Modelando dados amostrais complexos

Leo Bastos (PROCC/Fiocruz)

EAE2

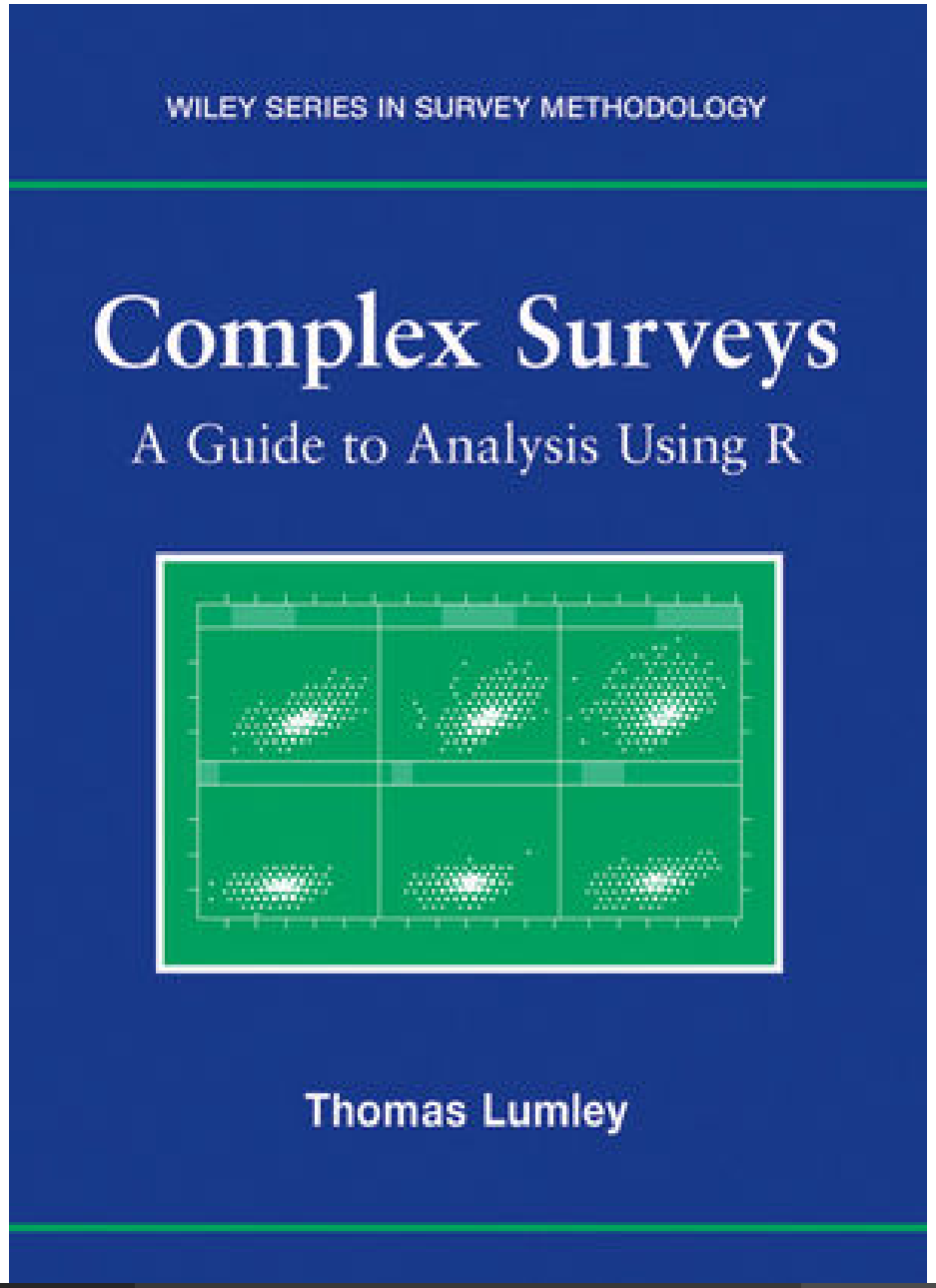
FIOCRUZ

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

FIOCRUZ

Referência principal



FIOCRUZ

Inferência estatística

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

FIOCRUZ

Exemplo

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

Figure 1: Região Metropolitana de Belo Horizonte à noite a partir da Estação Espacial Internacional. (Wikipedia)

FIOCRUZ

Tabagismo

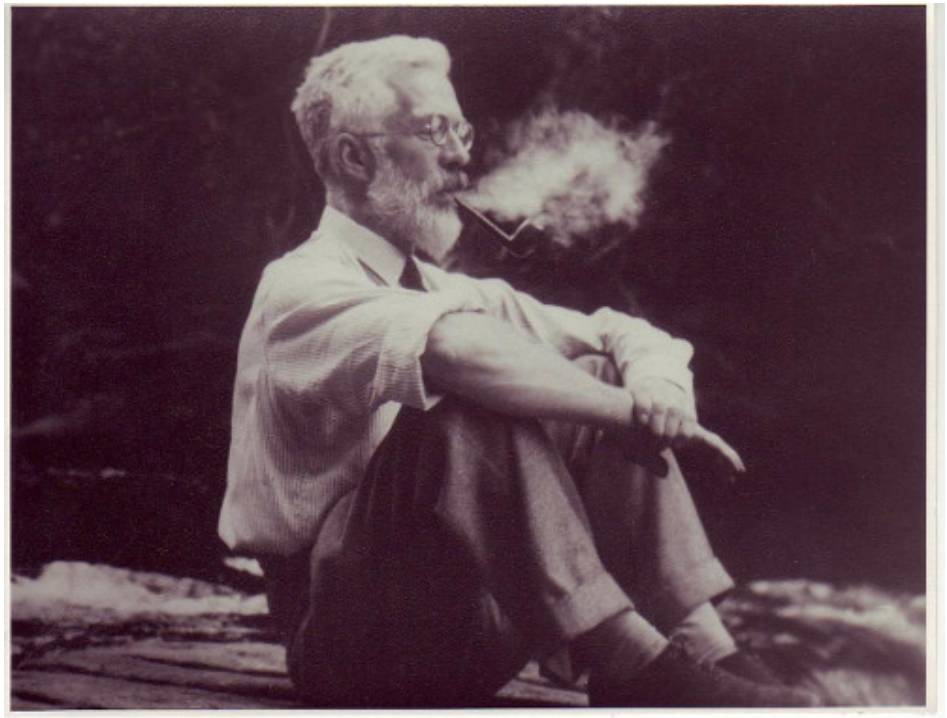


Figure 2: Estatístico famoso em 1946.

FIOCRUZ

Tabagismo no Mundo

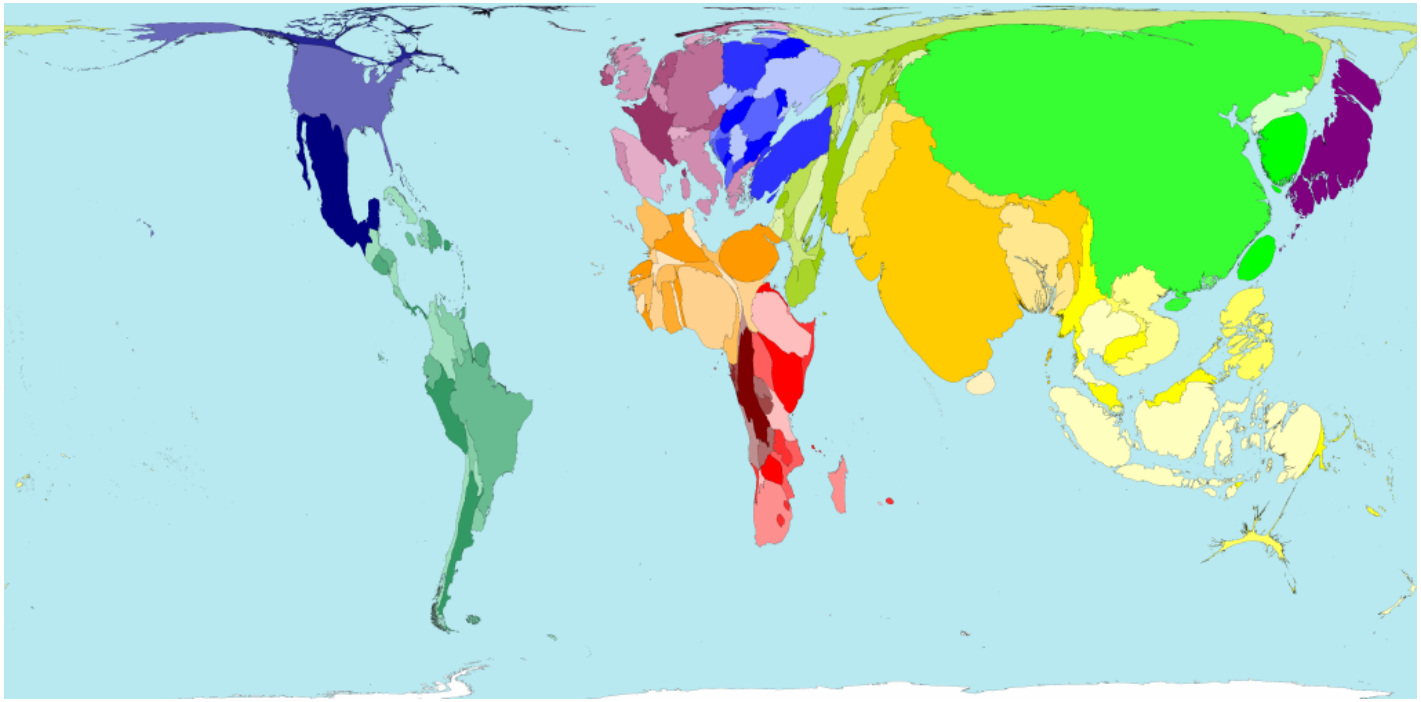


Figure 3: Cartograma da prevalência de tabagismo em homens no mundo.

FIOCRUZ

Tabagismo no Mundo

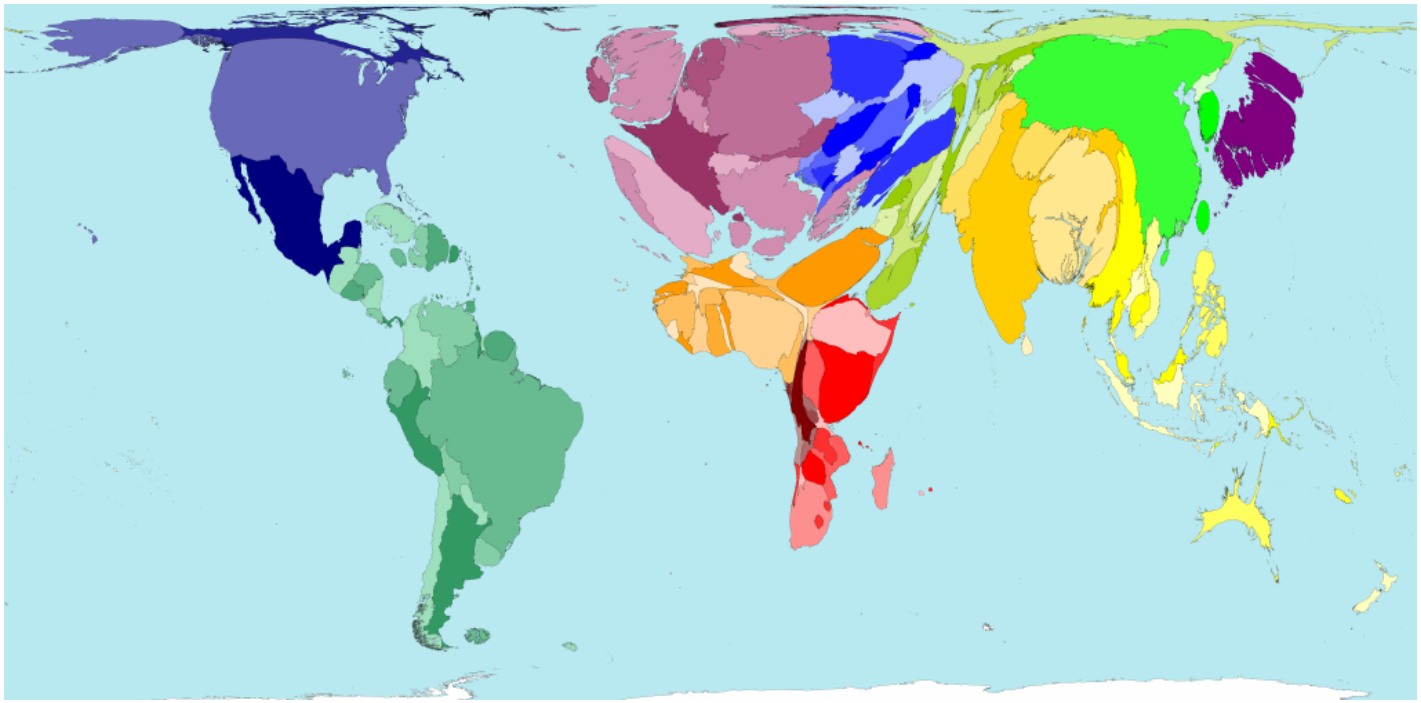


Figure 4: Cartograma da prevalência de tabagismo em mulheres no mundo.

FIOCRUZ

Como estimar o tabagismo em BH?

- Vamos coletar uma amostra ~~representativa~~ adequada para estimar a prevalência de tabagismo na 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 ser adequada?
- Se n fosse a população de BH, saberíamos exatamente a prevalência de fumantes em BH.

FIOCRUZ

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.

FIOCRUZ

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.

FIOCRUZ

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~~ adequada para estimar a prevalência de fumantes dos belorizontinos?

FIOCRUZ

Amostras probabilísticas

- "Selecionamos uma amostra aleatória de tamanho 1000"
- O conceito fundamental da inferência baseada no desenho é a **amostra aleatória** ou **amostra probabilística**
- **Lei Forte dos Grandes Números** nos permite afirmar que uma amostra de tamanho 1000 é representativa para a população de interesse, quando o interesse é estimar médias e/ou proporções.

$$P(\lim_{n\to\infty} |\bar{X} - \mu| < \epsilon) = 1$$

FIOCRUZ

Teorema Central do Limite

- No livro do Barry James tem uma nota sobre o nome do Teorema.
- Existem várias versões para o TCL (Levy, Lindeberg, Liapunov,..)
- Seja $X_1, X_2, \ldots$ v.a.s i.i.d. com média μ e variância finita $\sigma^2 > 0$, então

$$\frac{\bar{X} - \mu}{\sigma} \rightarrow_d N(0, 1/n)$$

- Independência pode ser uma suposição forte quando a população é finita.
- Baseado nesse resultado podemos, por exemplo, construir um IC de 95%.

FIOCRUZ

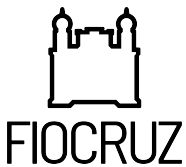
Versão especial do TCL para AAS sem reposição

- Existe uma versão especial do TCL para AAS sem reposição de uma população finita (Erdos and Renyi, 1959).

$$\frac{\bar{x}_n - \mu_N}{\sqrt{var(\bar{x}_n)}} \rightarrow_d N(0,1)$$

quando n cresce e $N - n$ ainda é grande.

- Se a amostra não for AAS, ainda pode-se ter aguma forma do TCL. Mas deve-se tomar cuidado ao construir intervalos e realizar testes de hipóteses.



Propriedades de uma amostra probabilística

- 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.

FIOCRUZ

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

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

FIOCRUZ

Peso amostral

- Se pegarmos uma amostra aleatória de tamanho 2500 de moradores da cidade do Rio (População $\approx$ 6.5 milhões pessoas), então cada pessoa teria chance aproximada de 3.84 em 10 mil de ser selecionada.
- Cada pessoa amostrada no Rio representaria 2600 moradores do Rio.
- Se por acaso, 500 pessoas amostradas forem fumantes, significa que no Rio tem 1.3 milhões de fumantes (500 * 2600).

FIOCRUZ

Peso amostral

- Se pegarmos uma amostra de tamanho 2500 de moradores de Niterói (Pop. $\approx$ 500 mil hab.), então cada morador teria probabilidade de 0.5%, ou 5 em mil, de ser selecionado.
- Cada pessoa amostrada em Niterói representaria 200 moradores de Niterói.
- Se por acaso, 500 pessoas amostradas forem fumantes, significa que Niterói tem 100 mil fumantes (200 * 500).

FIOCRUZ

Pesos

O GLOBO

elu

Downton Abbey
DAMAS DE HOJE

PANTALONA CURTA: A CALÇA DO VERÃO

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

FIOCRUZ

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 $\hat{T}_{HT}$

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

FIOCRUZ

Tipos de peso

- **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.

FIOCRUZ

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

FIOCRUZ

VIGITEL

PS: Microdado disponível para download no site do Datasus.

FIOCRUZ

VIGITEL

No Vigitel tem-se a pergunta:

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

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

FIOCRUZ

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

FIOCRUZ

Software, dados, e scripts

- O pacote *survey*
- Descritivas para tabagismo em BH

FIOCRUZ

Softwares

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

FIOCRUZ

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")
```

FIOCRUZ

Comandos básicos

```r
# Chamando a biblioteca survey
library(survey)
library(tidyverse) # Para manipular os dados
library(readxl) # Para ler arquivos .xls e .xlsx

# Lendo os microdados do Vigitel 2023
vigitel <- read_xlsx("Data/Vigitel-2023-peso-rake.xlsx")

BH <- vigitel |> filter(cidade == 3)

# Tamanho da amostra
nrow(BH)
## [1] 802
```

FIOCRUZ

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
##       324       478
```

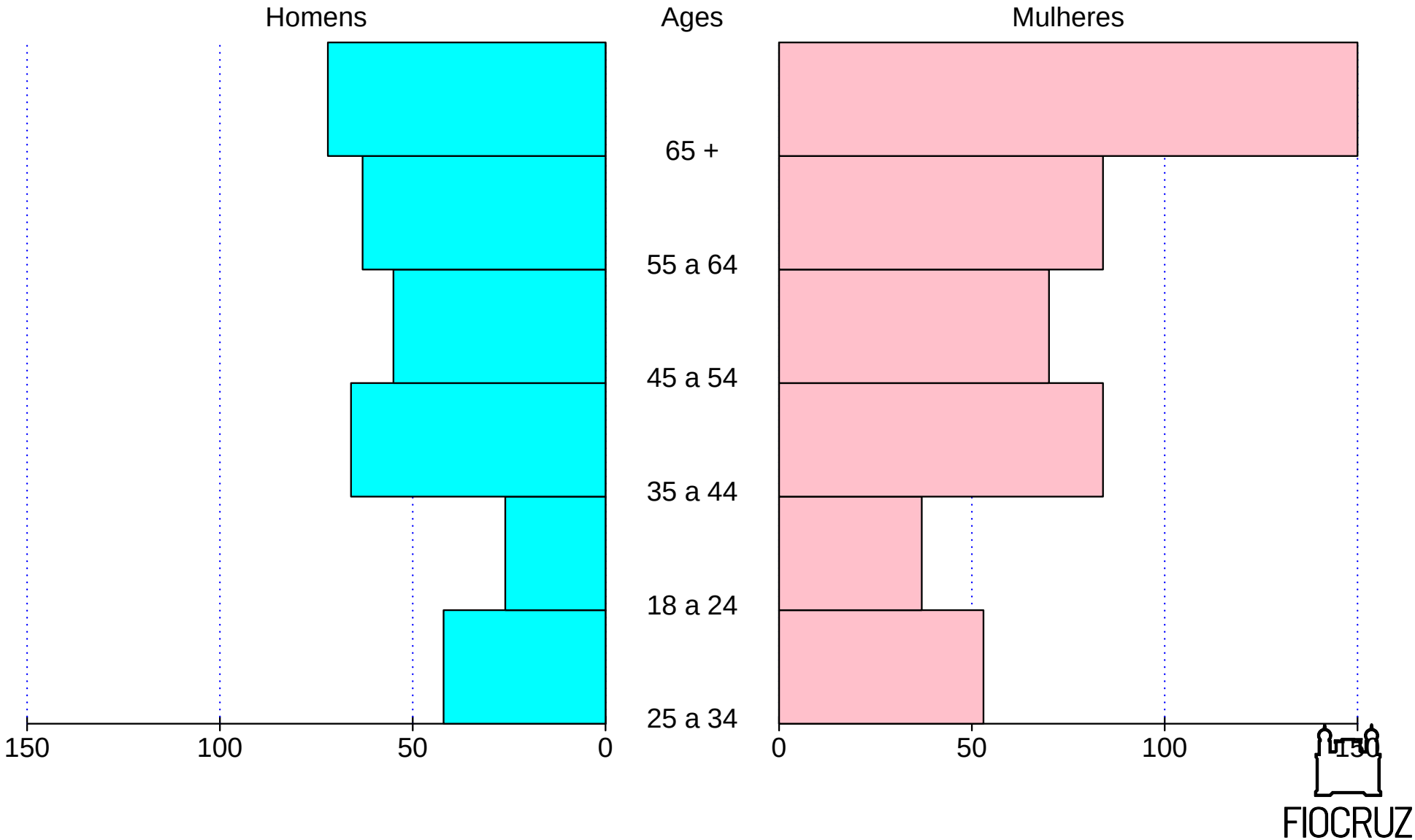
FIOCRUZ

Estimativas de tabagismo em BH, 2023 (por sexo)

```
# 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$sexo: Masculino
##       mean         sd
## 0.10185185 0.01680297
## ------------------------------------------------------------
## BH$sexo: Feminino
##       mean         sd
## 0.07740586 0.01222301
```

FIOCRUZ

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
```

FIOCRUZ

Descritivas para o Vigitel, BH, 2023

```
# Estimando o total de fumantes de BH em 2016
svytotal(~fumante, BH.svy)
##           total    SE
## fumante 183604 28966

# Estimando prevalência de tabagismo na capital
svymean(~fumante, BH.svy)
##             mean     SE
## fumante 0.09621 0.0147
```

FIOCRUZ

Estimativas de tabagismo em BH, 2023

```
# Tabagismo por sexo
svyby(formula = ~fumante, by = ~sexo, design = BH.svy,
      FUN = svymean)
##                  sexo    fumante         se
## Masculino Masculino 0.11935325 0.02665962
## Feminino   Feminino 0.07679441 0.01491016

# Tabagismo por escolaridade
svyby(formula = ~fumante, by = ~fesc, design = BH.svy,
      FUN = svymean)
##                            fesc    fumante         se
## 0 a 8 anos          0 a 8 anos 0.09267665 0.02336216
## 9 a 11 anos        9 a 11 anos 0.11781324 0.03047888
## 12 anos e mais 12 anos e mais 0.07720582 0.01914112
```

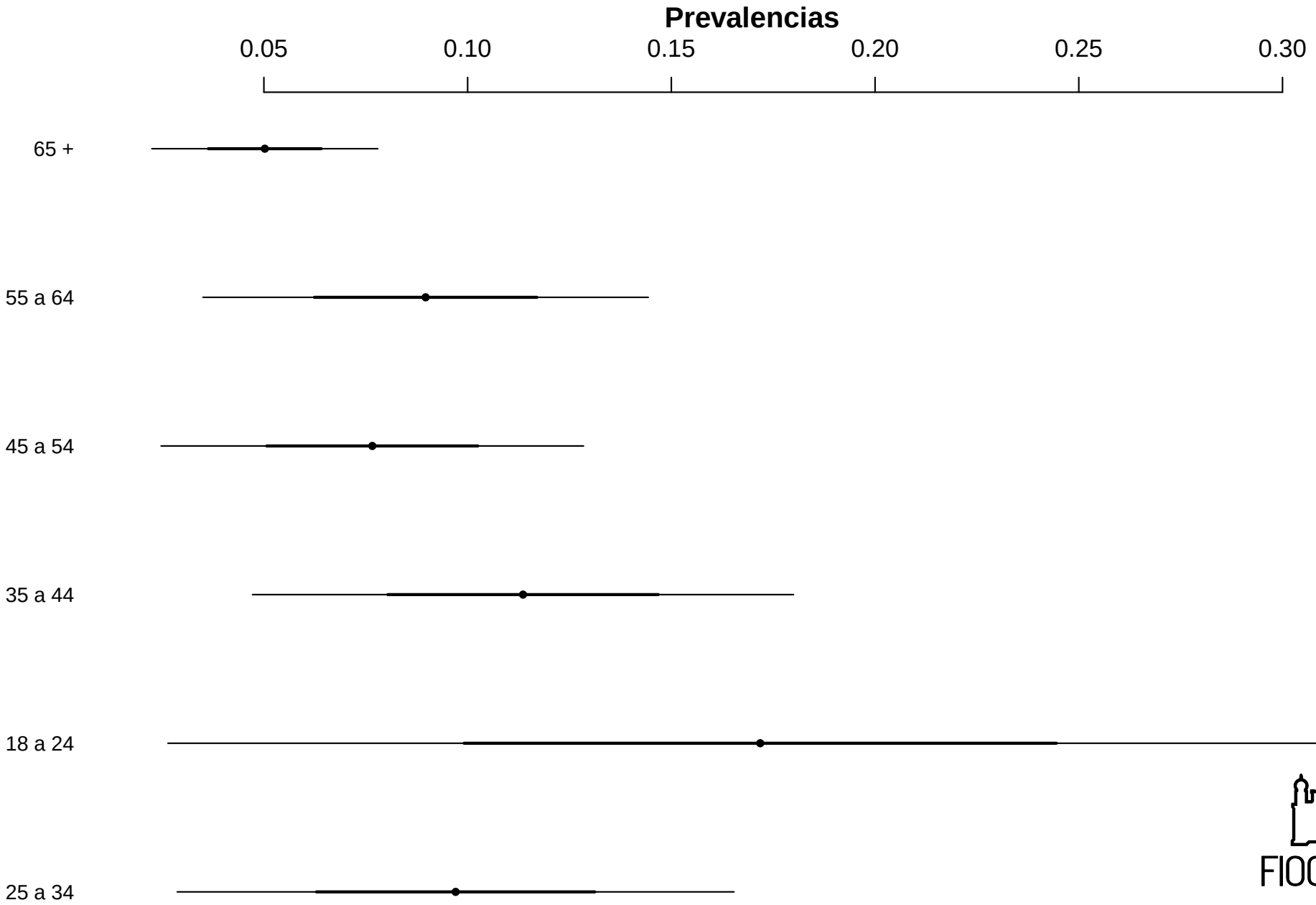
FIOCRUZ

Estimativas de tabagismo em BH, 2023 (por faixa etaria)

```
# Tabagismo por faixa etaria
svyby(formula = ~fumante, by = ~fet, design = BH.svy,
      FUN = svymean)[,-1]
##              fumante          se
## 25 a 34 0.09705258 0.03416331
## 18 a 24 0.17181559 0.07268760
## 35 a 44 0.11359869 0.03319699
## 45 a 54 0.07660731 0.02592153
## 55 a 64 0.08969375 0.02732338
## 65 +    0.05022370 0.01386329
```

FIOCRUZ

Prevalência de tabagismo segundo faixa etária



FIOCRUZ

Outras estatísticas descritivas usando o survey

- Médias e totais (svymean e svytotal)
- Quantis (svyquantile)
- Kappa, medida de concordância (svykappa)
- Gráficos
  - Histogramas (svyhist)
  - Boxplots (svyboxplot)
- etc

FIOCRUZ

Modelos estatísticos para amostras complexas usando o *survey*

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

FIOCRUZ

Modelos 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.

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

FIOCRUZ

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 função de **pseudo-verossimilhanca** (Lumley and Scott, 2017)

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

FIOCRUZ

Tabagismo em BH, 2023

Modelo para estimar o efeito do sexo no tabagismo

```
# Modelo
modelo <- fumante ~ fet

# Ajuste sem pesos
output0 <- glm(modelo, data = BH, family = binomial)

# Ajuste com pesos
output <- svyglm(formula = modelo,
                 family = binomial,
                 design = BH.svy)
## Warning in eval(family$initialize): non-integer #successes
```

FIOCRUZ

Tabagismo em BH, 2023

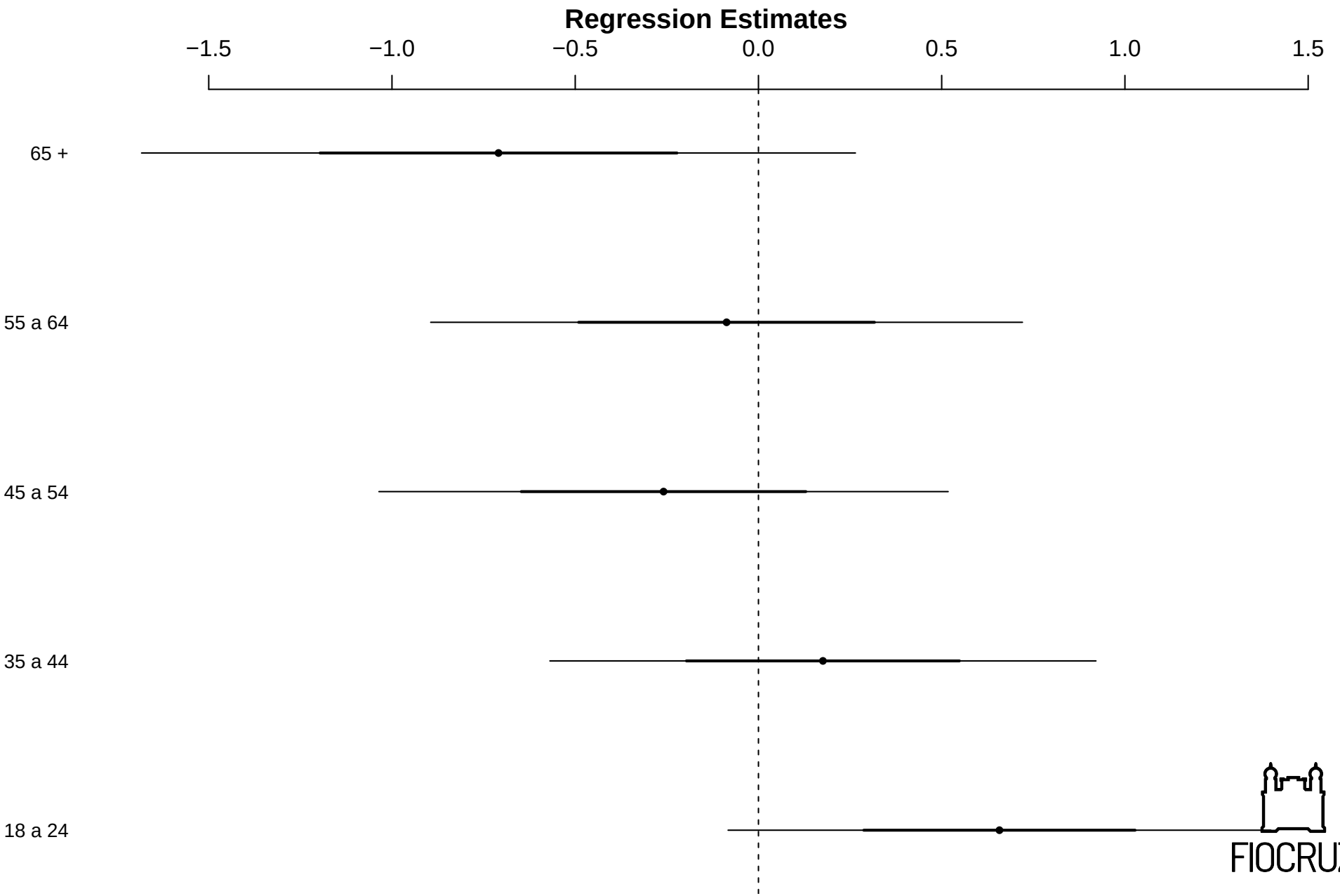
```
# Coeficientes estimados
cbind( Sem_Pesos = coef(output0), Com_Pesos = coef(output))
##                 Sem_Pesos   Com_Pesos
## (Intercept) -2.14006616 -2.23041139
## fet18 a 24  -0.31093893  0.65759732
## fet35 a 44  -0.05715841  0.17591360
## fet45 a 54  -0.30228087 -0.25895071
## fet55 a 64  -0.03468556 -0.08696858
## fet65 +     -0.48460243 -0.70932810
```

FIOCRUZ

Tabagismo em BH, 2023

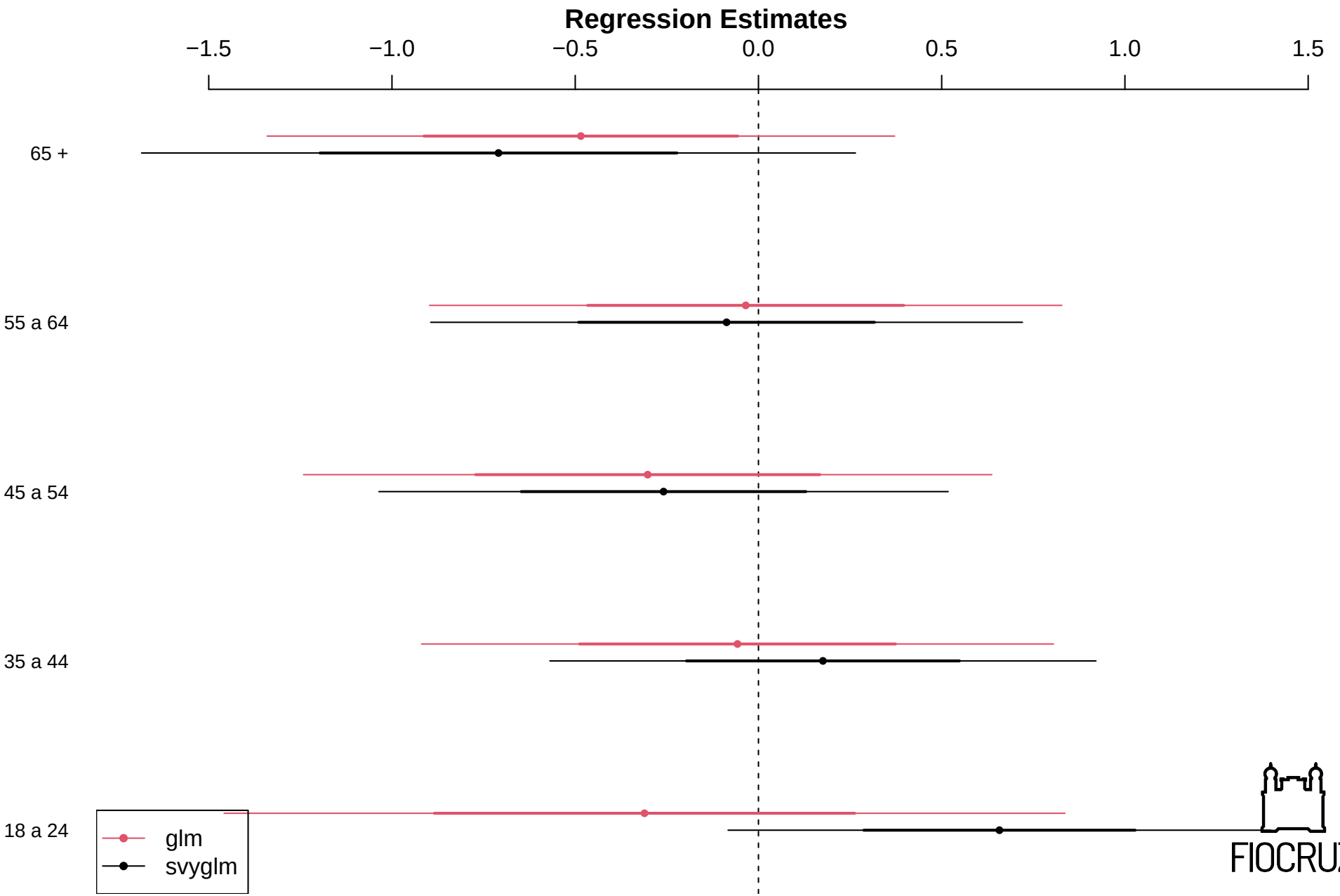
```
summary(output)
##
## Call:
## svyglm(formula = modelo, design = BH.svy, family = binomial
##
## Survey design:
## svydesign(id = ~1, strata = NULL, fpc = NULL, weights = ~pe
##     data = BH)
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.23041    0.38984  -5.721  1.5e-08 ***
## fet18 a 24   0.65760    0.64259   1.023    0.306
## fet35 a 44   0.17591    0.51056   0.345    0.731
## fet45 a 54  -0.25895    0.53503  -0.484    0.629
## fet55 a 64  -0.08697    0.51378  -0.169    0.866
## fet65 +     -0.70933    0.48625  -1.459    0.145
```

Coefplot



FIOCRUZ

Coefplot



FIOCRUZ

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

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

FIOCRUZ

Prevalências de tabagismo em BH, 2016

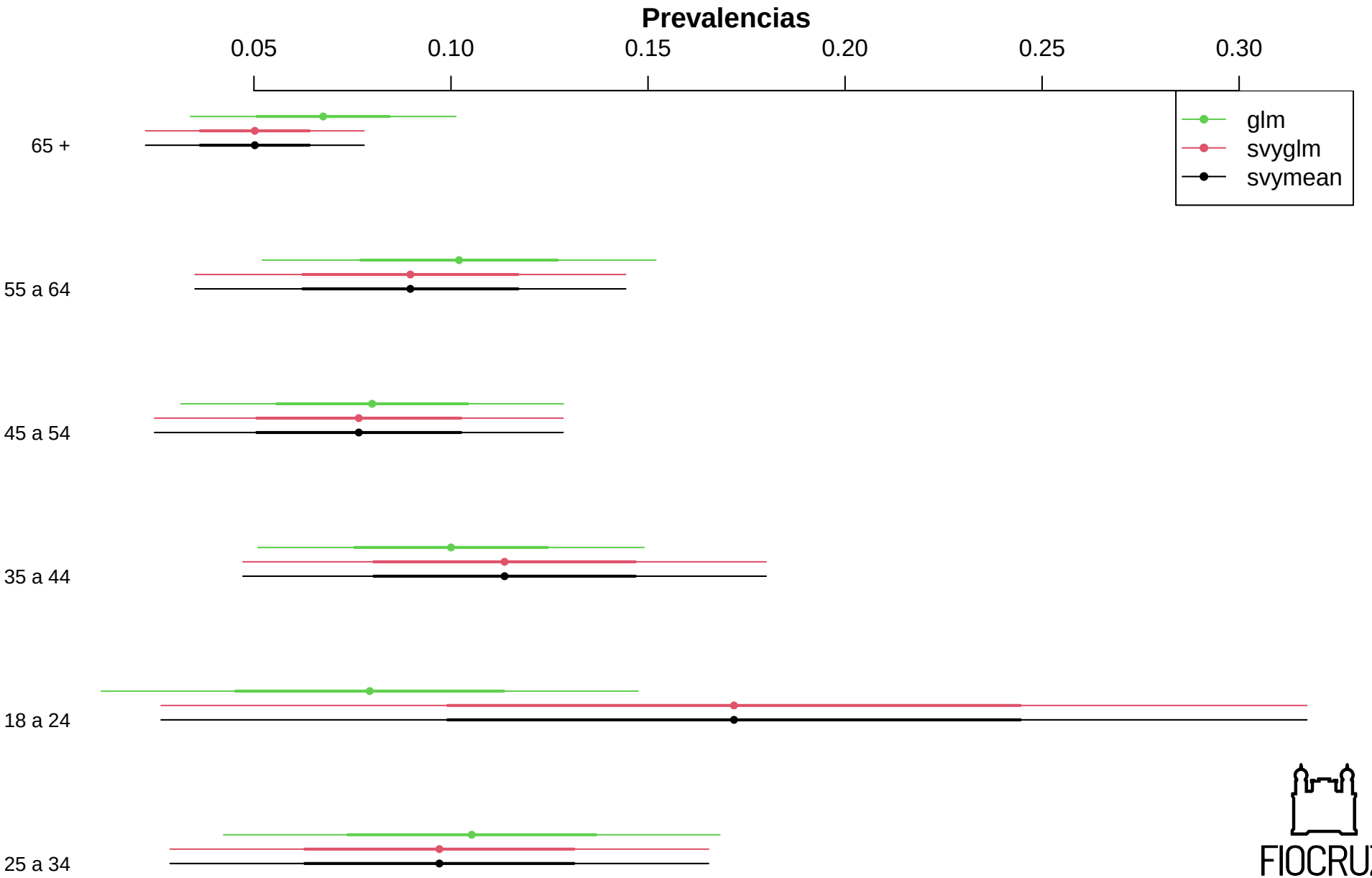
```
P1 <- svyby(formula = ~fumante, by = ~fet,
            design = BH.svy, FUN = svymean)

P2 <- predict( output, type = "response" ,se.fit = T,
               newdata = data.frame(
                 fet = levels(BH$fet)
                 )
               )

P3 <- predict( output0, type = "response" ,se.fit = T,
               newdata = data.frame(
                 fet = levels(BH$fet)
                 )
               )
```

FIOCRUZ

Prevalências de tabagismo em BH, 2023



FIOCRUZ

Controlando por outras variáveis

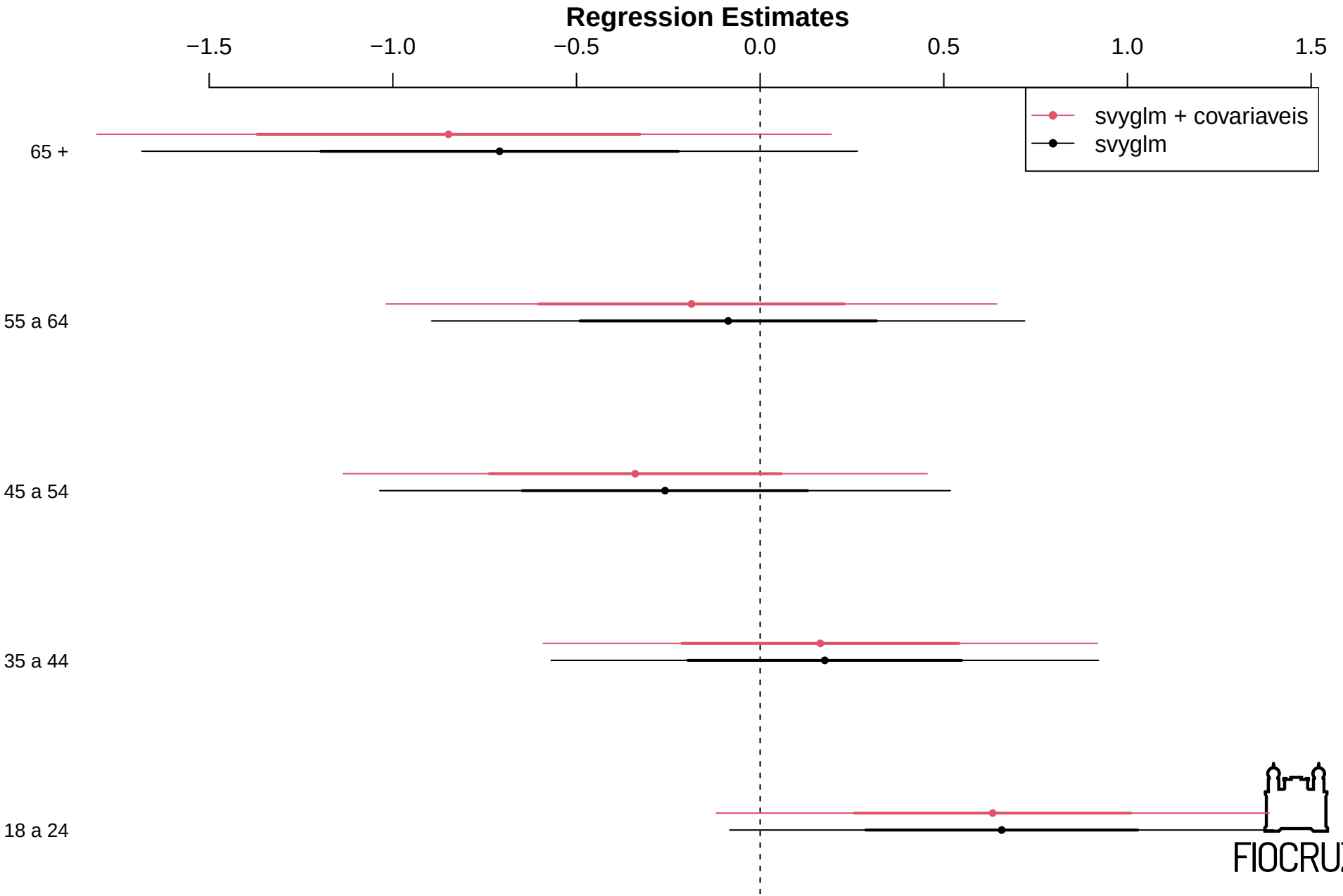
Modelo para estimar o efeito da idade no tabagismo controlando por sexo e escolaridade

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

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

FIOCRUZ

Coeficientes



Proposta de exercício

- Replicar o código com tabagismo em BH, 2023
- Trocar o desfecho (e cidade, olhar dicionário).
  - Ex. obesidade (obesid_i)
  - Hipertensão arterial (hart)
  - etc.
- Avaliar estimativas de um mesmo desfecho em 2013 e 2023
  - Microdados disponíveis em: https://svs.aids.gov.br/download/Vigitel/

FIOCRUZ

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

FIOCRUZ