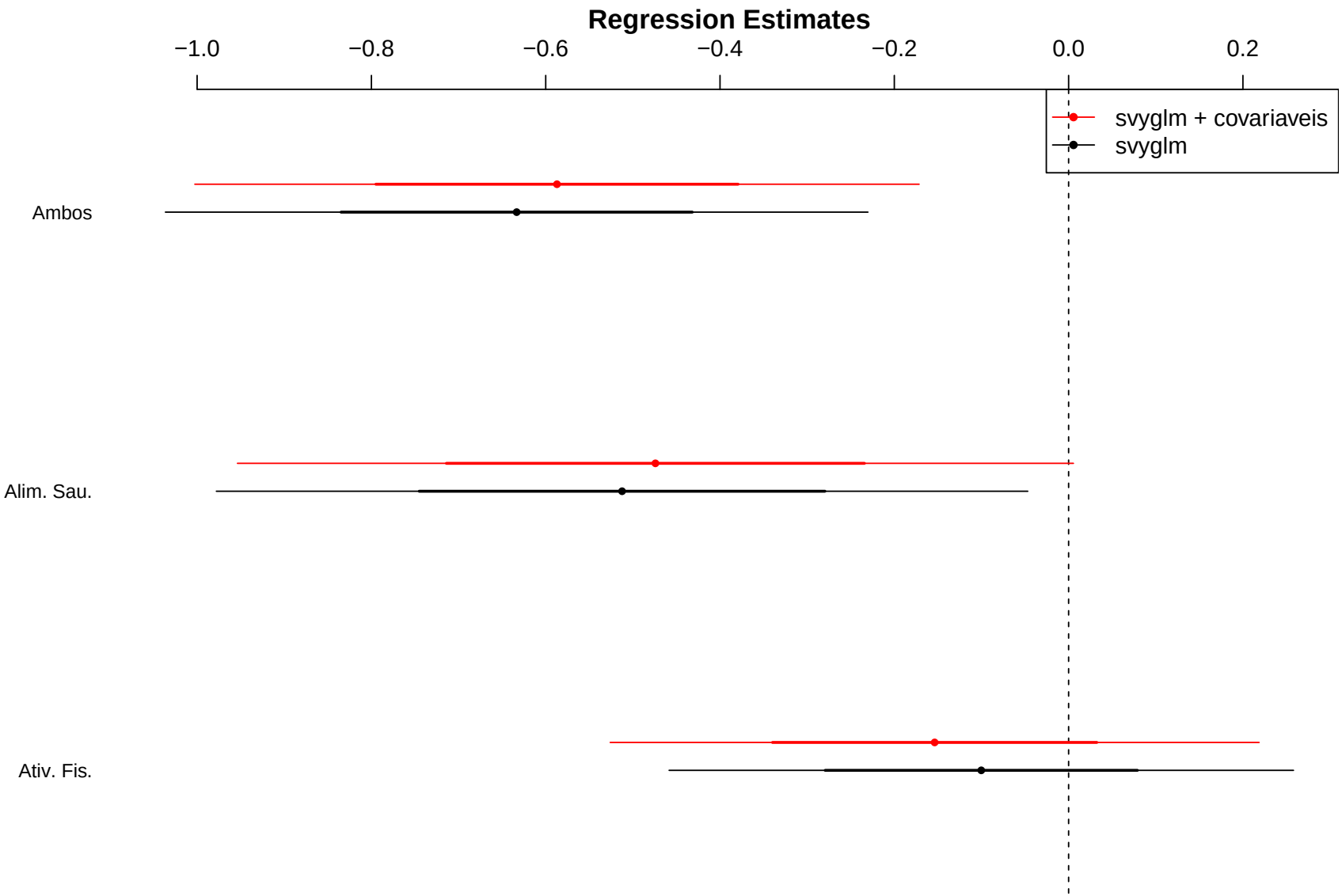
Controlando por outras variáveis

Modelo para estimar o efeito do comportamento saudável no tabagismo

```
# Modelo
modelo2 <- fumante ~ compsaude + sexo + fet + fesc

# Ajuste
output2 <- svyglm(formula = modelo2,
                  family = binomial,
                  design = BH.svy)
```

Coeficientes



Bayes (?)

- Modelos multinível incorporando o desenho
  - Toledo et al. (2017) – Dados da pesquisa Nacional do Crack
- Bayesian non-parametric
  - GP: Si, Pilai and Gelman (2015)
  - DP: Kunihama, Herring, Halpern and Dunson (2016)
- Pseudo-posterior
  - Savytsky and Toth (2016)

Principais Referências

1. Lumley, T. (2010) *Complex surveys: A guide to analysis using R*, Wiley.
2. Lumley, T. and Scott, A. (2017) Fitting Regression Models to Survey Data, *Statistical Science*, Vol. 32, No. 2, 265–278
3. Horvitz, D.G., Thompson, D.J., (1952) A generalization of sampling without replacement from a finite universe. *JASA*, 47, 663–685.
4. Si, Y., Pilai, N. and Gelman, A. (2015) Bayesian Nonparametric Weighted Sampling Inference. *Bayesian Analysis*, 10, Number 3, pp. 605–625
5. Kunihama, T., Herring, A., Halpern, C. and Dunson, D. (2016) Nonparametric Bayes modeling with sample survey weights, *SPL*, 113, 41–48
6. Savytsky, T. and Toth, D. (2016) Bayesian estimation under informative sampling. *EJS*, 10, 1677–1708