Análisis de encuestas de hogares con R

Modulo 6: Modelos lineales generalizados

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

CEPAL - Unidad de Estadísticas Sociales

Análisis de encuestas de hogares con R

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

Lectura de la base


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

Definir diseño de la muestra con `srvyr`


```
library(srvyr)

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

definir nuevas variables


```
diseno <- diseno %>% mutate(
  pobreza = ifelse(Poverty != "NotPoor", 1, 0),
  desempleo = ifelse(Employment == "Unemployed", 1, 0
```

Tablas de doble entrada para el tamaño


```
library(survey)
(tab_pobreza_sexo <- svyby(~factor(pobreza), ~Sex,
      FUN = svytotal, design = as.svrepdesign(diseno)
      se=F, na.rm=T, ci=T, keep.var=TRUE))
```

	Sex	factor(pobreza)0	factor(pobreza)1	se1	
Female	Female	48366	30824	2411	29
Male	Male	43032	28044	2522	30

```
(tab <- svytable(~pobreza + Sex, design = diseno))
```

pobreza/Sex	Female	Male
0	48366	43032
1	30824	28044

Tablas de doble entrada para el proporción


```
(tab_pobreza_sexo <- svyby(~factor(pobreza), ~Sex,
      FUN = svymean, design = as.svrepdesign(diseno),
      se=F, na.rm=T, ci=T, keep.var=TRUE))
```

	Sex	factor(pobreza)0	factor(pobreza)1	se1	se2
Female	Female	0.6108	0.3892	0.0316	0.0316
Male	Male	0.6054	0.3946	0.0366	0.0366

```
prop.table(tab, margin = 2)
```

pobreza/Sex	Female	Male
0	0.6108	0.6054
1	0.3892	0.3946

Prueba de independencia F

$$\hat{\pi}_{rc} = \frac{n_{r+}}{n_{++}} \times \frac{n_{+c}}{n_{++}}$$

$$\chi^2_{pearsom} = n_{++} \times \sum_r \sum_c \left(\frac{(p_{rc} - \hat{\pi}_{rc})^2}{\hat{\pi}_{rc}} \right)$$

$$G^2 = 2 \times n_{++} \times \sum_r \sum_c p_{cr} \times \ln \left(\frac{p_{rc}}{\hat{\pi}_{rc}} \right)$$

donde, R es el número de filas y C representa el número de columnas, la prueba tiene $(R-1) \times (C-1)$ grados de libertad.

Prueba de independencia F

$$\chi^2_{(R-S)} = \chi^2_{(Pearson)} / GDEFF$$

$$G^2_{(R-S)} = G^2 / GDEFF$$

con $GDEFF$ el efecto generalizado del diseño, esta dado por

$$GDEFF = \frac{\sum_r \sum_c (1 - p_{rc})\, d^2(p_{rc}) - \sum_r (1 - p_{r+})\, d^2(p_{r+}) - \sum_c (1 - p_{+c})\, d$$
$$(R-1)(C-1)$$

Prueba de independencia F

$$F_{R-S,Pearson} = \chi^2_{R-S} / \left[(R-1)(C-1) \right] \sim F_{(R-1)(C-1),(R-1)(C-1)df}$$

$$F_{R-S,LRT} = G^2_{R-S} / (C-1) \sim F_{(C-1),df}$$

donde C es el número de columnas de la tabla cruzada

Prueba de independencia ChiSq


```
summary(tab, statistic = "Chisq")
```

```
##          Sex
## pobreza Female  Male
##       0  48366 43032
##       1  30824 28044
##
##  Pearson's X^2: Rao & Scott adjustment
##
## data:  NextMethod()
## X-squared = 0.077, df = 1, p-value = 0.8
```

Prueba de independencia F


```
summary(tab, statistic = "F")
```

```
##          Sex
## pobreza Female  Male
##       0  48366 43032
##       1  30824 28044
##
##  Pearson's X^2: Rao & Scott adjustment
##
## data:  NextMethod()
## F = 0.056, ndf = 1, ddf = 119, p-value = 0.8
```

Estadístico de Wald

$$Q_{wald} = \hat{\boldsymbol{Y}}^t \left(\boldsymbol{H} \hat{\boldsymbol{V}} \left(\hat{\boldsymbol{N}} \right) \boldsymbol{H}^t \right)^{-1} \hat{\boldsymbol{Y}}$$

donde,

$$\hat{\boldsymbol{Y}} = \left(\hat{N} - E \right)$$

es un vector de $R \times C$ de diferencias entre los recuentos de celdas observadas y esperadas, esto es, $\hat{N}_{rc} - E_{rc}$

La matriz $\boldsymbol{H} \hat{\boldsymbol{V}} \left(\hat{\boldsymbol{N}} \right) \boldsymbol{H}^t$, representa la matriz de varianza-covarianza estimada para el vector de diferencias.

Estadístico de Wald

La matriz $\boldsymbol{H}$ es la inversa de la matriz $\boldsymbol{J}$ dada por:

$$\boldsymbol{J} = -\left[\frac{\delta^2 \ln PL(\boldsymbol{B})}{\delta^2 \boldsymbol{B}}\right] \mid \boldsymbol{B} = \hat{\boldsymbol{B}}$$

$$\sum_h \sum_a \sum_i x_{hai}^t x_{hai} w_{hai} \hat{\pi}_{hai}(\boldsymbol{B})(1 - \hat{\pi}_{hai}(\boldsymbol{B}))$$

Bajo la hipótesis nula, el estadístico

$$Q_{wald} \sim \chi^2_{(R-1)\times(C-1)}$$

Estadístico de Wald

$$F_{wald} = Q_{wald} \times \frac{df - (R-1)(C-1) + 1}{(R-1)(C-1)\, df} \sim F_{(R-1)(C-1),\, df-(R-1)(C-1)+1}$$

Prueba de independencia Wald


```
summary(tab, statistic = "Wald")
```

```
##          Sex
## pobreza Female  Male
##       0  48366 43032
##       1  30824 28044
##
##  Design-based Wald test of association
##
## data:  NextMethod()
## F = 0.056, ndf = 1, ddf = 119, p-value = 0.8
```

Prueba de independencia adjWald


```
summary(tab, statistic = "adjWald")
```

```
##          Sex
## pobreza Female  Male
##       0  48366 43032
##       1  30824 28044
##
##  Design-based Wald test of association
##
## data:  NextMethod()
## F = 0.056, ndf = 1, ddf = 119, p-value = 0.8
```

Prueba de independencia lincom


```
summary(tab, statistic = "lincom")
```

```
##          Sex
## pobreza Female  Male
##       0  48366 43032
##       1  30824 28044
##
##  Pearson's X^2: asymptotic exact distribution
##
## data:  NextMethod()
## X-squared = 0.077, p-value = 0.8
```

Prueba de independencia saddlepoint


```
summary(tab, statistic = "saddlepoint")
```

```
##          Sex
## pobreza Female  Male
##       0  48366 43032
##       1  30824 28044
##
##  Pearson's X^2: saddlepoint approximation
##
## data:  NextMethod()
## X-squared = 0.077, p-value = 0.8
```

Modelo log lineal para tablas de contingencia

$$\log(p_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY},$$

donde:

- $p_{ijk} =$ la proporción esperada en la celda bajo el modelo.

Modelo log lineal para tablas de contingencia

$$\log(p_{ijk}) = \mu + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY},$$

donde:

- p_{ijk} = la proporción esperada en la celda bajo el modelo.
- $\mu = \log(p_0) = \frac{1}{\#\ de\ celdas}$

Modelo log lineal para tablas de contingencia


```
mod1 <- svyloglin(~pobreza+Sex + pobreza:Sex , diseno
(s1 <- summary(mod1))
```

```
## Loglinear model: svyloglin(~pobreza + Sex + pobrez
##                       coef      se        p
## pobreza1          0.219673 0.06778 0.001192
## Sex1              0.052843 0.01625 0.001145
## pobreza1:Sex1     0.005583 0.02350 0.812175
```

Modelo log lineal para tablas de contingencia


```
mod2 <- svyloglin(~pobreza+Sex, diseno)
(s2 <- summary(mod2))
```

```
## Loglinear model: svyloglin(~pobreza + Sex, diseno)
##                 coef      se          p
## pobreza1 0.21997 0.06752 0.0011230
## Sex1     0.05405 0.01577 0.0006076
```

Modelo log lineal para tablas de contingencia


```
anova(mod1, mod2)
```

```
## Analysis of Deviance Table
##  Model 1: y ~ pobreza + Sex
## Model 2: y ~ pobreza + Sex + pobreza:Sex
## Deviance= 0.07719 p= 0.8126
## Score= 0.07719 p= 0.8126
```

Modelo de regresión logistica

$$g(\pi(x)) = logit(\pi(x)) = z = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = B_0 + B_1 x_1 + \cdots + B_p x$$

$$\hat{\pi}(\boldsymbol{x}) = \frac{\exp\left(\boldsymbol{X}\hat{\boldsymbol{B}}\right)}{1-\exp\left(\boldsymbol{X}\hat{\boldsymbol{B}}\right)} = \frac{\exp\left(\hat{B}_0 + \hat{B}_1 x_1 + \cdots + \hat{B} x_p\right)}{1-\exp\left(\hat{B}_0 + \hat{B}_1 x_1 + \cdots + \hat{B} x_p\right)}$$

$$PL(\boldsymbol{B} \mid X) = \prod_{i=1}^{n} \left\{ \pi(x_i)^{y_i} \left(1-\pi(x_i)\right)^{1-y_i} \right\}^{w_i}$$

con

$$\pi(x_i) = \frac{\exp(x_i \boldsymbol{B})}{1-\exp(x_i \boldsymbol{B})}$$

$$var\left(\hat{\boldsymbol{B}}\right) = \boldsymbol{J}^{-1} var\left(S\left(\hat{\boldsymbol{B}}\right)\right) \boldsymbol{J}^{-1}$$

Modelo de regresión logistica

$$S(B) = \sum_{h} \sum_{a} \sum_{i} w_{hai} \boldsymbol{D}_{hai}^{t} \left[\left(\pi_{hai} \left(\boldsymbol{B} \right) \right) \left(1 - \pi_{hai} \left(\boldsymbol{B} \right) \right) \right]^{-1} \left(y_{hai} - \pi_{hai} \left(\boldsymbol{B} \right) \right) = 0$$

$$D_{hai} = \frac{\delta \left(\pi_{hai} \left(\boldsymbol{B} \right) \right)}{\delta B_j}$$

$j = 0, \ldots, p$

Prueba de Wald para los parámetros del modelo

$$G = -2 \ln \left[\frac{L\left(\hat{\beta}_{MLE}\right)_{reduced}}{L\left(\hat{\beta}_{MLE}\right)_{full}} \right]$$

$$\hat{\psi} = \exp\left(\hat{B}_1\right)$$

$$CI\left(\psi\right) = \exp\left(\hat{B}_j \pm t_{df,1-\frac{\alpha}{2}} se\left(\hat{B}_j\right)\right)$$

Tablas de contingencia

Sex	pobreza	se	ci_l	ci_u
Female	0.3892	0.0316	0.3273	0.4512
Male	0.3946	0.0366	0.3228	0.4664

Zone	pobreza	se	ci_l	ci_u
Rural	0.4485	0.0561	0.3386	0.5585
Urban	0.3394	0.0320	0.2766	0.4022

Region	pobreza	se	ci_l	ci_u
Norte	0.3590	0.0555	0.2502	0.4677
Sur	0.3438	0.0435	0.2586	0.4291
Centro	0.3654	0.0786	0.2113	0.5195
Occidente	0.4008	0.0467	0.3092	0.4924
Oriente	0.4518	0.0886	0.2781	0.6255

Prueba de independencia ChiSq


```
## Pearson's X^2: Rao & Scott adjustment
pobreza_sex <-  svychisq(
  formula = ~pobreza + Sex, design = diseno)
tidy( pobreza_sex) %>% select(-method)
```

ndf	ddf	statistic	p.value
1	119	0.0565	0.8126

Prueba de independencia ChiSq


```
pobreza_Zona <- svychisq(
  formula = ~pobreza + Zone, design = diseno)
tidy(pobreza_Zona) %>% select(-method)
```

ndf	ddf	statistic	p.value
1	119	2.954	0.0883

Prueba de independencia ChiSq


```
pobreza_Region <- svychisq(
  formula = ~pobreza + Region, design = diseno)
   tidy(pobreza_Region) %>% select(-method)
```

ndf	ddf	statistic	p.value
3.008	358	0.4879	0.6914

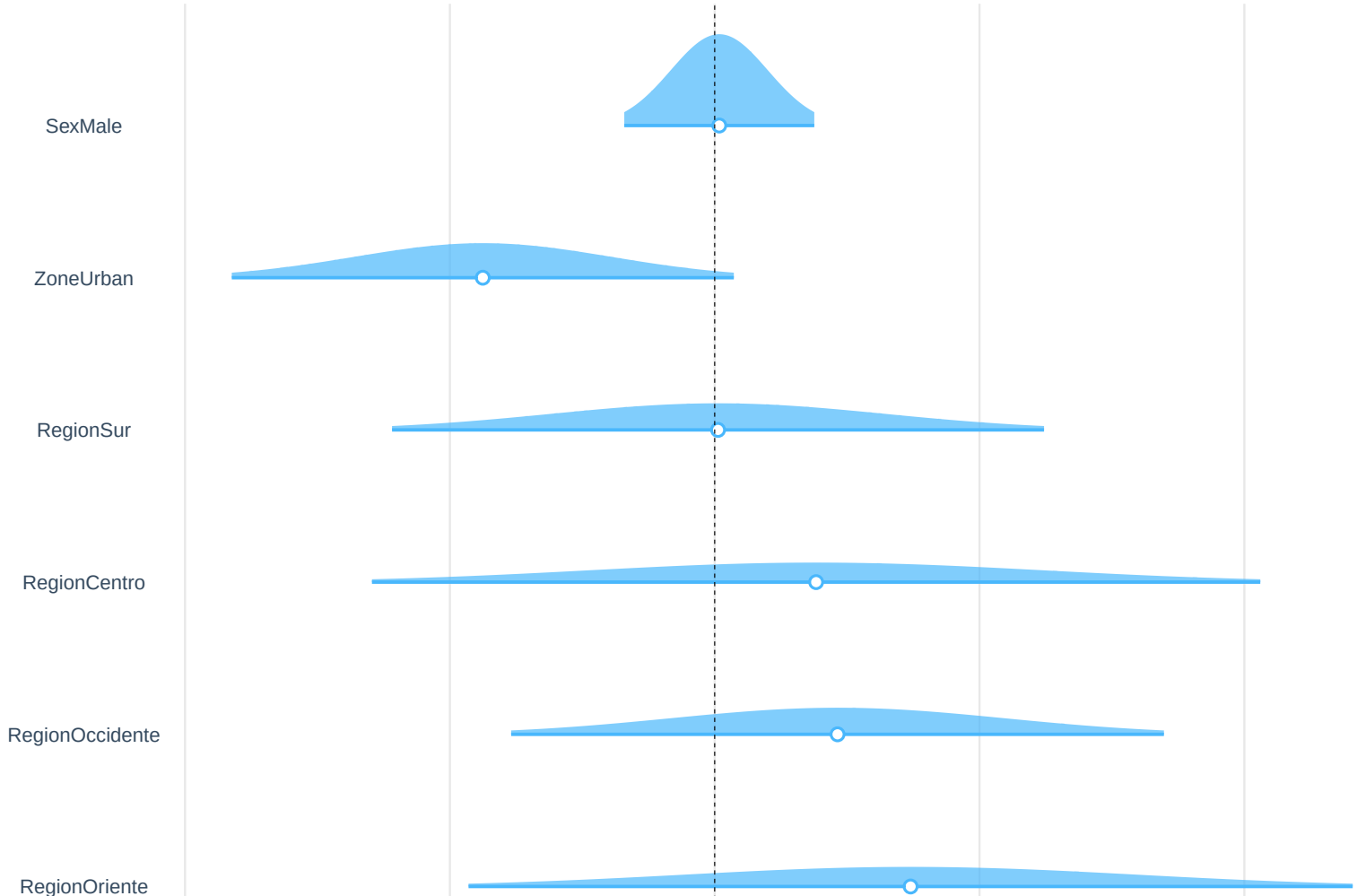
Modelo log lineal ajustado


```
mod_loglin <- svyglm(
  pobreza ~ Sex + Zone + Region,
   family=quasibinomial, design=diseno)
tidy(mod_loglin)
```

term	estimate	std.error	statistic	p.value
(Intercept)	-0.4082	0.2640	-1.5464	0.1248
SexMale	0.0086	0.0915	0.0945	0.9249
ZoneUrban	-0.4378	0.2418	-1.8106	0.0729
RegionSur	0.0063	0.3140	0.0201	0.9840
RegionCentro	0.1915	0.4279	0.4476	0.6553
RegionOccidente	0.2319	0.3144	0.7377	0.4622
RegionOriente	0.3699	0.4259	0.8686	0.3869

Plot de la distribución de los betas


```
plot_summs(mod_loglin,
           scale = TRUE, plot.distributions = TRUE)
```



Modelo log lineal ajustado


```
bind_cols(
data.frame(exp_estimado = exp(coef(mod_loglin))),
as.data.frame(exp(confint(mod_loglin)))
)
```

	exp_estimado	2.5 %	97.5 %
(Intercept)	0.6648	0.3941	1.122
SexMale	1.0087	0.8414	1.209
ZoneUrban	0.6454	0.3997	1.042
RegionSur	1.0063	0.5402	1.875
RegionCentro	1.2111	0.5188	2.827
RegionOccidente	1.2611	0.6764	2.351
RegionOriente	1.4476	0.6226	3.366

Estadístico de Wald sobre los parámetros


```
regTermTest(model = mod_loglin, ~Sex)
```

```
## Wald test for Sex
##  in svyglm(formula = pobreza ~ Sex + Zone + Region
##     family = quasibinomial)
## F =  0.00893  on  1  and  113  df: p= 0.92
```

```
regTermTest(model = mod_loglin, ~Zone)
```

```
## Wald test for Zone
##  in svyglm(formula = pobreza ~ Sex + Zone + Region
##     family = quasibinomial)
## F =  3.278  on  1  and  113  df: p= 0.073
```

```
regTermTest(model = mod_loglin, ~Region)
```

```
## Wald test for Region
```

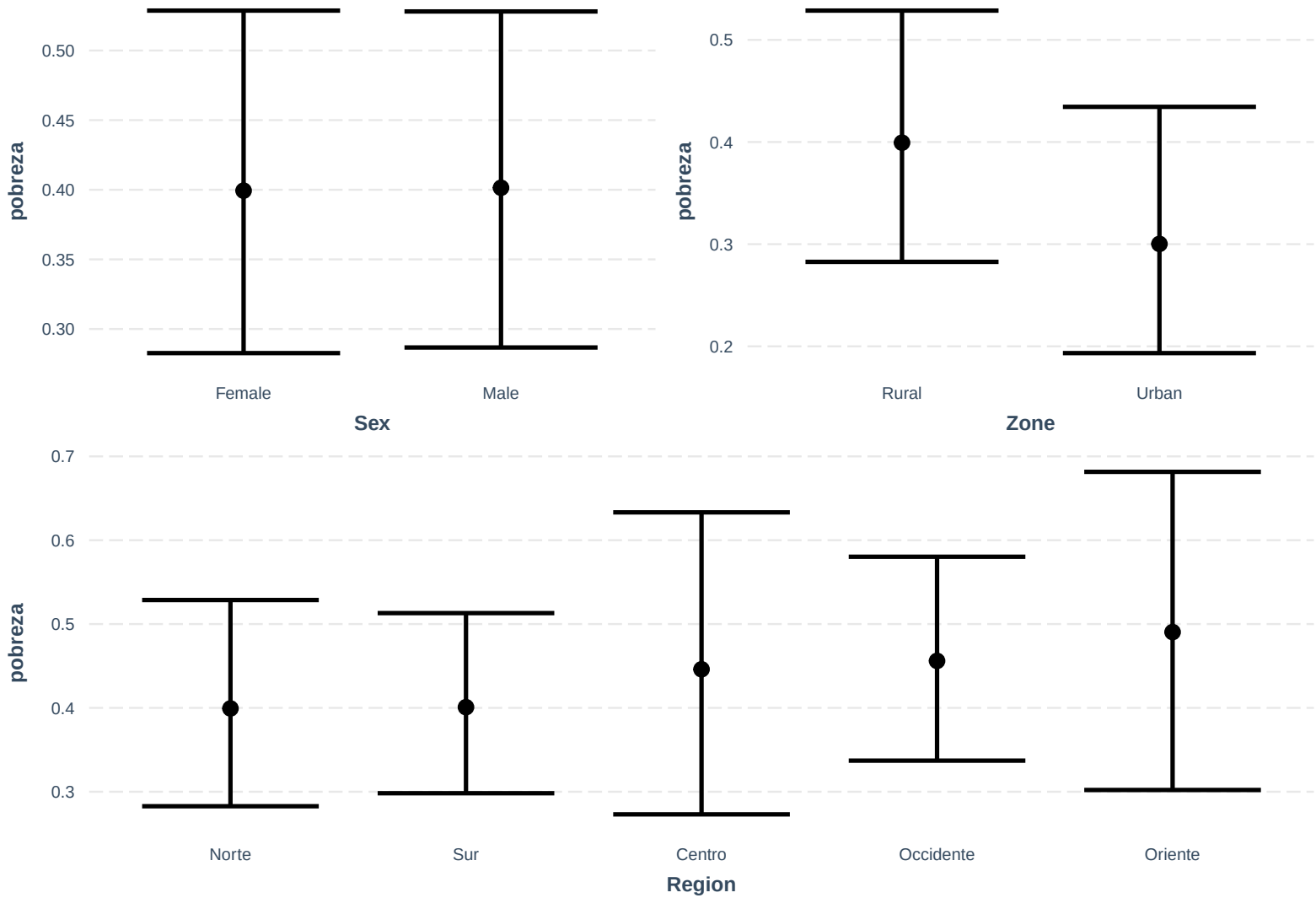
Efecto del modelo.


```
effe_sex <- effect_plot(mod_loglin, pred = Sex,
                        interval = TRUE)
effe_Zona <-effect_plot(mod_loglin, pred = Zone,
                        interval = TRUE)
effe_Region <- effect_plot(mod_loglin, pred = Region,
                           interval = TRUE)
(effe_sex |effe_Zona)/effe_Region
```

Efecto del modelo.

Análisis de encuestas de hogares con R

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



Modelo log lineal ajustado con interacciones


```
mod_loglin_int <- svyglm(pobreza ~ Sex + Zone + Regio
                                   Sex:Zone + Sex:Region,
                         family=quasibinomial, design=dis
tab_mod <- tidy(mod_loglin_int) %>% arrange(p.value)
tab_mod %>% slice(1:6)
```

term	estimate	std.error	statistic	p.valu
ZoneUrban	-0.4248	0.2562	-1.6580	0.100
(Intercept)	-0.4289	0.2849	-1.5055	0.135
SexMale:RegionSur	0.2871	0.2774	1.0348	0.303
RegionOriente	0.3843	0.4279	0.8980	0.371
RegionOccidente	0.3342	0.3783	0.8835	0.379
SexMale:RegionOccidente	-0.2302	0.2868	-0.8026	0.424

Modelo log lineal ajustado con interacciones

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

```
tab_mod %>% slice(7:12)
```

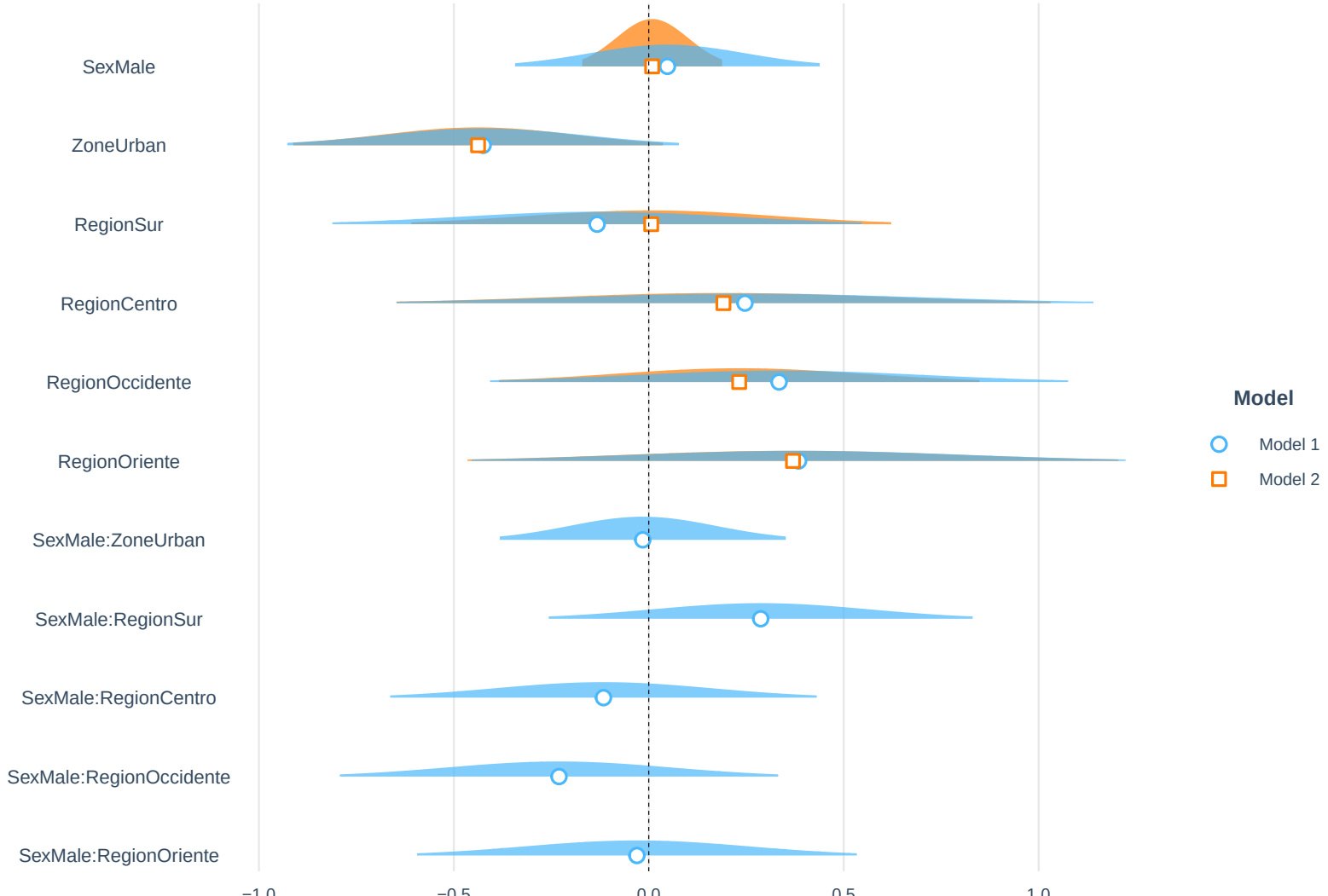
term	estimate	std.error	statistic	p.value
RegionCentro	0.2466	0.4560	0.5408	0.5897
SexMale:RegionCentro	-0.1162	0.2791	-0.4162	0.6781
RegionSur	-0.1325	0.3464	-0.3825	0.7028
SexMale	0.0478	0.1994	0.2399	0.8109
SexMale:RegionOriente	-0.0304	0.2878	-0.1057	0.9161
SexMale:ZoneUrban	-0.0154	0.1872	-0.0824	0.9345

Plot de la distribución de los betas

Análisis de encuestas de hogares con R

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

```
plot_summs(mod_loglin_int, mod_loglin, scale = TRUE,
```



Modelo log lineal ajustado

	exp_estimado	2.5 %	97.5 %
(Intercept)	0.6512	0.3702	1.145
SexMale	1.0490	0.7065	1.557
ZoneUrban	0.6539	0.3935	1.087
RegionSur	0.8759	0.4408	1.740
RegionCentro	1.2797	0.5183	3.160
RegionOccidente	1.3968	0.6599	2.957
RegionOriente	1.4685	0.6288	3.430
SexMale:ZoneUrban	0.9847	0.6795	1.427
SexMale:RegionSur	1.3325	0.7689	2.309
SexMale:RegionCentro	0.8903	0.5120	1.548
SexMale:RegionOccidente	0.7944	0.4499	1.403
SexMale:RegionOriente	0.9701	0.5484	1.716

Estadístico de Wald sobre los parámetros


```
regTermTest(model = mod_loglin_int, ~Sex)
```

```
## Wald test for Sex
##  in svyglm(formula = pobreza ~ Sex + Zone + Region
##     design = diseno, family = quasibinomial)
## F =  0.05753  on  1  and  108  df: p= 0.81
```

```
regTermTest(model = mod_loglin_int, ~Zone)
```

```
## Wald test for Zone
##  in svyglm(formula = pobreza ~ Sex + Zone + Region
##     design = diseno, family = quasibinomial)
## F =  2.749  on  1  and  108  df: p= 0.1
```

```
regTermTest(model = mod_loglin_int, ~Region)
```

```
## Wald test for Region
```

Estadístico de Wald sobre los parámetros


```
regTermTest(model = mod_loglin_int, ~Sex:Zone)
```

```
## Wald test for Sex:Zone
##  in svyglm(formula = pobreza ~ Sex + Zone + Region
##     design = diseno, family = quasibinomial)
## F =  0.006789  on  1  and  108  df: p= 0.93
```

```
regTermTest(model = mod_loglin_int, ~Sex:Region)
```

```
## Wald test for Sex:Region
##  in svyglm(formula = pobreza ~ Sex + Zone + Region
##     design = diseno, family = quasibinomial)
## F =  1.058  on  4  and  108  df: p= 0.38
```

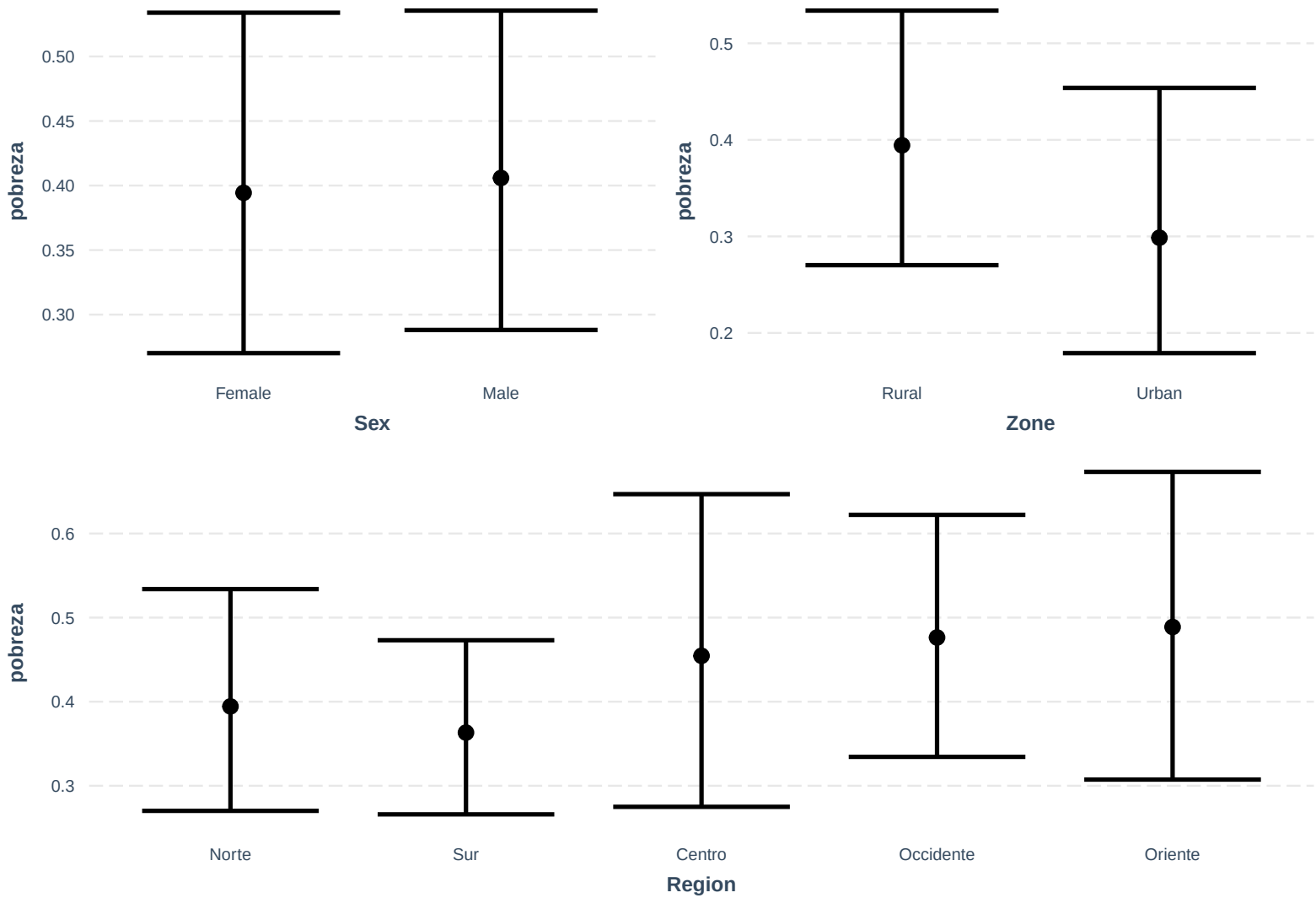
Efecto del modelo.


```
effe_sex <- effect_plot(mod_loglin_int, pred = Sex,
                        interval = TRUE)
effe_Zona <-effect_plot(mod_loglin_int, pred = Zone,
                        interval = TRUE)
effe_Region <- effect_plot(mod_loglin_int, pred = Reg
                           interval = TRUE)
(effe_sex |effe_Zona)/effe_Region
```

Efecto del modelo.

Análisis de encuestas de hogares con R

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



Modelo log lineal ajustado con Q_Weighting


```
fit_wgt <- lm(wk ~  Sex + Zone + Region , data = encu
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
  )
```

Modelo log lineal ajustado con Q_Weighting


```
diseno_qwgt <- diseno_qwgt %>% mutate(
  pobreza = ifelse(Poverty != "NotPoor", 1, 0))

mod_loglin_qwgt <- svyglm(pobreza ~ Sex + Zone + Regi
                          family=quasibinomial,
                          design=diseno_qwgt)
(tab_mod <- tidy(mod_loglin_qwgt) )
```

Modelo log lineal ajustado con Q_Weighting

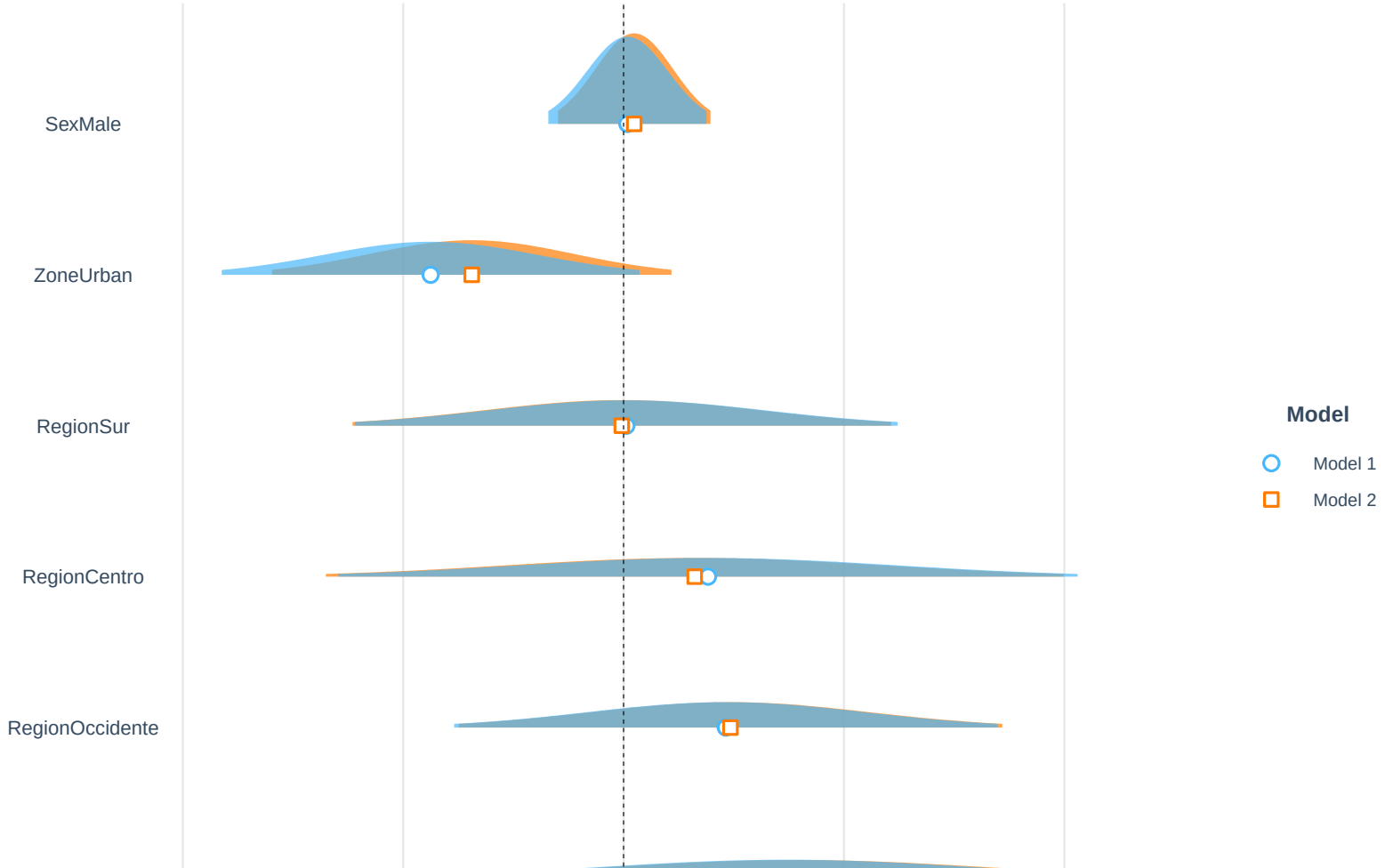
term	estimate	std.error	statistic	p.value
(Intercept)	-0.4644	0.2630	-1.7656	0.0802
SexMale	0.0241	0.0883	0.2726	0.7857
ZoneUrban	-0.3445	0.2311	-1.4903	0.1389
RegionSur	-0.0041	0.3116	-0.0130	0.9896
RegionCentro	0.1613	0.4270	0.3778	0.7063
RegionOccidente	0.2424	0.3147	0.7705	0.4426
RegionOriente	0.3937	0.4319	0.9115	0.3639

Plot de la distribución de los betas

Análisis de encuestas de hogares con R

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

```
plot_summs(mod_loglin, mod_loglin_qwgt,
           scale = TRUE, plot.distributions = TRUE)
```



Modelo log lineal ajustado

	exp_estimado	2.5 %	97.5 %
(Intercept)	0.6285	0.3732	1.058
SexMale	1.0244	0.8600	1.220
ZoneUrban	0.7086	0.4482	1.120
RegionSur	0.9960	0.5371	1.847
RegionCentro	1.1750	0.5043	2.738
RegionOccidente	1.2744	0.6832	2.377
RegionOriente	1.4824	0.6301	3.488

Estadístico de Wald sobre los parámetros


```
regTermTest(model = mod_loglin_qwgt, ~Sex)
```

```
## Wald test for Sex
##  in svyglm(formula = pobreza ~ Sex + Zone + Region
##     family = quasibinomial)
## F =  0.0743  on  1  and  113  df: p= 0.79
```

```
regTermTest(model = mod_loglin_qwgt, ~Zone)
```

```
## Wald test for Zone
##  in svyglm(formula = pobreza ~ Sex + Zone + Region
##     family = quasibinomial)
## F =  2.221  on  1  and  113  df: p= 0.14
```

```
regTermTest(model = mod_loglin_qwgt, ~Region)
```

```
## Wald test for Region
```

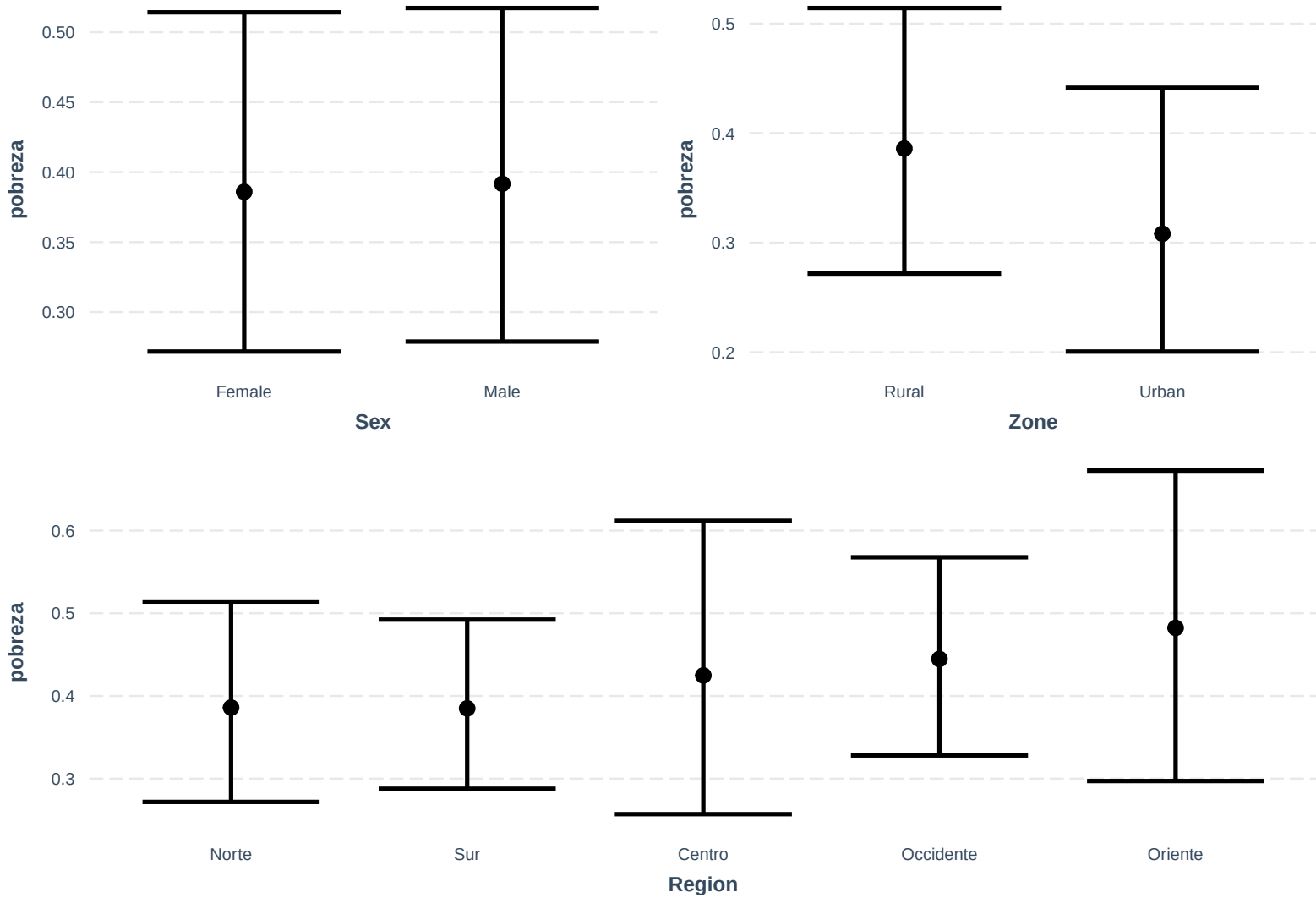
Efecto del modelo.


```
effe_sex <- effect_plot(mod_loglin_qwgt, pred = Sex,
                        interval = TRUE)
effe_Zona <-effect_plot(mod_loglin_qwgt, pred = Zone,
                        interval = TRUE)
effe_Region <- effect_plot(mod_loglin_qwgt, pred = Re
                           interval = TRUE)
(effe_sex |effe_Zona)/effe_Region
```

Efecto del modelo.

Análisis de encuestas de hogares con R

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



¡Gracias!

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

Email: andres.gutierrez@cepal.org