

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

Análisis de encuestas de hogares con R

Modulo 5: Modelos de regresión

Andrés Gutiérrez, Ph.D.
Stalyn Guerrero M.Sc.

CEPAL - Unidad de Estadísticas Sociales

1 Procesando múltiples bases.

Lectura de la base

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

```
encuesta <- readRDS("../Data/encuesta.rds")  
data("BigCity", package = "TeachingSampling")
```

Definir diseño de la muestra con srvyr

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

```
library(srvyr)
options(survey.lonely.psu="adjust")
diseno <- encuesta %>%
  as_survey_design(
    strata = Stratum,
    ids = PSU,
    weights = wk,
    nest = T
  )
```

Sub-grupos

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

Extraer sub-grupos de la encuesta.

```
sub_Urbano <- diseno %>% filter(Zone == "Urban")
sub_Rural <- diseno %>% filter(Zone == "Rural")
sub_Mujer <- diseno %>% filter(Sex == "Female")
sub_Hombre <- diseno %>% filter(Sex == "Male")
```

Modelo de regresión

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn Guerrero
M.Sc.

Procesando múltiples bases.

$$y = \beta_0 + \beta_1 x + \epsilon$$

$$E(y | x) = B_0 + B_1 x$$

donde \$ B = [B_0, B1]\$ y el estimador de \$ B \$ esta dado por:

$$\hat{B} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}$$

$$F(B) = \sum_{i=1}^N (y_i - \mathbf{x}_i \mathbf{B})^2$$

$$\widehat{WSSE}_{pop} = \sum_h^H \sum_{\alpha}^{a_h} \sum_{i=1}^{n_{h\alpha}} w_{h\alpha i} (y_{hai} - \mathbf{x}_{h\alpha i} \mathbf{B})^2$$

Modelo nulo (Q_W)

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

```
modNul <- svyglm(Income ~ 1, design = diseno)
fit_Nul <- lm(wk ~ 1, data = encuesta)
qw <- predict(fit_Nul)

encuesta %>% mutate(wk1 = wk/qw)

diseno_qwgt <- encuesta %>%
  as_survey_design(
    strata = Stratum,
    ids = PSU,
    weights = wk1,
    nest = T
)
modNul_qw <- svyglm(Income ~ 1, design = diseno_qwgt)
```

Scaterplot con los datos poblacionales

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn Guerrero
M.Sc.

Procesando múltiples bases.

```
library(ggplot2); library(ggpmisc)
plot_BigCity <-
  ggplot(data = BigCity,
         aes(x = Expenditure, y = Income)) +
  geom_point() +
  geom_smooth(method = "lm",
              se = FALSE,
              formula = y ~ x) +
  theme_cepal()

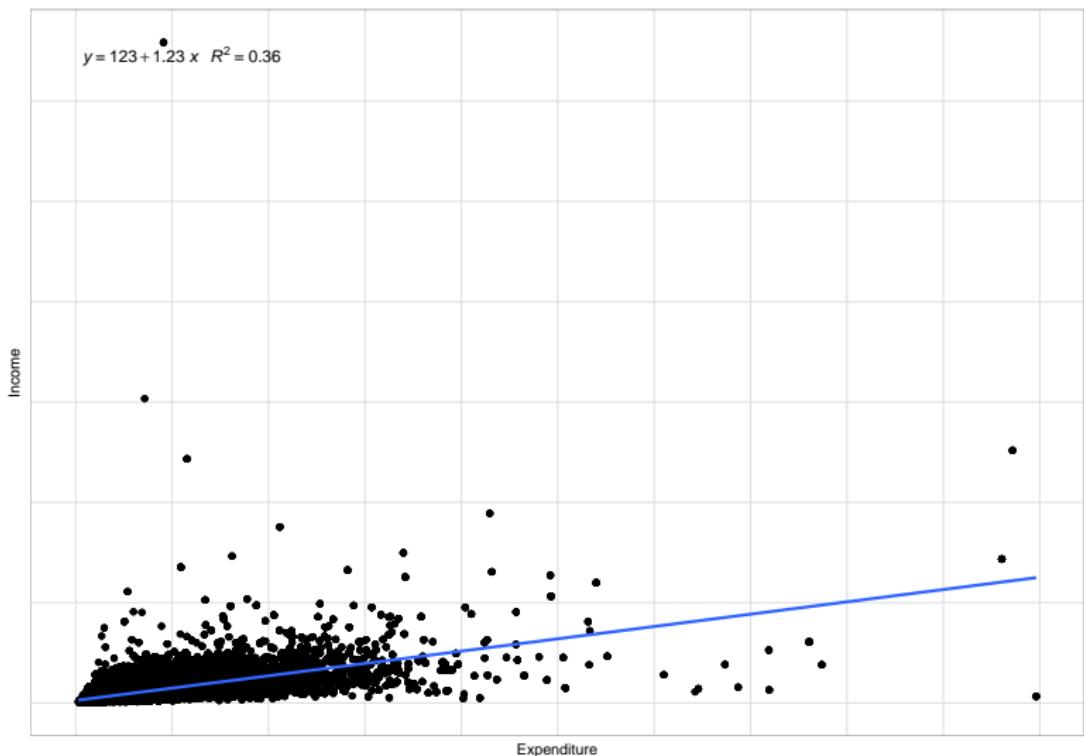
plot_BigCity + stat_poly_eq(formula = y~x,
                            aes(label = paste(..eq.label..,
                            ..rr.label.., sep = "~~~")),
                            parse = TRUE)
```

Scaterplot con los datos poblacionales

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.



Modelo poblacional

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

```
library(modelsummary)
fit <- lm(Income ~ Expenditure, data = BigCity)
modelsummary(list(Pob = fit), statistic = NULL,
            title = "Modelo BigCity",
            output = "markdown",
            gof_omit = 'BIC|Log|AIC|F' )
```

Table 1: Modelo BigCity

	Pob
(Intercept)	123.337
Expenditure	1.229
Num.Obs.	150266
R2	0.359
R2 Adj.	0.359
RMSE	461.74

Scaterplot con los datos encuesta sin ponderar

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

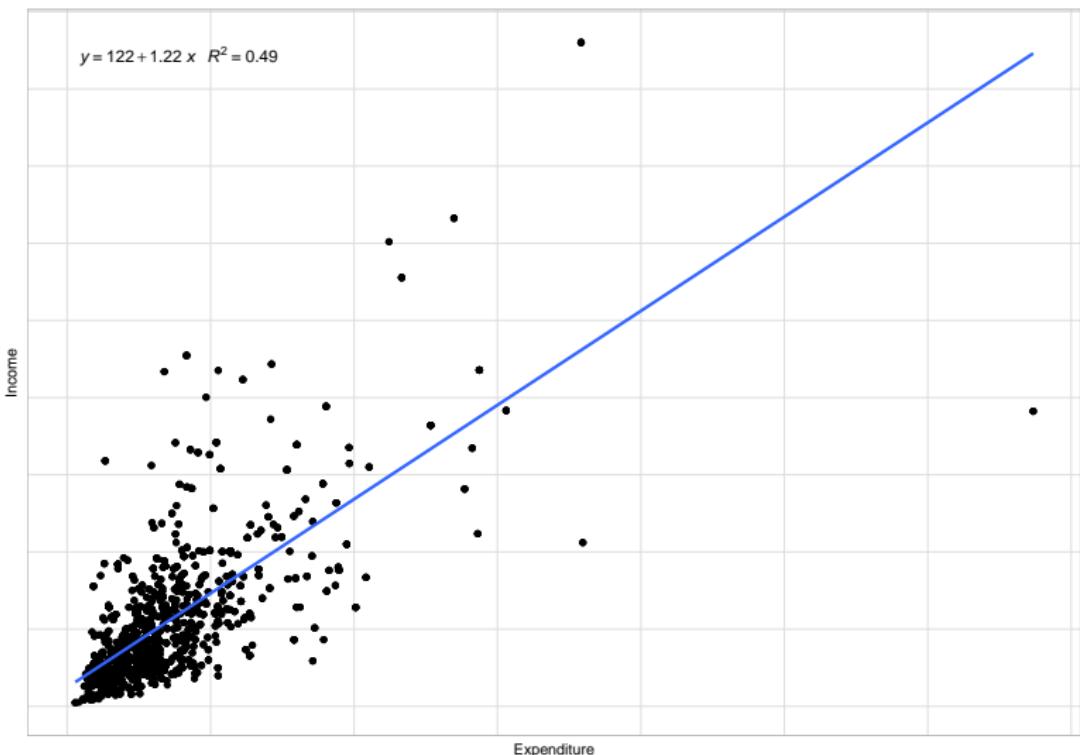
```
plot_sin <-
  ggplot(data = encuesta,
         aes(x = Expenditure, y = Income)) +
  geom_point() +
  geom_smooth(method = "lm",
              se = FALSE,
              formula = y ~ x) +
  theme_cepal()
plot_sin + stat_poly_eq(formula = y~x,
aes(label = paste(..eq.label..,
..rr.label.., sep = "~~~")), parse = TRUE)
```

Scaterplot con los datos encuesta sin ponderar

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.



Modelo sin ponderar

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

```
fit_sinP <- lm(Income ~ Expenditure, data = encuesta)
stargazer(fit_sinP, header = FALSE,
           title = "Modelo encuesta Sin ponderar",
           style = "ajps")
```

Modelo sin ponderar

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

Table 2: Modelo encuesta Sin ponderar

	Income
Expenditure	1.220*** (0.025)
Constant	121.500*** (11.410)
N	2605
R-squared	0.487
Adj. R-squared	0.487
Residual Std. Error	345.000 (df = 2603)
F Statistic	2473.000*** (df = 1; 2603)

***p < .01; **p < .05; *p < .1

Scaterplot con los datos encuesta ponderado

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

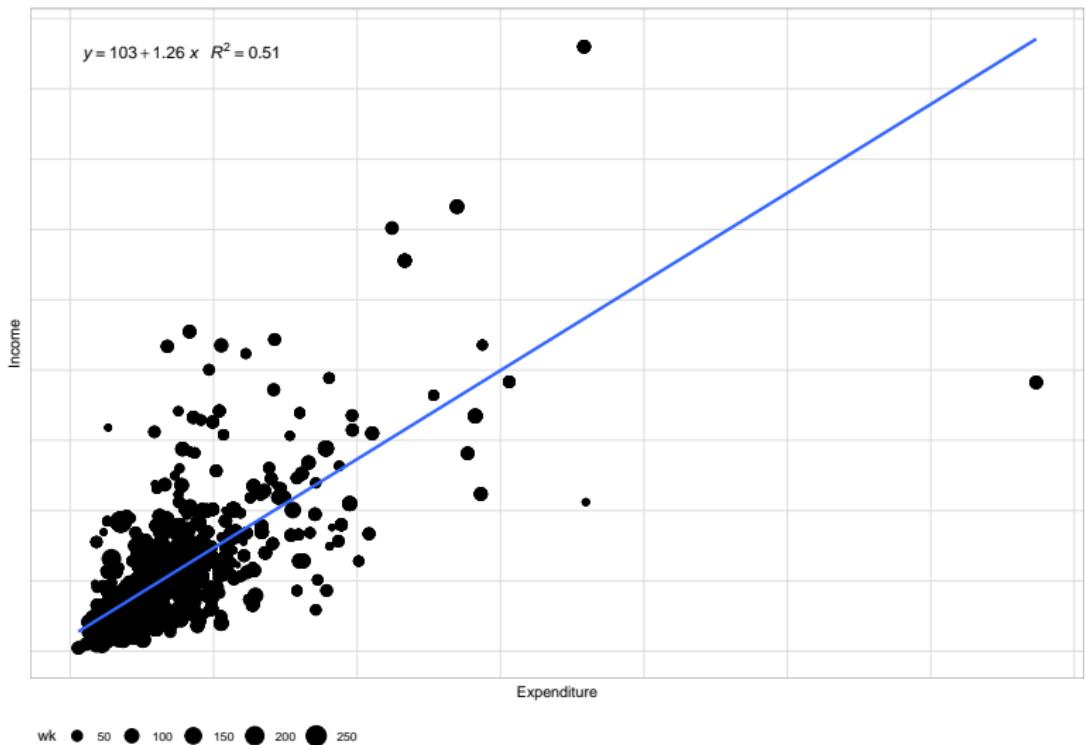
```
plot_Ponde <-
  ggplot(data = encuesta,
         aes(x = Expenditure, y = Income)) +
  geom_point(aes(size = wk)) +
  geom_smooth(method = "lm",
              se = FALSE,
              formula = y ~ x,
              mapping = aes(weight = wk)) +
  theme_cepal()
plot_Ponde + stat_poly_eq(formula = y~x,
                           aes(weight = wk,
                               label = paste(..eq.label..,
                               ..rr.label.., sep = "~~~")),
                           parse = TRUE)
```

Scaterplot con los datos encuesta sin ponderar

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.



Modelo ponderado lm

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

```
fit_Ponde <- lm(Income ~ Expenditure,  
                  data = encuesta, weights = wk)  
stargazer(fit_Ponde, header = FALSE,  
          title = "Modelo encuesta ponderada",  
          style = "ajps")
```

Modelo ponderado lm

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

Table 3: Modelo encuesta ponderada

	Income
Expenditure	1.263*** (0.024)
Constant	103.100*** (11.260)
N	2605
R-squared	0.509
Adj. R-squared	0.509
Residual Std. Error	2627.000 (df = 2603)
F Statistic	2703.000*** (df = 1; 2603)

***p < .01; **p < .05; *p < .1

Modelo ponderado svyglm

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

```
fit_svy <- svyglm(Income ~ Expenditure, design = diseño)
modNul <- svyglm(Income ~ 1, design = diseño)
s1 <- summary(fit_svy)
s0 <- summary(modNul)
```

Calculo del R^2

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

$$R^2 = 1 - \frac{SSE}{SST}$$

donde

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = (y_i - \mathbf{x}_i \mathbf{B})^2$$

$$R^2_{weighted} = 1 - \frac{WSSE}{WSST}$$

Calculo del R^2

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

```
s1$dispersion
```

```
##      variance      SE
## [1,]    119563 19005
```

```
s0$dispersion
```

```
##      variance      SE
## [1,]    243719 47168
```

(**R2** = 1-78320/149477)

```
## [1] 0.476
```

```
n = sum(diseno_qwgt$variables$wk)
```

(**R2Adj** = 1-((1-R2)*(n-1)/(n-1-1)))

Resumen del Modelo

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

Table 4: Modelo encuesta ponderada, svyglm

	Income
Expenditure	1.263*** (0.191)
Constant	103.100 (64.780)
N	2605
AIC	38281.000

***p < .01; **p < .05; *p < .1

Comparando los resultados

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

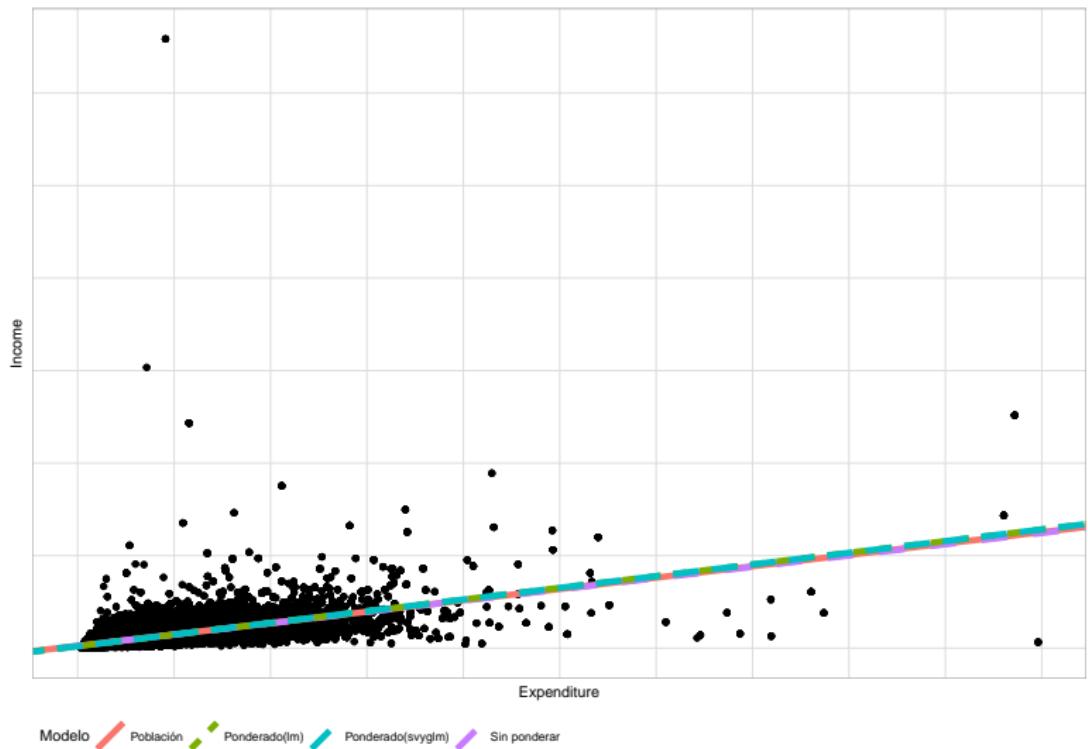
```
df_model <- data.frame(  
    intercept = c(coefficients(fit)[1],  
                  coefficients(fit_sinP)[1],  
                  coefficients(fit_Ponde)[1],  
                  coefficients(fit_svy)[1]),  
    slope = c(coefficients(fit)[2],  
              coefficients(fit_sinP)[2],  
              coefficients(fit_Ponde)[2],  
              coefficients(fit_svy)[2]),  
    Modelo = c("Población", "Sin ponderar",  
              "Ponderado(lm)", "Ponderado(svyglm)"))  
plot_BigCity + geom_abline( data = df_model,  
    mapping = aes( slope = slope,  
                  intercept = intercept, linetype = Modelo,  
                  color = Modelo ), size = 2  
)
```

Comparando los resultados

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn Guerrero
M.Sc.

Procesando múltiples bases.



Comparando los resultados

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

	Pob
(Intercept)	123.337
Expenditure	1.229
Num.Obs.	150266
R2	0.359
R2 Adj.	0.359
AIC	2270206.0
F	84052.758
RMSE	461.74

Comparando los resultados

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

	Sin Pond	Ponde(lm)	Ponde(svyglm)
(Intercept)	121.516	103.136	103.136
	p = (0.000)	p = (0.000)	p = (0.114)
Expenditure	1.220	1.263	1.263
	p = (0.000)	p = (0.000)	p = (0.000)
Num.Obs.	2605	2605	2605
R2	0.487	0.509	0.509
R2 Adj.	0.487	0.509	-9.826
AIC	37841.7	38281.0	123.5
F	2472.919	2702.978	43.885
RMSE	344.88	345.09	345.09

Metodología de los Q_Weighting de pfefferman

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn Guerrero
M.Sc.

Procesando múltiples bases.

```
fit_wgt <- lm(wk ~ Expenditure, data = encuesta)
wgt_hat <- predict(fit_wgt)
encuesta %>% mutate(wk2 = wk/wgt_hat)

diseno_qwgt <- encuesta %>%
  as_survey_design(
    strata = Stratum,
    ids = PSU,
    weights = wk2,
    nest = T
  )
```

Modelos empleando los Q_Weighting

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

```
library(tidyr)
fit_svy_qwgt <- svyglm(Income ~ Expenditure,
                         design = diseno_qwgt)
modNul <- svyglm(Income ~ 1, design = diseno_qwgt)
s0 <- summary(modNul)
s1 <- summary(fit_svy_qwgt)
tidy(fit_svy_qwgt)
```

term	estimate	std.error	statistic	p.value
(Intercept)	109.123	68.7290	1.588	0.115
Expenditure	1.246	0.2014	6.188	0.000

Calculo del R^2

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

```
s1$dispersion
```

```
##      variance      SE
## [1,]    123018 20322
```

```
s0$dispersion
```

```
##      variance      SE
## [1,]    250708 50082
```

```
(R2 = 1-78053/148800)
```

```
## [1] 0.4755
```

```
n = sum(diseno_qwgt$variables$wk2)
(R2Adj = 1-((1-R2)*(n-1)/(n-1-1)))
```

Modelos empleando los Q_Weighting

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

Table 7: Comprando Modelos con Q Weighting

	svyglm(wgt)	svyglm(qwgt)
(Intercept)	103.136	109.123
	p = (0.114)	p = (0.115)
Expenditure	1.263	1.246
	p = (0.000)	p = (0.000)
Num.Obs.	2605	2605
R2	0.509	0.509
R2 Adj.	-9.826	-9.828
AIC	123.5	141.5
F	43.885	38.294
RMSE	345.09	344.96

Modelo escogido

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

```
diseno_qwgt %<-% mutate(Age2 = Age^2)
mod_svy <- svyglm(
  Income ~ Expenditure + Zone + Sex + Age2 ,
  design = diseno_qwgt)
s1 <- summary(mod_svy)
s0 <- summary(modNul)
stargazer(mod_svy, header = FALSE, single.row = T,
           title = "Modelo",
           style = "ajps", omit.stat=c("bic", "ll"))
```

Resumen del Modelo escogido

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

Table 8: Modelo

	Income
Expenditure	1.208*** (0.209)
ZoneUrban	67.050 (41.340)
SexMale	21.330 (15.820)
Age2	0.008 (0.006)
Constant	66.980 (66.210)
N	2605
AIC	38332.000

***p < .01; **p < .05; *p < .1

Calculo del R^2

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

```
s1$dispersion
```

```
##      variance      SE
## [1,] 121651 20098
```

```
s0$dispersion
```

```
##      variance      SE
## [1,] 250708 50082
```

```
(R2 = 1-76821/148800)
```

```
## [1] 0.4837
```

```
n = sum(diseno_qwgt$variables$wk2)
(R2Adj = 1-((1-R2)*(n-1)/(n-1-1)))
```

Residuales estandarizados

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

$$r_{pi} = \left(y_i - \mu_i (\hat{\boldsymbol{B}}_w) \right) \sqrt{\frac{w_i}{V(\hat{\mu}_i)}}$$

$$H = W^{1/2} X \left(X^T W X \right)^{-1} W^{1/2}$$

donde

$$W = \text{diag} \left\{ \frac{w_1}{V(\hat{\mu}_1) [g'(\mu_1)]^2}, \dots, \frac{w_n}{V(\hat{\mu}_1) [g'(\mu_n)]^2} \right\}$$

con g es una función de enlace que es especificada mediante un Modelo lineal generalizado.

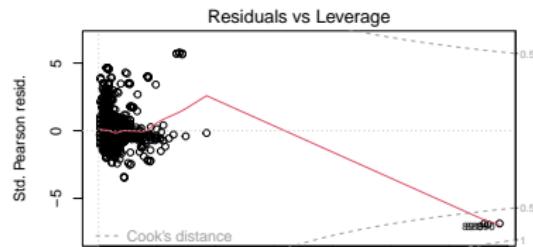
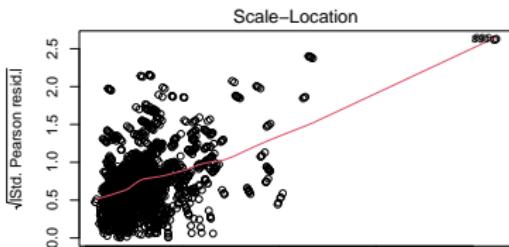
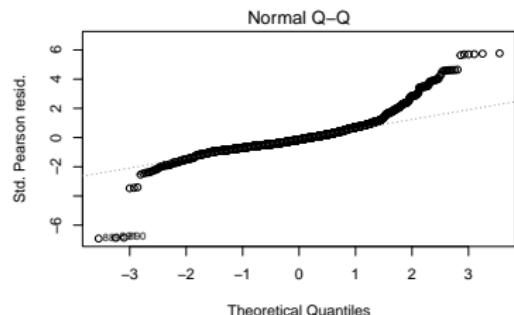
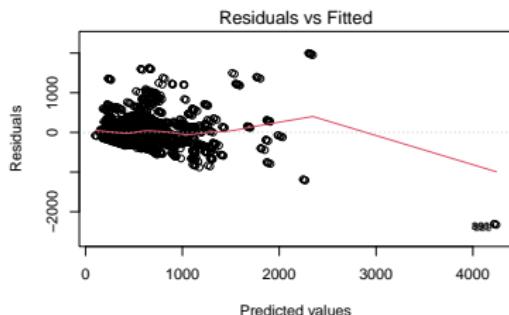
Diagnóstico del Modelo

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn Guerrero
M.Sc.

Procesando múltiples bases.

```
par(mfrow = c(2,2))
plot(mod_svy)
```



Pruebas de normalidad

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

- H_0 : Los errores proviene de una distribución normal.

Algunas librerías que podemos emplear son:

```
library(normtest) #REALIZA 5 PRUEBAS
library(nortest)  #REALIZA 10 PRUEBAS
library(moments) #REALIZA 1 PRUEBA
```

Extrayendo residuales standarizados

```
library(svystdiags)
stdresids = as.numeric(svystdres(mod_svy)$stdresids)
diseno_qwgt$variables %<-% mutate(stdresids = stdresi
```

Pruebas de normalidad

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

- H_0 : Los errores proviene de una distribución normal.
- H_1 : Los errores no proviene de una distribución normal.

Algunas librerías que podemos emplear son:

```
library(normtest) #REALIZA 5 PRUEBAS
library(nortest)  #REALIZA 10 PRUEBAS
library(moments) #REALIZA 1 PRUEBA
```

Extrayendo residuales standarizados

```
library(svystdiags)
stdresids = as.numeric(svystdres(mod_svy)$stdresids)
diseno_qwgt$variables %<-% mutate(stdresids = stdresi
```

Histograma de los residuales

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

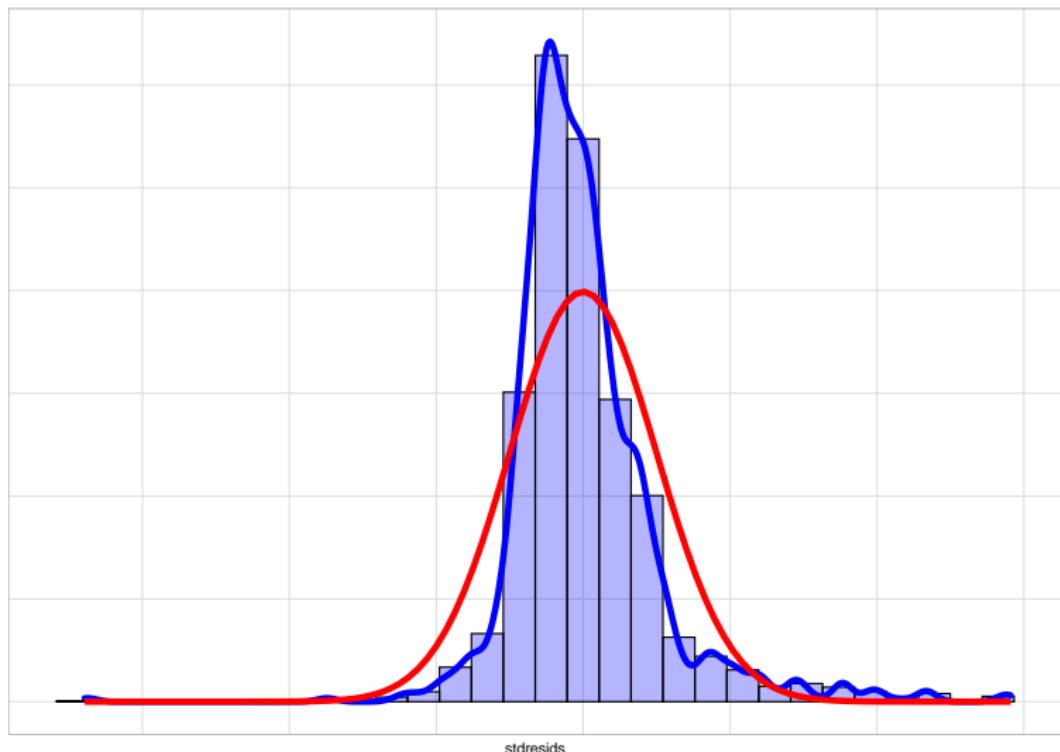
```
ggplot(data = diseno_qwgt$variables,  
       aes(x = stdresids)) +  
  geom_histogram(aes(y = ..density..), colour = "black",  
                 fill = "blue", alpha = 0.3) +  
  geom_density(size = 2, colour = "blue") +  
  geom_function(fun = dnorm, colour = "red", size = 2) +  
  theme_cepal() + labs(y = "")
```

Histograma de los residuales

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn Guerrero
M.Sc.

Procesando múltiples bases.



Pruebas de normalidad Kolmogorov-Smirnov

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

```
nortest::lillie.test(diseno_qwgt$variables$stdresids)

##  
##  Lilliefors (Kolmogorov-Smirnov) normality test  
##  
## data:  disen0_qwgt$variables$stdresids  
## D = 0.12, p-value <2e-16
```

Varianza constante

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

```
library(patchwork)
diseno_qwgt$variables %<>%
  mutate(pred = predict(mod_svy))
g2 <- ggplot(data = diseno_qwgt$variables,
              aes(x = Expenditure, y = stdresids))+
  geom_point() +
  geom_hline(yintercept = 0) + theme_cepal()
g3 <- ggplot(data = diseno_qwgt$variables,
              aes(x = Age2, y = stdresids))+
  geom_point() +
  geom_hline(yintercept = 0) + theme_cepal()
```

Varianza constante

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

```
g4 <- ggplot(data = diseno_qwgt$variables,  
              aes(x = Zone, y = stdresids))+  
      geom_point() +  
      geom_hline(yintercept = 0) + theme_cepal()  
g5 <- ggplot(data = diseno_qwgt$variables,  
              aes(x = Sex, y = stdresids))+  
      geom_point() + geom_hline(yintercept = 0) +  
      theme_cepal()
```

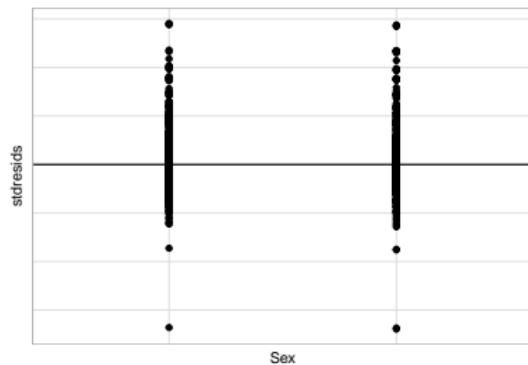
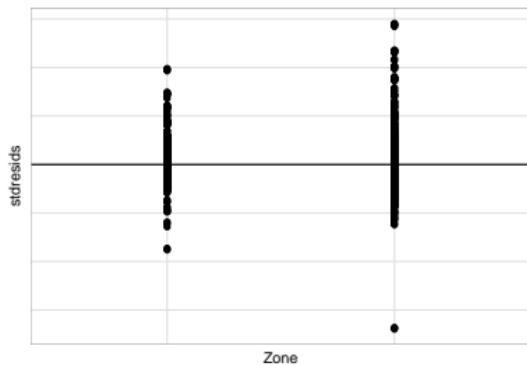
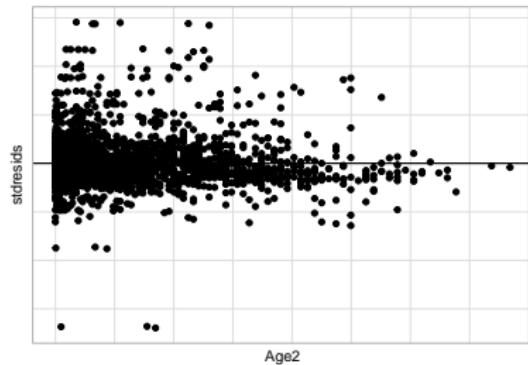
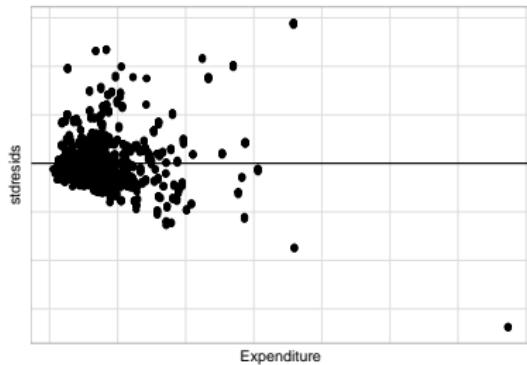
(g2|g3)/(g4|g5)

Varianza constante

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn Guerrero
M.Sc.

Procesando múltiples bases.



Distancia de cook

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

$$c_i = \frac{w_i^* w_i e_i^2}{p \phi V(\hat{\mu}_i) (1 - h_{ii})^2} \mathbf{x}_i^t \left[\widehat{\text{Var}} \left(U_w (\hat{\mathbf{B}}_w) \right) \right]^{-1} \mathbf{x}_i$$

donde,

- w_i^* = Pesos de la encuesta.

Distancia de cook

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

$$c_i = \frac{w_i^* w_i e_i^2}{p \phi V(\hat{\mu}_i) (1 - h_{ii})^2} \mathbf{x}_i^t \left[\widehat{\text{Var}} \left(U_w (\hat{\mathbf{B}}_w) \right) \right]^{-1} \mathbf{x}_i$$

donde,

- w_i^* = Pesos de la encuesta.
- w_i Elementos por fuera de la diagonal de la matriz hat

Distancia de cook

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn Guerrero
M.Sc.

Procesando múltiples bases.

$$c_i = \frac{w_i^* w_i e_i^2}{p \phi V(\hat{\mu}_i) (1 - h_{ii})^2} \mathbf{x}_i^t \left[\widehat{\text{Var}} \left(U_w (\hat{\mathbf{B}}_w) \right) \right]^{-1} \mathbf{x}_i$$

donde,

- w_i^* = Pesos de la encuesta.
- w_i Elementos por fuera de la diagonal de la matriz hat
- e_i = residuales

Distancia de cook

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

$$c_i = \frac{w_i^* w_i e_i^2}{p \phi V(\hat{\mu}_i) (1 - h_{ii})^2} \mathbf{x}_i^t \left[\widehat{\text{Var}} \left(U_w (\hat{\mathbf{B}}_w) \right) \right]^{-1} \mathbf{x}_i$$

donde,

- w_i^* = Pesos de la encuesta.
- w_i Elementos por fuera de la diagonal de la matriz hat
- e_i = residuales
- p = número de parámetros del Modelo de regresión.

Distancia de cook

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn Guerrero
M.Sc.

Procesando múltiples bases.

$$c_i = \frac{w_i^* w_i e_i^2}{p\phi V(\hat{\mu}_i)(1 - h_{ii})^2} \mathbf{x}_i^t \left[\widehat{\text{Var}}(U_w(\hat{\mathbf{B}}_w)) \right]^{-1} \mathbf{x}_i$$

donde,

- w_i^* = Pesos de la encuesta.
- w_i Elementos por fuera de la diagonal de la matriz hat
- e_i = residuales
- p = número de parámetros del Modelo de regresión.
- ϕ = parámetro de dispersión en el glm

Distancia de cook

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn Guerrero
M.Sc.

Procesando múltiples bases.

$$c_i = \frac{w_i^* w_i e_i^2}{p \phi V(\hat{\mu}_i) (1 - h_{ii})^2} \mathbf{x}_i^t \left[\widehat{\text{Var}}(U_w(\hat{\mathbf{B}}_w)) \right]^{-1} \mathbf{x}_i$$

donde,

- w_i^* = Pesos de la encuesta.
- w_i Elementos por fuera de la diagonal de la matriz hat
- e_i = residuales
- p = número de parámetros del Modelo de regresión.
- ϕ = parámetro de dispersión en el glm
- $\widehat{\text{Var}}(U_w(\hat{\mathbf{B}}_w))$ = estimación de varianza linealizada de la ecuación de puntuación, que se utiliza para pseudo MLE en Modelos lineales generalizados ajustados a datos de encuestas de muestras complejas

Distancia de cook

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

Una vez que se ha determinado el valor de la D de Cook para un elemento de muestra individual, se puede calcular la siguiente estadística de prueba para evaluar la importancia de la estadística D :

$$\frac{(df - p + 1) \times c_i}{df} \doteq F_{(p, df-p)}$$

donde $df =$ grados de libertad basados en el diseño.

Por otro lado, la literatura considera a las observaciones influyentes cuando sean mayores a 2 o 3

Detección de observaciones influyentes (Distancia de cook)

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn Guerrero
M.Sc.

Procesando múltiples bases.

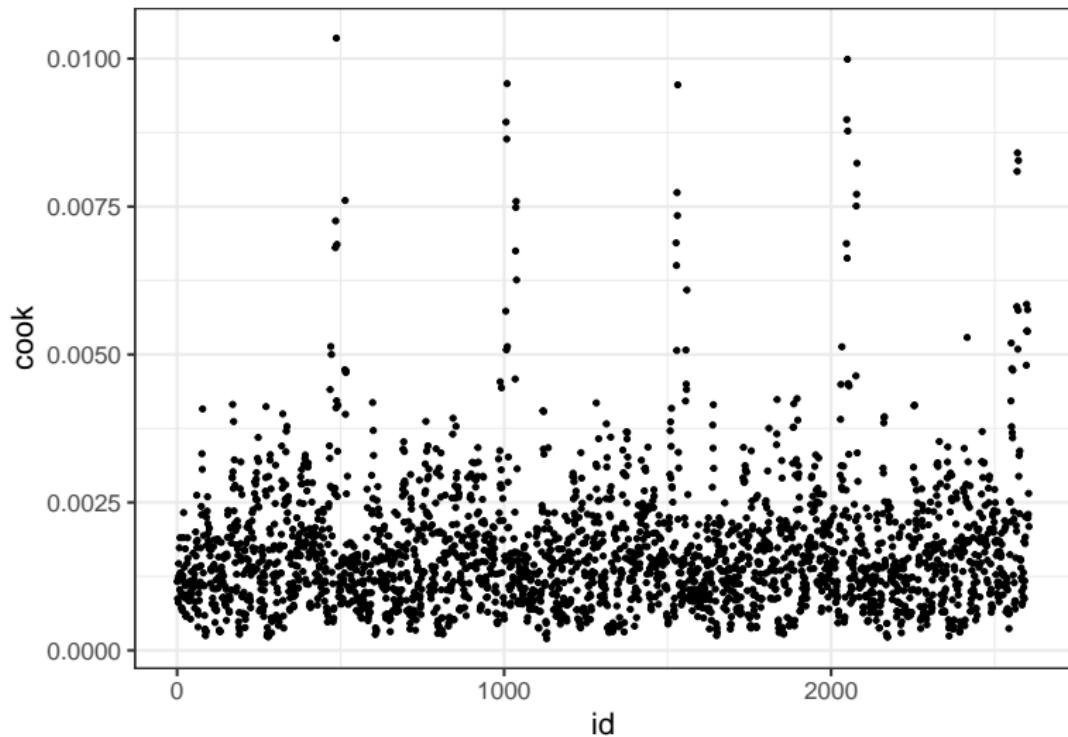
```
d_cook = data.frame(cook = svyCooksD(mod_svy),  
                     id = 1:length(svyCooksD(mod_svy)))  
  
ggplot(d_cook, aes(y = cook, x = id)) + geom_point()  
theme_bw(20)
```

Detección de observaciones influyentes (Distancia de cook)

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn Guerrero
M.Sc.

Procesando múltiples bases.



$D_f Beta_{(i)}$

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

$$D_f Beta_{(i)} = \hat{B} - \hat{B}_{(i)} = \frac{\mathbf{A}^{-1} \mathbf{X}_{(i)}^t \hat{e}_i w_i}{1 - h_{ii}}$$

Donde $\mathbf{A} = \mathbf{X}^t \mathbf{W} \mathbf{X}$ $\hat{B}_{(i)}$ es el vector de parámetros estimados una vez se ha eliminado la i-ésima observación, h_{ii} es el correspondiente elemento de la diagonal de H y \hat{e}_i es el residual de la i-ésima observación.

$D_f Beta_{(i)}$

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

$$D_f Beta_{(i)} = \frac{c_{ji} e_i / (1 - h_{ii})}{\sqrt{v(\hat{B}_j)}}$$

donde:

- c_{ji} = es el j -í-estimo elemento de $\mathbf{A}^{-1} w_i^2 \mathbf{X}_{(i)} \mathbf{X}_{(i)}^t \mathbf{A}^{-1}$

Procesando
múltiples
bases.

$D_f Beta_{(i)}$

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

$$D_f Beta_{(i)} = \frac{c_{ji} e_i / (1 - h_{ii})}{\sqrt{v(\hat{B}_j)}}$$

donde:

- c_{ji} = es el j -í-estimo elemento de $\mathbf{A}^{-1} w_i^2 \mathbf{X}_{(i)} \mathbf{X}_{(i)}^t \mathbf{A}^{-1}$
- El estimador de $v(\hat{B}_j)$ basado en el Modelo se obtiene como: $v_m(\hat{B}_j) = \hat{\sigma} \sum_{i=1}^n c_{ji}^2$ con $\hat{\sigma} = \sum_{i \in s} w_i e^2 / (\hat{N} - p)$
y $\hat{N} = \sum_{i \in s} w_i$

Procesando
múltiples
bases.

$D_f Beta_{(i)}$

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn Guerrero
M.Sc.

$$D_f Beta_{(i)} = \frac{c_{ji} e_i / (1 - h_{ii})}{\sqrt{v(\hat{B}_j)}}$$

donde:

- c_{ji} = es el j -ésimo elemento de $\mathbf{A}^{-1} w_i^2 \mathbf{X}_{(i)} \mathbf{X}_{(i)}^t \mathbf{A}^{-1}$
- El estimador de $v(\hat{B}_j)$ basado en el Modelo se obtiene como: $v_m(\hat{B}_j) = \hat{\sigma} \sum_{i=1}^n c_{ji}^2$ con $\hat{\sigma} = \sum_{i \in s} w_i e^2 / (\hat{N} - p)$ y $\hat{N} = \sum_{i \in s} w_i$
- La i -ésima observación es influyente para B_j si $|D_f Beta_{(i)}| \geq \frac{z}{\sqrt{n}}$ con $z = 2$ o 3

Procesando múltiples bases.

$D_f Beta_{(i)}$

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn Guerrero
M.Sc.

$$D_f Beta_{(i)} = \frac{c_{ji} e_i / (1 - h_{ii})}{\sqrt{v(\hat{B}_j)}}$$

donde:

- c_{ji} = es el j -ésimo elemento de $\mathbf{A}^{-1} w_i^2 \mathbf{X}_{(i)} \mathbf{X}_{(i)}^t \mathbf{A}^{-1}$
- El estimador de $v(\hat{B}_j)$ basado en el Modelo se obtiene como: $v_m(\hat{B}_j) = \hat{\sigma} \sum_{i=1}^n c_{ji}^2$ con $\hat{\sigma} = \sum_{i \in s} w_i e^2 / (\hat{N} - p)$ y $\hat{N} = \sum_{i \in s} w_i$
- La i -ésima observación es influyente para B_j si $|D_f Beta_{(i)}| \geq \frac{z}{\sqrt{n}}$ con $z = 2$ o 3
- Como alternativa puede usar $t_{0.025, n-p} / \sqrt{(n)}$ donde $t_{0.025, n-p}$ es el percentil 97.5

Procesando múltiples bases.

Detección de observaciones influyentes $(D_f Beta_{(i)j})$

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

```
d_dfbetas = data.frame(t(svydfbetas(mod_svy)$Dfbetas)
colnames(d_dfbetas) <- paste0("Beta_", 1:5)
d_dfbetas %>% slice(1:10L)
```

Beta_1	Beta_2	Beta_3	Beta_4	Beta_5
0.0005	-2e-04	0.0020	-0.0045	-0.0075
-0.0005	-1e-04	0.0013	0.0026	-0.0030
-0.0008	-1e-04	0.0008	0.0022	0.0008
-0.0004	-1e-04	0.0011	-0.0031	0.0007
-0.0008	0e+00	0.0007	0.0021	0.0014
0.0009	5e-04	-0.0036	-0.0063	0.0097
0.0026	4e-04	-0.0031	-0.0076	-0.0027
0.0011	3e-04	-0.0028	0.0077	-0.0042
0.0029	3e-04	-0.0030	-0.0078	-0.0050
-0.0003	4e-04	0.0012	-0.0037	-0.0040

Detección de observaciones influyentes (D_f Betas_{(i)j})

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

```
d_dfbetas$id <- 1:nrow(d_dfbetas)
d_dfbetas <- reshape2::melt(d_dfbetas, id.vars = "id"
cutoff <- svydfbetas(mod_svy)$cutoff
d_dfbetas %<>%
  mutate(
    Criterio = ifelse(abs(value) > cutoff, "Si", "No"))

tex_label <- d_dfbetas %>%
  filter(Criterio == "Si") %>%
  arrange(desc(abs(value))) %>%
  slice(1:10L)
```

Detección de observaciones influyentes (D_f Betas_{(i)j})

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

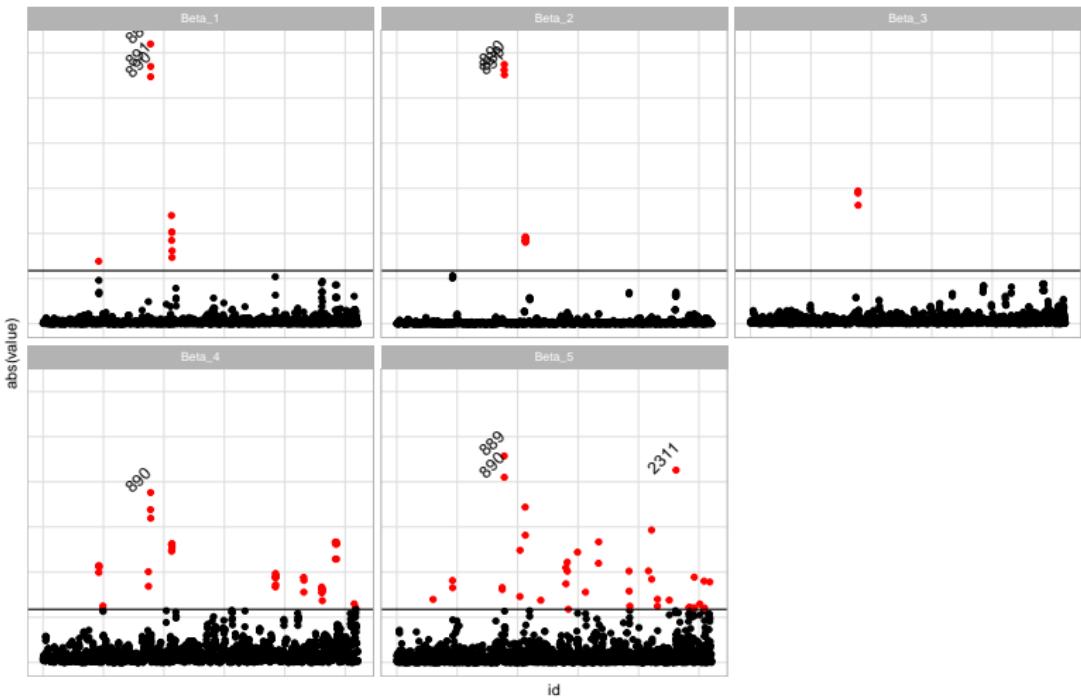
```
ggplot(d_dfbetas, aes(y = abs(value), x = id)) +  
  geom_point(aes(col = Criterio)) +  
  geom_text(data = tex_label,  
            angle = 45,  
            vjust = -1,  
            aes(label = id)) +  
  geom_hline(aes(yintercept = cutoff)) +  
  facet_wrap(. ~ variable, nrow = 2) +  
  scale_color_manual(  
    values = c("Si" = "red", "No" = "black")) +  
  theme_cepal()
```

Detección de observaciones influyentes ($D_f Beta_{(i)j}$)

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn Guerrero
M.Sc.

Procesando múltiples bases.



Criterion • Si • No

Matriz H asociada al PMLE

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

- La matriz asociada al Estimador de Pseudo Máxima Verosimilitud (PMLE) de $\hat{\boldsymbol{B}}$ es $\boldsymbol{H} = \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^{-t}\mathbf{W}$ cuya diagonal esta dado por $h_{ii} = \mathbf{x}_i^t \mathbf{A}^{-1} \mathbf{x}_i^{-t} w_i$.

Matriz H asociada al PMLE

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

- La matriz asociada al Estimador de Pseudo Máxima Verosimilitud (PMLE) de $\hat{\boldsymbol{B}}$ es $\boldsymbol{H} = \boldsymbol{X}\boldsymbol{A}^{-1}\boldsymbol{X}^{-t}\boldsymbol{W}$ cuya diagonal esta dado por $h_{ii} = \boldsymbol{x}_i^t\boldsymbol{A}^{-1}\boldsymbol{x}_i^{-t}w_i$.
- Una observación puede ser grande y, como resultado, influir en las predicciones, cuando un x_i es considerablemente diferente del promedio ponderado $\bar{x}_w = \sum_{i \in s} w_i \boldsymbol{x}_i / \sum_{i \in s} w_i$.

Detección de observaciones influyentes (h_{ii})

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

```
vec_hat <- svyhat(mod_svy, doplot = FALSE)
d_hat = data.frame(hat = vec_hat,
                    id = 1:length(vec_hat))
d_hat %>% mutate(
  C_cutoff = ifelse(hat > (3 * mean(hat)), "Si", "No"))

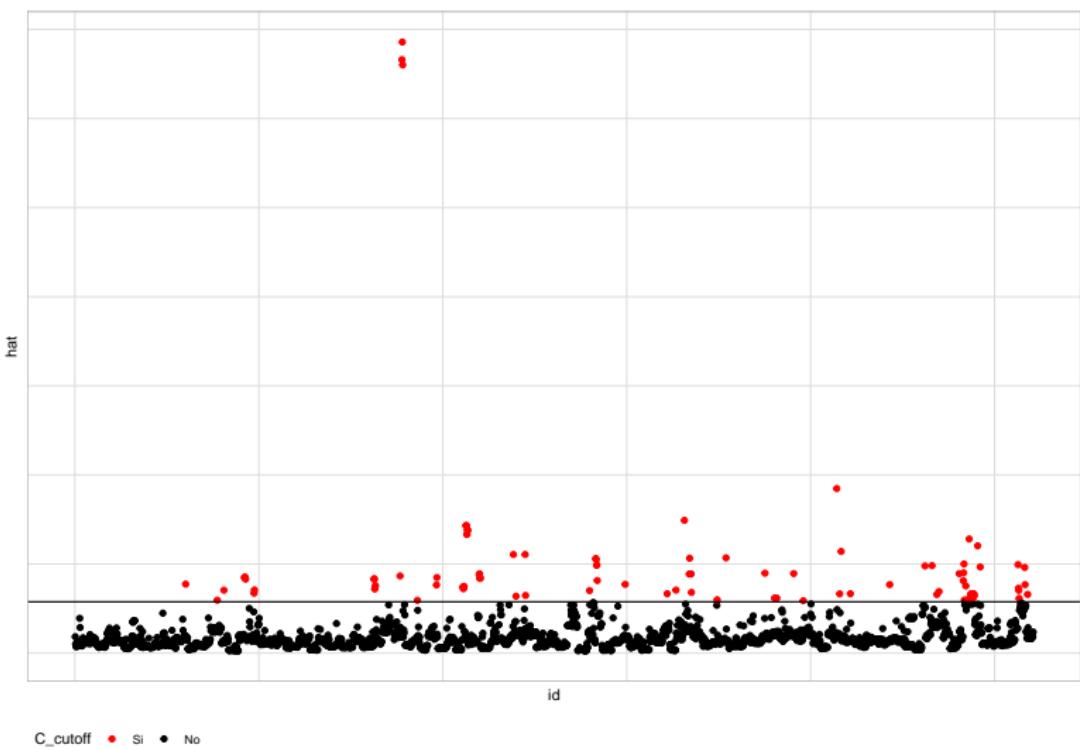
ggplot(d_hat, aes(y = hat, x = id)) +
  geom_point(aes(col = C_cutoff)) +
  geom_hline(yintercept = (3 * mean(d_hat$hat))) +
  scale_color_manual(
    values = c("Si" = "red", "No" = "black"))+
  theme_cepal()
```

Detección de observaciones influyentes (h_{ii})

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn Guerrero
M.Sc.

Procesando múltiples bases.



Estadístico $D_f Fits_{(i)}$

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

$$D_f Fits_{(i)} = \frac{h_{ii} e_i / (1 - h_{ii})}{\sqrt{v(\hat{\beta}_j)}}$$

Donde, $\sqrt{v(\hat{\beta}_j)}$ puede ser aproximada por el diseño o el Modelo. La i-ésima observación se considera influyente en el ajuste del Modelo si $|DfFits(i)| \geq z \sqrt{\frac{p}{n}}$ con $z = 2$ o 3

Detección de observaciones influyentes (D_f $Fits_{(i)}$)

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn Guerrero
M.Sc.

Procesando múltiples bases.

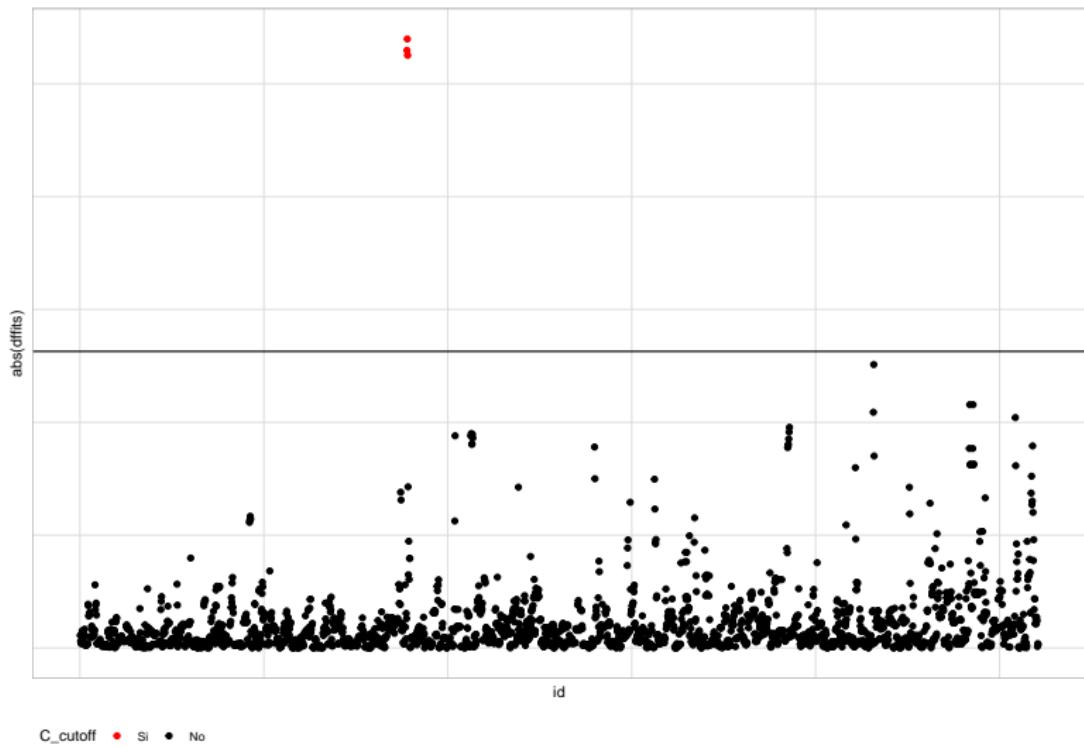
```
d_dffits = data.frame(  
  dffits = svydffits(mod_svy)$Dffits,  
  id = 1:length(svydffits(mod_svy)$Dffits))  
  
cutoff <- svydffits(mod_svy)$cutoff  
  
d_dffits %<-% mutate(  
  C_cutoff = ifelse(abs(dffits) > cutoff, "Si", "No")  
  ggplot(d_dffits, aes(y = abs(dffits), x = id)) +  
    geom_point(aes(col = C_cutoff)) +  
    geom_hline(yintercept = cutoff) +  
    scale_color_manual(  
      values = c("Si" = "red", "No" = "black"))+  
    theme_cepal()
```

Detección de observaciones influyentes (D_f $Fits_{(i)}$)

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn Guerrero
M.Sc.

Procesando múltiples bases.



Inferencia sobre los parámetros del Modelo

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

$$t = \frac{\hat{\beta}_k - \beta_k}{se(\hat{\beta}_k)} \sim t_{n-p}$$

$$\hat{B} \pm t_{(1-\frac{\alpha}{2}, df)} \times se(\hat{B})$$

Estimación del dato

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando múltiples bases.

$$\hat{E}(y_i | \mathbf{x}_{obs,i}) = \mathbf{x}_{obs,i} \hat{\beta}$$

$$\hat{E}(y_i | \mathbf{x}_{obs,i}) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}$$

term	estimate	std.error	statistic	p.value
(Intercept)	66.9825	66.2141	1.012	0.3139
Expenditure	1.2082	0.2093	5.774	0.0000
ZoneUrban	67.0536	41.3446	1.622	0.1076
SexMale	21.3320	15.8214	1.348	0.1802
Age2	0.0085	0.0057	1.491	0.1386

$$\hat{E}(y_i | \mathbf{x}_{obs,i}) = 91.6319 + 1.0893x_{1i} + 48.7667x_{2i} + 8.0933x_{3i} + 0.0115$$

Estimación del dato

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

(Intercept)	Expenditure	ZoneUrban	SexMale	Age2
1	346.3	0	1	4624
1	346.3	0	0	3136
1	346.3	0	0	576
1	346.3	0	1	676
1	346.3	0	0	9
1	392.2	0	0	3721
1	392.2	0	0	529

$$\hat{y}_i = 91.6319 + 1.0893(247.9) + 48.7667(0) + 8.0933(1) + 0.0115(55^2) = 404.6$$

Estimando el IC de predicción

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn Guerrero
M.Sc.

Procesando múltiples bases.

$$\text{var} \left(\hat{E} (y_i | \mathbf{x}_{obs,i}) \right) = \mathbf{x}_{obs,i}^t \text{cov} (\boldsymbol{\beta}) \mathbf{x}_{obs,i}$$

`vcov(mod_svy)`

	(Intercept)	Expenditure	ZoneUrban	SexMale	Age2
(Intercept)	4384.3129	-12.6820	462.7656	306.4189	-0.2019
Expenditure	-12.6820	0.0438	-4.1970	-0.4944	0.0006
ZoneUrban	462.7656	-4.1970	1709.3771	-131.0058	-0.0629
SexMale	306.4189	-0.4944	-131.0058	250.3161	-0.0036
Age2	-0.2019	0.0006	-0.0629	-0.0036	0.0000

Estimando el IC de predicción

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

```
xobs <- model.matrix(mod_svy) %>%
  data.frame() %>% slice(1) %>% as.matrix()

cov_beta <- vcov(mod_svy) %>% as.matrix()

as.numeric(sqrt((xobs) %*% cov_beta %*% t(xobs)))

## [1] 44.46
```

Intervalo de confianza para la predicción

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

$$\mathbf{x}_{obs,i} \hat{\beta} \pm t_{\left(1 - \frac{\alpha}{2}, n-p\right)} \sqrt{var(\hat{E}(y_i | \mathbf{x}_{obs,i}))}$$

Utilizando la función predict

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

```
pred <- data.frame(predict(mod_svy, type = "link"))
pred_IC <- data.frame(
  confint(predict(mod_svy, type = "link")))
colnames(pred_IC) <- c("Lim_Inf", "Lim_Sup")
pred <- bind_cols(pred, pred_IC)
pred$Expenditure <- encuesta$Expenditure
pred %>% slice(1:6L)
```

Utilizando la función predict

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

link	SE	Lim_Inf	Lim_Sup	Expenditure
545.9	44.46	458.8	633.0	346.3
512.0	33.67	446.0	578.0	346.3
490.3	29.26	433.0	547.7	346.3
512.5	37.08	439.8	585.2	346.3
485.5	29.19	428.3	542.7	346.3
572.4	42.08	489.9	654.8	392.2

Scaterplot de la predicción

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

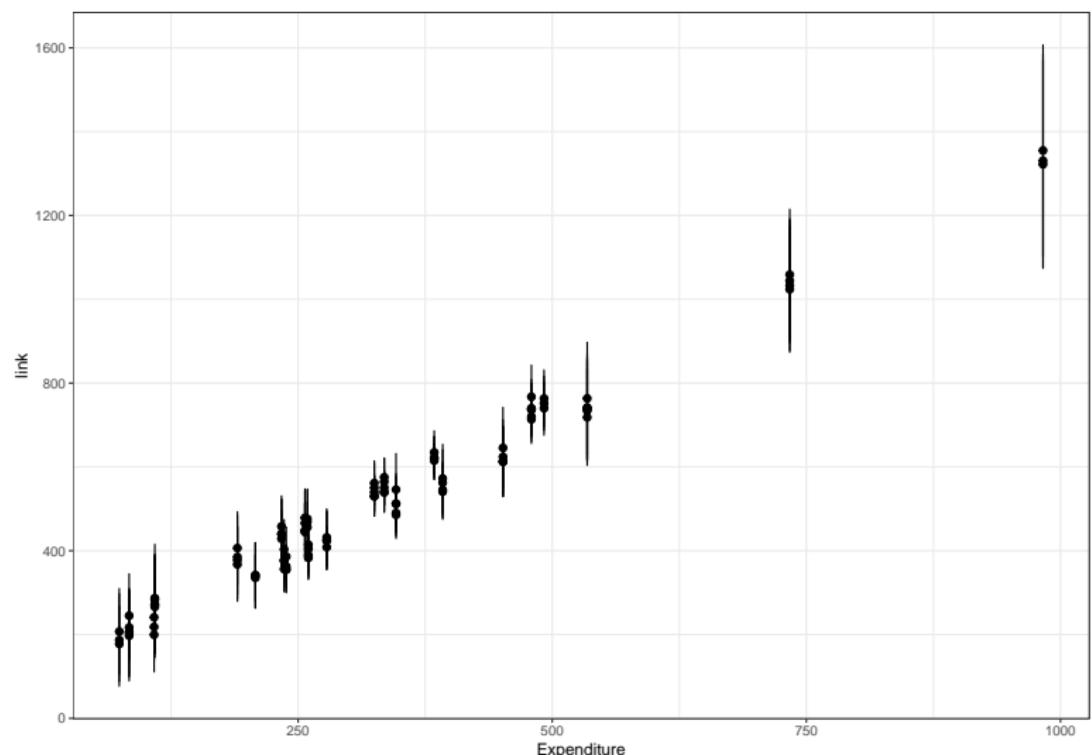
```
pd <- position_dodge(width = 0.2)
ggplot(pred %>% slice(1:100L),
       aes(x = Expenditure , y = link)) +
  geom_errorbar(aes(ymin = Lim_Inf,
                     ymax = Lim_Sup),
                width = .1,
                linetype = 1) +
  geom_point(size = 2, position = pd) +
  theme_bw()
```

Scaterplot de la predicción

Análisis de encuestas de hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.



Predictión fuera de las observaciones.

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn Guerrero
M.Sc.

Procesando múltiples bases.

```
datos_nuevos <- data.frame(Expenditure = 1600,
                             Age2 = 40^2, Sex = "Male",
                             Zone = "Urban")
```

$$\hat{y}_i = 91.6319 + 1.0893(1600) + 48.7667(0) + 8.0933(1) + 0.0115(40^2) = 1910$$

$$var \left(\hat{E} (y_i | \mathbf{x}_{obs,i}) \right) = \mathbf{x}_{obs,i}^t cov (\beta) \mathbf{x}_{obs,i} + \hat{\sigma}_{yx}^2$$

```
x_noObs = matrix(c(1, 1600, 1, 1, 40^2), nrow = 1)
as.numeric(sqrt(x_noObs %*% cov_beta %*% t(x_noObs)))
```

```
## [1] 257.4
```

Intervalo de confianza para la predicción

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

$$\mathbf{x}_{obs,i} \hat{\beta} \pm t_{\left(1 - \frac{\alpha}{2}, n-p\right)} \sqrt{var \left(\hat{E} (y_i | \mathbf{x}_{obs,i}) \right) + \hat{\sigma}_{yx}^2}$$

Predictión fuera de las observaciones.

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

```
predict(mod_svy, newdata = datos_nuevos, type = "lin  
##      link    SE  
## 1 2102 257  
  
confint(predict(mod_svy,newdata = datos_nuevos))
```

2.5 %	97.5 %
1598	2607

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

Procesando múltiples bases.

Lectura de múltiples bases

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

Para realizar la lectura de múltiples bases debemos conocer las rutas donde estas estan guardadas para ello empleamos la función `file.list` del paquete base, que nos permite tener un listado completo de los archivos.

```
(data_path <- list.files("Z:/BC/", full.names = TRUE,  
                         pattern = "2020") %>%  
  tibble(path = .) %>%  
  mutate( pais = gsub("Z:\\\\BC\\\\(.*)_.*", "\\\\1",  
                     x = path)))
```

Note que utiliza la función `gsub` para separar el nombre del país de la ruta.

Lectura de múltiples bases

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

path	país
Z:/BC/ARG_2020N.dta	ARG
Z:/BC/BOL_2020N.dta	BOL
Z:/BC/BRA_2020N1.dta	BRA
Z:/BC/CHL_2020N.dta	CHL
Z:/BC/COL_2020N1.dta	COL
Z:/BC/CRI_2020N1.dta	CRI
Z:/BC/DOM_2020N1.dta	DOM
Z:/BC/ECU_2020N.dta	ECU
Z:/BC/MEX_2020N1.dta	MEX
Z:/BC/PER_2020N.dta	PER
Z:/BC/PRY_2020N.dta	PRY
Z:/BC/SLV_2020N.dta	SLV
Z:/BC/URY_2020N.dta	URY

Lectura de encuestas

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

Para la lectura de los archivos, se procede de la siguiente forma.

```
require(purrr)
require(haven)
data_path %<>%
  mutate(encuesta = path %>% map(~read_dta(.x) %>%
    transmute(upm = `_upm`,
              estrato =`_estrato`,
              sexo, areageo2,lp,li,ingcorte,
              fep =`_fep`)))
```

La función `map` es utilizada para trabajar con los elementos de una lista. Las variables seleccionadas son sexo, área geográfica (`areageo2`), Linea de pobreza (`lp`), Linea de indigencia (`li`), Ingreso persona (`ingcorte`) y factor de expansión por persona (`fep`).

Lectura de encuestas (resultado)

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

	path <i><chr></i>	país <i><chr></i>	encuesta <i><list></i>
1	Z:/BC/ARG_2020N.dta	ARG	<i><tibble [43,767 x 6]></i>
2	Z:/BC/BOL_2020N.dta	BOL	<i><tibble [37,092 x 6]></i>
3	Z:/BC/BRA_2020N1.dta	BRA	<i><tibble [355,436 x 6]></i>
4	Z:/BC/CHL_2020N.dta	CHL	<i><tibble [185,437 x 6]></i>
5	Z:/BC/COL_2020N1.dta	COL	<i><tibble [747,822 x 6]></i>
6	Z:/BC/CRI_2020N1.dta	CRI	<i><tibble [25,530 x 6]></i>
7	Z:/BC/DOM_2020N1.dta	DOM	<i><tibble [71,378 x 6]></i>
8	Z:/BC/ECU_2020N.dta	ECU	<i><tibble [30,646 x 6]></i>
9	Z:/BC/MEX_2020N1.dta	MEX	<i><tibble [315,743 x 6]></i>
10	Z:/BC/PER_2020N.dta	PER	<i><tibble [120,346 x 6]></i>
11	Z:/BC/PRY_2020N.dta	PRY	<i><tibble [17,582 x 6]></i>
12	Z:/BC/SLV_2020N.dta	SLV	<i><tibble [37,030 x 6]></i>
13	Z:/BC/URY_2020N.dta	URY	<i><tibble [145,166 x 6]></i>

El resultado es objeto tipo **tibble** el cual permite observar de forma compacta el contenido de una lista e indica el tipo y tamaño de cada objeto en la contenido en la lista.

Definir el diseño

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

Ahora se debe definir un diseño para cada encuesta, para nuestro ejemplo se define el diseño muestral.

```
options(survey.lonely.psu="adjust")
data_path %<>% mutate(
  diseño = encuesta %>%
    map(~as_survey_design(.data = .x,
      ids = upm,
      strata = estrato,
      weights = fep,
      nest = T
    )))

```

Definir el diseño (resultado)

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando múltiples bases.

```
# A tibble: 13 x 4
  path          pais encuesta      diseno
  <chr>        <chr> <list>        <list>
1 Z:/BC/ARG_2020N.dta ARG <tibble [43,767 x 6]> <tbl_svy[,6]>
2 Z:/BC/BOL_2020N.dta BOL <tibble [37,092 x 6]> <tbl_svy[,6]>
3 Z:/BC/BRA_2020N1.dta BRA <tibble [355,436 x 6]> <tbl_svy[,6]>
4 Z:/BC/CHL_2020N.dta CHL <tibble [185,437 x 6]> <tbl_svy[,6]>
5 Z:/BC/COL_2020N1.dta COL <tibble [747,822 x 6]> <tbl_svy[,6]>
6 Z:/BC/CRI_2020N1.dta CRI <tibble [25,530 x 6]> <tbl_svy[,6]>
7 Z:/BC/DOM_2020N1.dta DOM <tibble [71,378 x 6]> <tbl_svy[,6]>
8 Z:/BC/ECU_2020N.dta ECU <tibble [30,646 x 6]> <tbl_svy[,6]>
9 Z:/BC/MEX_2020N1.dta MEX <tibble [315,743 x 6]> <tbl_svy[,6]>
10 Z:/BC/PER_2020N.dta PER <tibble [120,346 x 6]> <tbl_svy[,6]>
11 Z:/BC/PRY_2020N.dta PRY <tibble [17,582 x 6]> <tbl_svy[,6]>
12 Z:/BC/SLV_2020N.dta SLV <tibble [37,030 x 6]> <tbl_svy[,6]>
13 Z:/BC/URY_2020N.dta URY <tibble [145,166 x 6]> <tbl_svy[,6]>
```

Estimación de las personas debajo de la linea de pobreza en multiples encuestas

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando múltiples bases.

```
library(tidyr)
data_path[1:4,] %>% mutate(
  model = map(diseno,
    ~svyglm(ingcorte~sexo+areageo2,.x) %>%
      coef() %>%
      data.frame(estimado = .) %>%
      tibble::rownames_to_column(var = "Coef")
  dplyr::select(pais,model) %>% unnest(model)
```

Estimación de las personas con ingresos de bajo de la linea de pobreza en multiples encuestas (Resultado)

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn Guerrero
M.Sc.

Procesando múltiples bases.

pais	Coef	estimado
ARG	(Intercept)	2.326e+04
ARG	sexo	-9.794e+00
BOL	(Intercept)	2.925e+03
BOL	sexo	-1.193e+02
BOL	areageo2	-9.089e+02
BRA	(Intercept)	2.618e+03
BRA	sexo	-8.258e+01
BRA	areageo2	-8.552e+02
CHL	(Intercept)	5.765e+05
CHL	sexo	-3.308e+04
CHL	areageo2	-1.167e+05

Alternativa para el procesamiento de múltiples archivos.

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

En ocasiones solo se desea obtener un resultado rápido para realizar un reporte o una comparación rápida de información, en estas ocasiones no es necesario guardar en la memoria de R toda la encuesta, por esta razón se ilustra una alternativa de procesamiento de múltiples archivos.

■ **Paso 1** Leer archivo y organizar encuestas.

Alternativa para el procesamiento de múltiples archivos.

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

En ocasiones solo se desea obtener un resultado rápido para realizar un reporte o una comparación rápida de información, en estas ocasiones no es necesario guardar en la memoria de R toda la encuesta, por esta razón se ilustra una alternativa de procesamiento de múltiples archivos.

- **Paso 1** Leer archivo y organizar encuestas.
- **Paso 2** Definir diseño muestral.

Alternativa para el procesamiento de múltiples archivos.

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

En ocasiones solo se desea obtener un resultado rápido para realizar un reporte o una comparación rápida de información, en estas ocasiones no es necesario guardar en la memoria de R toda la encuesta, por esta razón se ilustra una alternativa de procesamiento de múltiples archivos.

- **Paso 1** Leer archivo y organizar encuestas.
- **Paso 2** Definir diseño muestral.
- **Paso 3** Procesar información.

Alternativa para el procesamiento de múltiples archivos.

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

En ocasiones solo se desea obtener un resultado rápido para realizar un reporte o una comparación rápida de información, en estas ocasiones no es necesario guardar en la memoria de R toda la encuesta, por esta razón se ilustra una alternativa de procesamiento de múltiples archivos.

- **Paso 1** Leer archivo y organizar encuestas.
- **Paso 2** Definir diseño muestral.
- **Paso 3** Procesar información.
- **Paso 4** Organizar y presentar resultados.

Creando función para el procesamiento de múltiples archivos.

Análisis de encuestas de hogares con R

Andrés Gutiérrez,
Ph.D.
Stalyn Guerrero
M.Sc.

Procesando múltiples bases.

```
options(survey.lonely.psu="adjust")
model_aux <- function(input_file){
  ## Paso 1
  encuesta <- read_dta(input_file) %>%
    transmute(upm = `_upm`, estrato =`estrato`,
              sexo, areageo2,lp,li,ingcorte, fep =`_fep`)
  ## Paso 2
  diseno <- as_survey_design(.data = encuesta,
                               ids = upm,
                               strata = estrato,
                               weights = fep,
                               nest = T)
  ## Paso 3
  s <- svyglm(ingcorte~sexo+areageo2,diseno) %>% summary()
  s$coefficients %>%
    data.frame() %>%
    tibble::rownames_to_column(var = "Coef")
}
```

Procesando encuestas múltiples

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

Para el *Paso 4* realizamos la siguiente sintaxis.

```
list.files("Z:/BC/", full.names = TRUE,  
          pattern = "2020") [1:3] %>%  
  map_df(~model_aux(.x))
```

Los resultados se muestran en el orden de lectura de los archivos

Coef	Estimate	Std..Error	t.value	Pr...t..
(Intercept)	23259.139	0.00	1.891e+14	NaN
sexo	-9.794	0.00	-8.409e+12	NaN
(Intercept)	2924.814	95.81	3.053e+01	0
sexo	-119.293	18.58	-6.421e+00	0
areageo2	-908.884	63.35	-1.435e+01	0
(Intercept)	2618.067	57.75	4.533e+01	0
sexo	-82.575	11.62	-7.105e+00	0
areageo2	-855.212	26.95	-3.173e+01	0

¡Gracias!

Análisis de
encuestas de
hogares con R

Andrés
Gutiérrez,
Ph.D.
Stalyn
Guerrero
M.Sc.

Procesando
múltiples
bases.

Email: andres.gutierrez@cepal.org