Análisis de encuestas de hogares con R

Modulo 8: Modelos multinivel

CEPAL - Unidad de Estadísticas Sociales

Lectura de la base


```r
encuesta <- readRDS("../Data/encuesta.rds")
```

Creando theme_cepal


```
theme_cepal <- function(...) theme_light(10) +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank(),
        axis.text.y = element_blank(),
        axis.ticks.y = element_blank(),
        legend.position="bottom",
        legend.justification = "left",
        legend.direction="horizontal",
        plot.title = element_text(size = 20, hjust =
        ...)
```

Introducción a los modelos multinivel.

Seleccionando una muestra de seis estratos

```
encuesta_plot <- encuesta %>%
  dplyr::select(HHID, Stratum) %>% unique() %>%
  group_by(Stratum)  %>% tally() %>%
  arrange(desc(n)) %>% dplyr::select(-n) %>%
  slice(1:6L) %>%
  inner_join(encuesta) %>% filter(Expenditure <700) %
  dplyr::select(Income, Expenditure, Stratum,
         Sex, Region, Zone)
encuesta_plot  %>% slice(1:10L)
```

Introducción a los modelos multinivel.

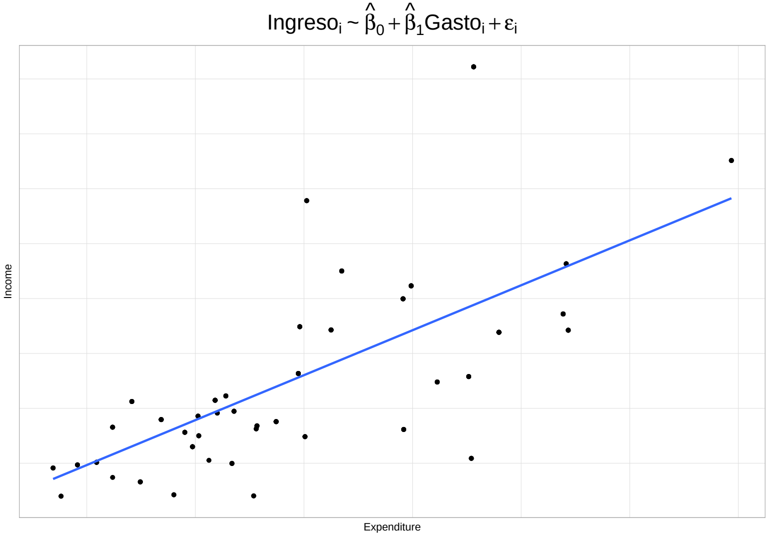
Income	Expenditure	Stratum	Sex	Region	Zone
697.3	296.1	idStrt017	Male	Norte	Rural
697.3	296.1	idStrt017	Female	Norte	Rural
697.3	296.1	idStrt017	Male	Norte	Rural
697.3	296.1	idStrt017	Female	Norte	Rural
526.8	294.8	idStrt017	Male	Norte	Rural
526.8	294.8	idStrt017	Female	Norte	Rural
526.8	294.8	idStrt017	Female	Norte	Rural
526.8	294.8	idStrt017	Male	Norte	Rural
526.8	294.8	idStrt017	Male	Norte	Rural
526.8	294.8	idStrt017	Female	Norte	Rural

Introducción a los modelos multinivel.

Modelo lineal sin considerar el efecto de los estratos.

```
ggplot(data = encuesta_plot,
       aes(y = Income, x = Expenditure)) +
  geom_jitter() +
  theme( legend.position="none",
         plot.title = element_text(hjust = 0.5)) +
  geom_smooth( formula = y ~ x,
               method = "lm",  se = F) +
  ggtitle(
latex2exp::TeX("$Ingreso_{i}\\sim\\hat{\\beta}_{0}+\\hat{\\beta}_{1}Gasto_{i}+\\epsilon_{i}$"))
  theme_cepal()
```

Introducción a los modelos multinivel.



Introducción a los modelos multinivel.

Estimando un intercepto aleatorio

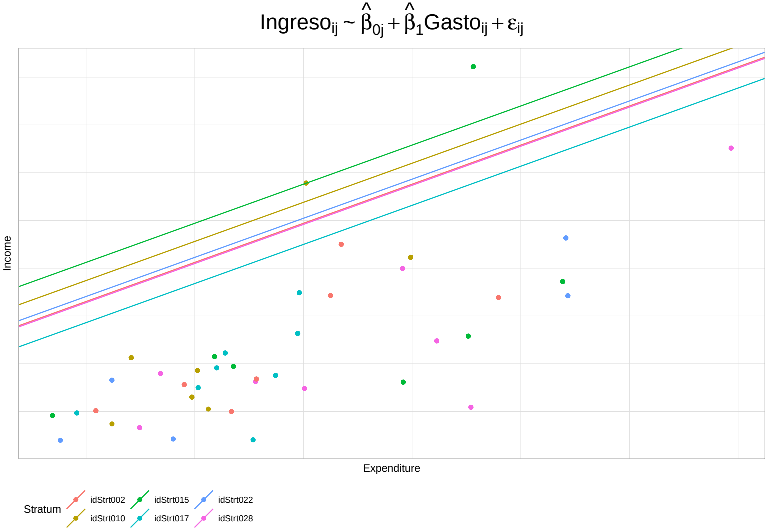
```
B1 <- coef(lm(Income ~ Expenditure, data = encuesta_p
(coef_Mod <- encuesta_plot %>% group_by(Stratum) %>%
  summarise(B0 = coef(lm(Income ~ 1))[1]) %>%
  mutate(B1 = B1))
```

Stratum	B0	B1
idStrt002	496.9	1.637
idStrt010	584.7	1.637
idStrt015	660.6	1.637
idStrt017	408.3	1.637
idStrt022	517.9	1.637
idStrt028	492.1	1.637

Introducción a los modelos multinivel.


```
ggplot(data = encuesta_plot,
       aes(y = Income, x = Expenditure,
           colour = Stratum)) +
  geom_jitter() + theme(legend.position="none",
    plot.title = element_text(hjust = 0.5)) +
  geom_abline(data = coef_Mod,
              mapping=aes(slope=B1,
                          intercept=B0,
                          colour = Stratum)) +
  ggtitle(
    latex2exp::TeX("$Ingreso_{ij}\\sim\\hat{\\beta}_{0j}+\\hat{\\beta}_{1}Gasto_{ij}+\\epsilon_{
  theme_cepal()
```

Introducción a los modelos multinivel.



Introducción a los modelos multinivel.

Estimando una pendiente aleatoria para cada estrato

```
B0 <- coef(lm(Income ~ Expenditure,
               data = encuesta_plot))[1]
(coef_Mod <- encuesta_plot %>% group_by(Stratum) %>%
  summarise(
    B1 = coef(lm(Income ~ -1 + Expenditure))[1]) %>%
  mutate(B0 = B0))
```

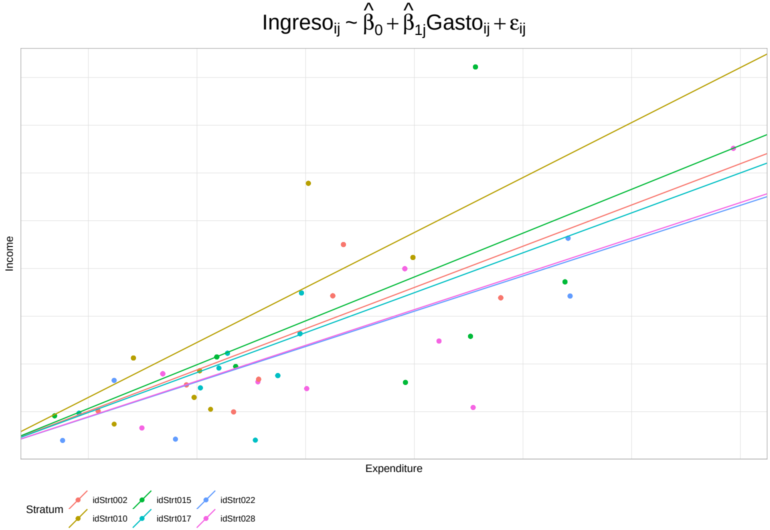
Stratum	B1	B0
idStrt002	1.727	29.56
idStrt010	2.303	29.56
idStrt015	1.837	29.56
idStrt017	1.672	29.56
idStrt022	1.478	29.56
idStrt028	1.495	29.56

Creando el grafico con pendientes aleatorias

```
ggplot(data = encuesta_plot,
       aes(y = Income, x = Expenditure,
           colour = Stratum)) +
  geom_jitter() + theme(legend.position="none",
    plot.title = element_text(hjust = 0.5)) +
  geom_abline(data = coef_Mod,
              mapping=aes(slope=B1,
                          intercept=B0, colour = Stratum)) +
  ggtitle(
    latex2exp::TeX("$Ingreso_{ij}\\sim\\hat{\\beta}_{0}+\\hat{\\beta}_{1j}Gasto_{ij}+\\epsilon_{
  theme_cepal()
```

Introducción a los modelos multinivel.

Análisis de encuestas de hogares con R

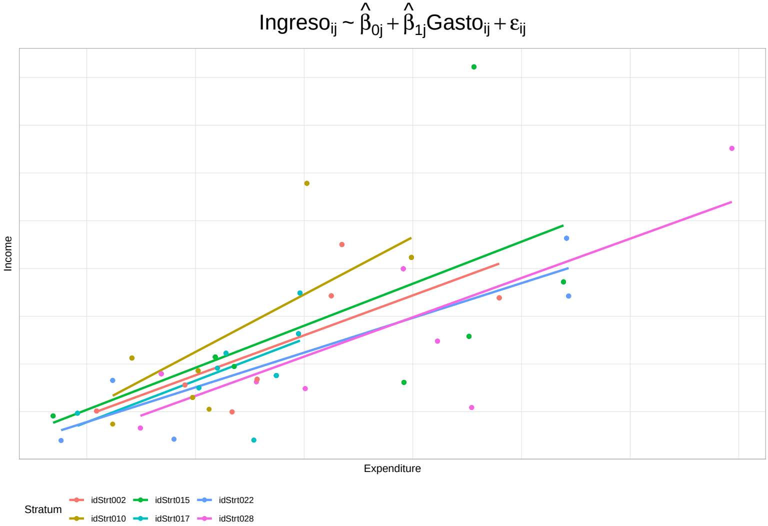


Introducción a los modelos multinivel.

Creando un gráfico con intercepto y pendientes aleatorias.

```
ggplot(data = encuesta_plot,
       aes(y = Income, x = Expenditure,
           colour = Stratum)) +
    geom_smooth( formula = y ~ x, method = "lm", se = F) +
  geom_jitter() + theme(legend.position="none",
    plot.title = element_text(hjust = 0.5)) +
  ggtitle(
    latex2exp::TeX("$Ingreso_{ij}\\sim\\hat{\\beta}_{0j}+\\hat{\\beta}_{1j}Gasto_{ij}+\\epsilon_
  theme_cepal()
```

Introducción a los modelos multinivel.



Introducción a los modelos multinivel.

Dos tipos de índices son relevantes en los análisis multinivel:

- Los coeficientes de regresión, generalmente denominados como los parámetros fijos del modelo.

Cualquier análisis de regresión multinivel siempre debe comenzar con el cálculo de las estimaciones de varianza de Nivel 1 y Nivel 2 para la variable dependiente.

Introducción a los modelos multinivel.

Dos tipos de índices son relevantes en los análisis multinivel:

- Los coeficientes de regresión, generalmente denominados como los parámetros fijos del modelo.
- Las estimaciones de la varianza, generalmente denominadas parámetros aleatorios del modelo.

Cualquier análisis de regresión multinivel siempre debe comenzar con el cálculo de las estimaciones de varianza de Nivel 1 y Nivel 2 para la variable dependiente.

Introducción a los modelos multinivel.

- El primer paso recomendado en el análisis de regresión multinivel consiste en una descomposición de la varianza de la variable dependiente en los diferentes niveles.

Ejemplo La varianza del ingreso se descompondrá en dos componentes:

Estos dos componentes de varianza se pueden obtener una regresión multinivel.

Introducción a los modelos multinivel.

- El primer paso recomendado en el análisis de regresión multinivel consiste en una descomposición de la varianza de la variable dependiente en los diferentes niveles.

Ejemplo La varianza del ingreso se descompondrá en dos componentes:

- La varianza dentro dentro del estrato

Estos dos componentes de varianza se pueden obtener una regresión multinivel.

Introducción a los modelos multinivel.

- El primer paso recomendado en el análisis de regresión multinivel consiste en una descomposición de la varianza de la variable dependiente en los diferentes niveles.

Ejemplo La varianza del ingreso se descompondrá en dos componentes:

- La varianza dentro dentro del estrato
- la varianza entre los estratos.

Estos dos componentes de varianza se pueden obtener una regresión multinivel.

Introducción a los modelos multinivel.

Un modelo básico es:

$y_{ij} = \beta_{0j} + \epsilon_{ij}$

$\beta_{0j} = \gamma_{00} + \tau_{0j}$

- $y_{ij} =$ Los ingresos de la persona i en el estrato j.

donde, $\tau_{0j} \sim N\left(0, \sigma_{\tau}^{2}\right)$ y $\epsilon_{ij} \sim N\left(0, \sigma_{\epsilon}^{2}\right)$.

La correlación intra clásica esta dada por:

$$\rho = \frac{\sigma_{\tau}^{2}}{\sigma_{\tau}^{2} + \sigma_{\epsilon}^{2}}$$

Introducción a los modelos multinivel.

Un modelo básico es:

$y_{ij} = \beta_{0j} + \epsilon_{ij}$

$\beta_{0j} = \gamma_{00} + \tau_{0j}$

- $y_{ij} =$ Los ingresos de la persona i en el estrato j.
- $\beta_{0j} =$ El intercepto en el estrato j.

donde, $\tau_{0j} \sim N\left(0, \sigma^2_{\tau}\right)$ y $\epsilon_{ij} \sim N\left(0, \sigma^2_{\epsilon}\right)$.

La correlación intra clásica esta dada por:

$$\rho = \frac{\sigma^2_{\tau}}{\sigma^2_{\tau} + \sigma^2_{\epsilon}}$$

Introducción a los modelos multinivel.

Un modelo básico es:

$y_{ij} = \beta_{0j} + \epsilon_{ij}$

$\beta_{0j} = \gamma_{00} + \tau_{0j}$

- $y_{ij} =$ Los ingresos de la persona i en el estrato j.
- $\beta_{0j} =$ El intercepto en el estrato j.
- ϵ_{ij} El residual de la persona i en el estrato j.

donde, $\tau_{0j} \sim N\left(0, \sigma_{\tau}^{2}\right)$ y $\epsilon_{ij} \sim N\left(0, \sigma_{\epsilon}^{2}\right)$.

La correlación intra clásica esta dada por:

$$\rho = \frac{\sigma_{\tau}^{2}}{\sigma_{\tau}^{2} + \sigma_{\epsilon}^{2}}$$

Introducción a los modelos multinivel.

Un modelo básico es:

$y_{ij} = \beta_{0j} + \epsilon_{ij}$

$\beta_{0j} = \gamma_{00} + \tau_{0j}$

- $y_{ij} =$ Los ingresos de la persona i en el estrato j.
- $\beta_{0j} =$ El intercepto en el estrato j.
- ϵ_{ij} El residual de la persona i en el estrato j.
- $\gamma_{00} =$ El intercepto en general.

donde, $\tau_{0j} \sim N\left(0, \sigma_{\tau}^2\right)$ y $\epsilon_{ij} \sim N\left(0, \sigma_{\epsilon}^2\right)$.

La correlación intra clásica esta dada por:

$$\rho = \frac{\sigma_{\tau}^2}{\sigma_{\tau}^2 + \sigma_{\epsilon}^2}$$

Introducción a los modelos multinivel.

Un modelo básico es:

$y_{ij} = \beta_{0j} + \epsilon_{ij}$

$\beta_{0j} = \gamma_{00} + \tau_{0j}$

- $y_{ij} =$ Los ingresos de la persona i en el estrato j.
- $\beta_{0j} =$ El intercepto en el estrato j.
- ϵ_{ij} El residual de la persona i en el estrato j.
- $\gamma_{00} =$ El intercepto en general.
- $\tau_{0j} =$ Efecto aleatorio para el intercepto.

donde, $\tau_{0j} \sim N\left(0, \sigma^2_{\tau}\right)$ y $\epsilon_{ij} \sim N\left(0, \sigma^2_{\epsilon}\right)$.

La correlación intra clásica esta dada por:

$$\rho = \frac{\sigma^2_{\tau}}{\sigma^2_{\tau} + \sigma^2_{\epsilon}}$$

Modelos multinivel en muestras complejas.

- Aunque existe evidencia suficiente de que las ponderaciones de muestreo deben usarse en el modelado multinivel (MLM) para obtener estimaciones no sesgadas[1], y también sobre cómo deben usarse estas ponderaciones en los análisis de un solo nivel, hay poca discusión en la literatura sobre qué y cómo usar pesos de muestreo en MLM.

[1] Cai, T. (2013). Investigation of ways to handle sampling weights for multilevel model analyses. Sociological Methodology, 43(1), 178-219.

Modelos multinivel en muestras complejas.

- Aunque existe evidencia suficiente de que las ponderaciones de muestreo deben usarse en el modelado multinivel (MLM) para obtener estimaciones no sesgadas[1], y también sobre cómo deben usarse estas ponderaciones en los análisis de un solo nivel, hay poca discusión en la literatura sobre qué y cómo usar pesos de muestreo en MLM.
- Actualmente, diferentes autores recomiendan cuatro enfoques diferentes sobre cómo usar los pesos de muestreo en modelos jerárquicos.

[1] Cai, T. (2013). Investigation of ways to handle sampling weights for multilevel model analyses. Sociological Methodology, 43(1), 178-219.

Introducción modelos multinivel.

- Pfefermann et al. (1998) y Asparouhov (2006) aconsejan utilizar un enfoque de pseudomáxima verosimilitud para calcular estimaciones dentro y entre los diferentes niveles utilizando la técnica de maximización de mínimos cuadrados generalizados ponderados por probabilidad (PWGLS) para obtener estimaciones no sesgadas.[2][3]

---

[2]Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., & Rasbash, J. (1998). Weighting for unequal selection probabilities in multilevel models. Journal of the Royal Statistical Society: series B (statistical methodology), 60(1), 23-40.

[3]Asparouhov, T. (2006). General multi-level modeling with sampling weights. Communications in Statistics—Theory and Methods, 35(3), 439-460.

Introducción modelos multinivel.

- Rabe-Hesketh y Skrondal (2006) proporcionan técnicas de maximización de expectativas para maximizar la pseudoverosimilitud[4]

[4] Asparouhov, T., & Muthen, B. (2006, August). Multilevel modeling of complex survey data. In Proceedings of the joint statistical meeting in Seattle (pp. 2718-2726).

Estimación de pseudo máxima verosimilitud

La función de log-verosimilitud para la población esta dada por:

$$L_U\left(\theta\right)=\sum_{i\in U}\log\left[f\left(\boldsymbol{y}_i;\theta\right)\right]$$

El estimador de máxima verosimilitud esta dada por:

$$\frac{\partial L_U\left(\theta\right)}{\partial\theta}=0$$

La dificultad que encontramos aquí, es transferir los pesos muéstrales a los niveles inferiores, por ejemplo UPMs -> Stratum.

Estimación de pseudo máxima verosimilitud

Pfeffermann et al. (1998) argumentaron que debido a la estructura de datos agrupados, ya no se asume que las observaciones sean independientes y que la probabilidad logarítmica se convierta en una suma entre los elementos de nivel uno y dos en lugar de una simple suma de las contribuciones de los elementos.

Modelo Nulo

asuma que la información dentro del estrato esta definida por el intercepto.

$$Ingreso_{ij} = \beta_{0j} + \epsilon_{ij}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Stratum_j + \tau_{1j}$$

Ajuste de pesos (alternativa a los Modelo Qweighted)

Estimado el modelo con Qweighted

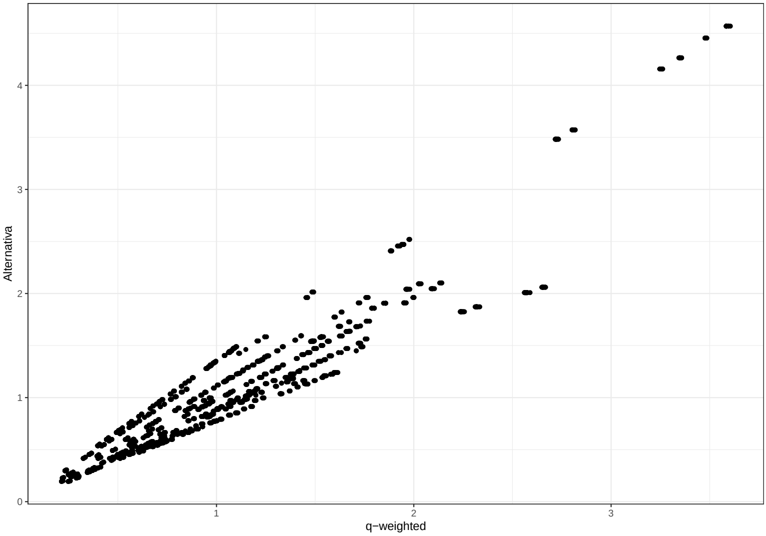
```
mod_qw <- lm(wk ~ Age + Sex + Region + Zone,
             data = encuesta)
encuesta$wk2 <-   encuesta$wk/predict(mod_qw)
# Alternativa los Qweighted
n = nrow(encuesta)
encuesta <- encuesta %>% mutate(wk3 = n*wk/sum(wk))
encuesta %>% summarise(fep = sum(wk),
                       q_wei = sum(wk2),
                       fep2 = sum(wk3) )
```

fep	q_wei	fep2
150266	2602	2605

Comparando los pesos.


```
ggplot(encuesta, aes(x = wk2, y = wk3)) +
  geom_point() + theme_bw() +
  labs(x = "q-weighted", y = "Alternativa")
```

Comparando los pesos.



Modelo Nulo

La estimación del modelo se hace con la función `lmer` de la libreria *lmer*

```
library(lme4)

mod_null  <- lmer( Income  ~ ( 1  |  Stratum ),
                       data  =  encuesta,
                  weights  =  wk2 )

mod_null2  <- lmer( Income  ~ ( 1  |  Stratum ),
                       data  =  encuesta,
                  weights  =  wk3 )
```

Modelo Nulo

Comparando los modelos obtenidos.

```
coef_mod_null <- bind_cols(coef( mod_null )$Stratum,
          coef(mod_null2 )$Stratum)
colnames(coef_mod_null) <- c("Intercept Mod 1",
                             "Intercept Mod 2")
coef_mod_null %>% slice(1:12)
```

Modelo Nulo

	Intercept Mod 1	Intercept Mod 2
idStrt001	630.7	630.1
idStrt002	505.4	506.2
idStrt003	481.3	484.7
idStrt004	959.6	954.5
idStrt005	514.6	515.9
idStrt006	433.8	438.2
idStrt007	467.5	470.5
idStrt008	371.6	376.4
idStrt009	207.6	218.1
idStrt010	591.6	592.1
idStrt011	588.8	588.3
idStrt012	352.0	361.2

Modelo Nulo


```
mod_null
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: Income ~ (1 | Stratum)
##    Data: encuesta
## Weights: wk2
## REML criterion at convergence: 39356
## Random effects:
##  Groups   Name        Std.Dev.
##  Stratum  (Intercept) 281
##  Residual             408
## Number of obs: 2605, groups:  Stratum, 119
## Fixed Effects:
## (Intercept)
##         584
```

Modelo Nulo

Correlación intraclases

```
#library(sjstats)
sjstats::icc(mod_null)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.322
##    Unadjusted ICC: 0.322
```

Modelo Nulo

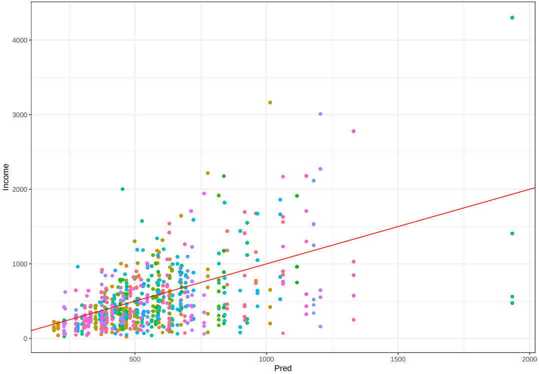
Predicción dentro de los estrato es constante.

```
(tab_pred <- data.frame(Pred = predict(mod_null),
            Income = encuesta$Income,
            Stratum = encuesta$Stratum)) %>% distinct(
  slice(1:6L) # Son las pendientes aleatorias
```

	Pred	Income	Stratum
1	630.7	409.87	idStrt001
6	630.7	823.75	idStrt001
10	630.7	90.92	idStrt001
13	630.7	135.33	idStrt001
18	630.7	336.19	idStrt001
22	630.7	1539.75	idStrt001

Scaterplot de y vs $\hat{y}$

Si la predicción es correcta se espera estar sobre la linea de 45°



Modelo con intercepto aleatoria

Consideremos el siguiente modelo

$$Ingreso_{ij} = \beta_0 + \beta_{1j} Gasto_{ij} + \epsilon_{ij}$$

donde β_{1j} esta dado como

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Stratum_j + \tau_{1j}$$

```
mod_Int_Aleatorio <- lmer(
  Income ~ Expenditure  + (1 | Stratum),
  data = encuesta, weights  =  wk2)
sjstats::icc(mod_Int_Aleatorio)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.196
##    Unadjusted ICC: 0.102
```

Modelo con intercepto aleatoria

Para cada estrato se tiene las siguientes estimaciones de β_{1j}

```
coef(mod_Int_Aleatorio)$Stratum %>% slice(1:8L)
```

	(Intercept)	Expenditure
idStrt001	248.257	1.202
idStrt002	152.988	1.202
idStrt003	139.765	1.202
idStrt004	292.650	1.202
idStrt005	-42.165	1.202
idStrt006	46.766	1.202
idStrt007	2.841	1.202
idStrt008	103.346	1.202

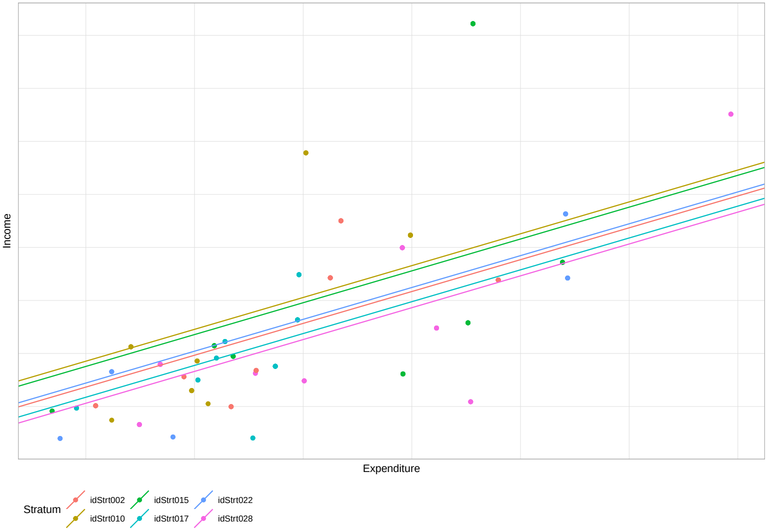
Modelo con intercepto aleatoria

Organizando los coeficientes para el gráfico.

```
Coef_Estimado <- inner_join(
  coef(mod_Int_Aleatorio)$Stratum %>%
       add_rownames(var = "Stratum"),
encuesta_plot %>% select(Stratum) %>% distinct())

ggplot(data = encuesta_plot,
       aes(y = Income, x = Expenditure,
           colour = Stratum)) +
  geom_jitter() + theme(legend.position="none",
    plot.title = element_text(hjust = 0.5)) +
  geom_abline(data = Coef_Estimado,
              mapping=aes(slope=Expenditure,
                          intercept=`(Intercept)`,
                          colour = Stratum))+
  theme_cepal()
```

Modelo con intercepto aleatoria



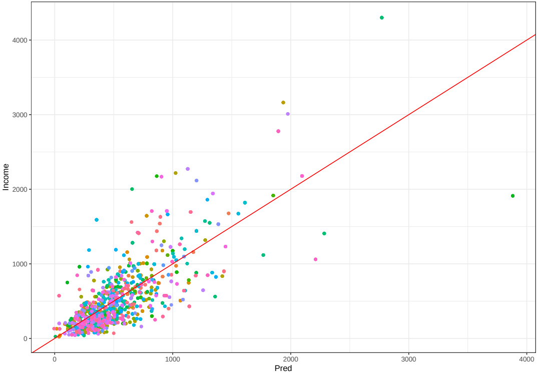
Predicción del modelo


```
(tab_pred <- data.frame(
  Pred = predict(mod_Int_Aleatorio),
            Income = encuesta$Income,
            Stratum = encuesta$Stratum)) %>% distinct(
  slice(1:6L) # Son las pendientes aleatorias
```

	Pred	Income	Stratum
1	664.4	409.87	idStrt001
6	719.6	823.75	idStrt001
10	337.3	90.92	idStrt001
13	348.9	135.33	idStrt001
18	560.9	336.19	idStrt001
22	890.5	1539.75	idStrt001

Scaterplot de y vs $\hat{y}$

La predicción esta más cerca a la linea de 45 grados.



Modelo con intercepto y pendiente aleatoria

$$Ingreso_{ij} = \beta_{0j} + \beta_{1j} Gasto_{ij} + \epsilon_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Stratum_{j} + \tau_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Stratum_{j} + \tau_{1j}$$

```
mod_Pen_Aleatorio <- lmer(
  Income ~ Expenditure  + (1 + Expenditure| Stratum),
  data = encuesta, weights  =  wk2)

sjstats::icc(mod_Pen_Aleatorio)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.700
```

Modelo con intercepto y pendiente aleatoria


```
coef(mod_Pen_Aleatorio)$Stratum %>% slice(1:10L)
```

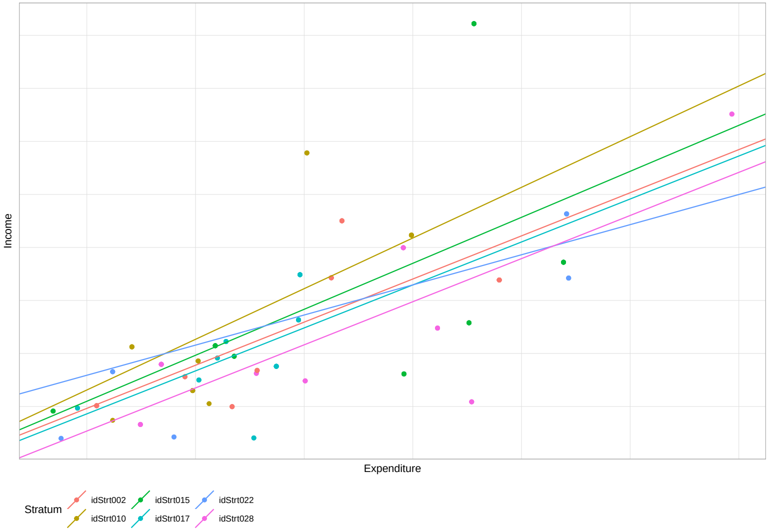
	(Intercept)	Expenditure
idStrt001	-232.75	2.7843
idStrt002	30.20	1.6268
idStrt003	152.46	1.1621
idStrt004	229.66	1.3471
idStrt005	-96.03	1.2946
idStrt006	31.79	1.2003
idStrt007	38.05	1.0764
idStrt008	168.67	0.8971
idStrt009	32.73	0.7396
idStrt010	71.10	1.9112

Modelo con intercepto y pendiente aleatoria


```
Coef_Estimado <- inner_join(
  coef(mod_Pen_Aleatorio)$Stratum %>%
       add_rownames(var = "Stratum"),
encuesta_plot %>% select(Stratum) %>% distinct())

ggplot(data = encuesta_plot,
       aes(y = Income, x = Expenditure,
           colour = Stratum)) +
  geom_jitter() + theme(legend.position="none",
    plot.title = element_text(hjust = 0.5)) +
  geom_abline(data = Coef_Estimado,
              mapping=aes(slope=Expenditure,
                          intercept=`(Intercept)`,
                          colour = Stratum))+
  theme_cepal()
```

Modelo con intercepto y pendiente aleatoria

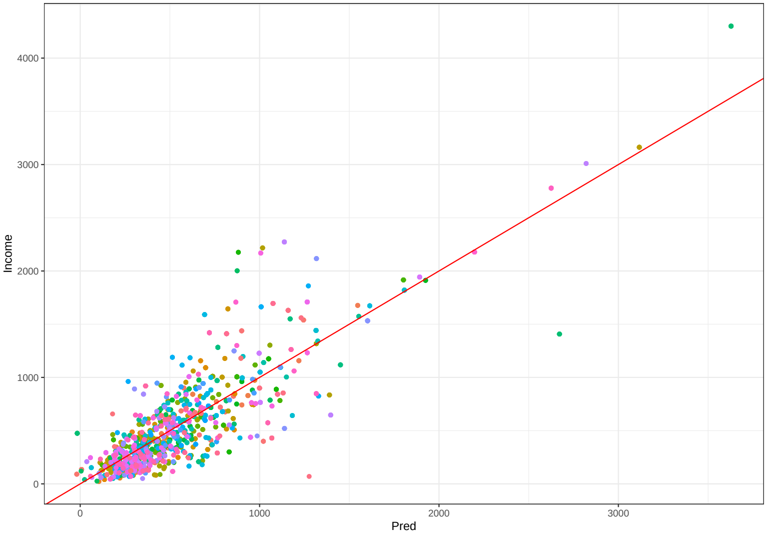


Predicción del modelo


```
(tab_pred <- data.frame(Pred = predict(mod_Pen_Aleato
            Income = encuesta$Income,
            Stratum = encuesta$Stratum)) %>% distinct(
  slice(1:6L) # Son las pendientes aleatorias
```

	Pred	Income	Stratum
1	731.5694	409.87	idStrt001
6	859.3694	823.75	idStrt001
10	-26.5154	90.92	idStrt001
13	0.5481	135.33	idStrt001
18	491.6731	336.19	idStrt001
22	1255.2708	1539.75	idStrt001

Scaterplot de y vs $\hat{y}$



Modelo con intercepto y pendiente aleatoria

$$Ingreso_{ij} = \beta_{0j} + \beta_{1j} Gasto_{ij} + \beta_{2j} Zona_{ij} + \epsilon_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Stratum_j + \gamma_{02}\mu_j + \tau_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Stratum_j + \gamma_{12}\mu_j + \tau_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21} Stratum_j + \gamma_{12}\mu_j + \tau_{2j}$$

donde μ_j es el gasto medio en el estrato j.

Modelo con intercepto y pendiente aleatoria

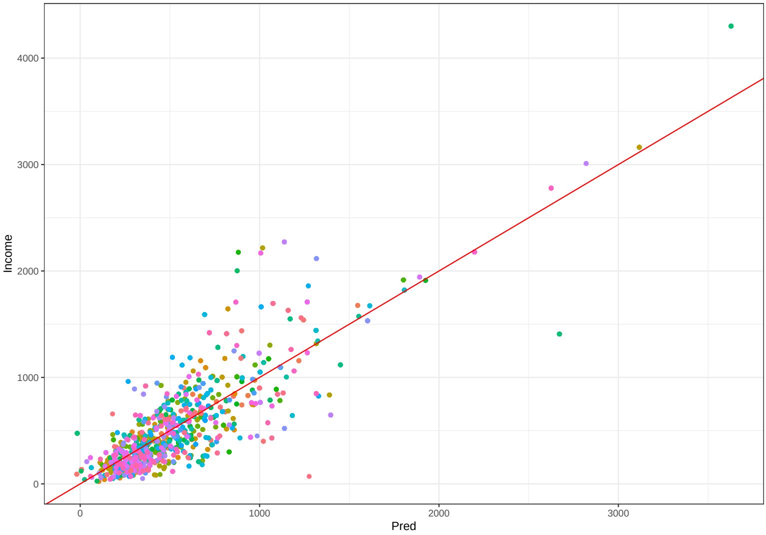
```
media_estrato <- encuesta %>% group_by(Stratum) %>%
  summarise(mu = mean(Expenditure))
encuesta <- inner_join(encuesta,
                        media_estrato, by = "Stratum")

mod_Pen_Aleatorio2 <- lmer(
  Income ~ 1 + Expenditure + Zone + mu +
    (1 + Expenditure + Zone + mu | Stratum ),
    data = encuesta, weights  =  wk2)
sjstats::icc(mod_Pen_Aleatorio2)
```

```
## [1] NA
```

```
(tab_pred <- data.frame(Pred = predict(mod_Pen_Aleato
           Income = encuesta$Income,
           Stratum = encuesta$Stratum)) %>% distinct(
```

Scaterplot de y vs $\hat{y}$



Modelo con intercepto y pendiente aleatoria


```
as.data.frame( model.matrix(mod_Pen_Aleatorio2)) %>%
  distinct()
```

Modelo con intercepto y pendiente aleatoria


```
(Coef_Estimado <- inner_join(
  coef(mod_Pen_Aleatorio2)$Stratum %>%
    add_rownames(var = "Stratum"),
  encuesta_plot %>% select(Stratum, Zone) %>% distinc
))
```

Stratum	(Intercept)	Expenditure	ZoneUrban	mu	Zo
idStrt002	51.05	1.592	28.98	-0.1221	Urb
idStrt010	95.38	1.980	147.67	-0.6678	Urb
idStrt015	36.61	1.749	-154.18	-0.0318	Rur
idStrt017	55.41	1.577	43.06	-0.1365	Rur
idStrt022	41.23	1.133	26.51	0.2705	Urb
idStrt028	50.22	1.568	-81.52	0.0029	Urb

Modelo con intercepto y pendiente aleatoria


```
(Coef_Estimado %<>% inner_join(
  media_estrato, by = "Stratum"))
```

Stratum	(Intercept)	Expenditure	ZoneUrban	mu.x	Zone	mu.y
idStrt002	51.05	1.592	28.98	-0.1221	Urban	286.2
idStrt010	95.38	1.980	147.67	-0.6678	Urban	255.8
idStrt015	36.61	1.749	-154.18	-0.0318	Rural	357.0
idStrt017	55.41	1.577	43.06	-0.1365	Rural	244.8
idStrt022	41.23	1.133	26.51	0.2705	Urban	524.0
idStrt028	50.22	1.568	-81.52	0.0029	Urban	337.1

El modelo para el estrato *idStrt001* viene dado por:

$$\hat{y}_{ij} = 154.4 + 1.7418 Expenditure_{ij} + 77.353 Zone_{ij} + (-0.6954)\,\mu_j$$

$$\hat{y}_{ij} = 154.4 + 1.7418 Expenditure + 77.353\,(0) + (-0.6954)\,(255.2)$$

$$\hat{y}_{ij} = -23.07 + 1.7418 Expenditure$$

Modelo con intercepto y pendiente aleatoria

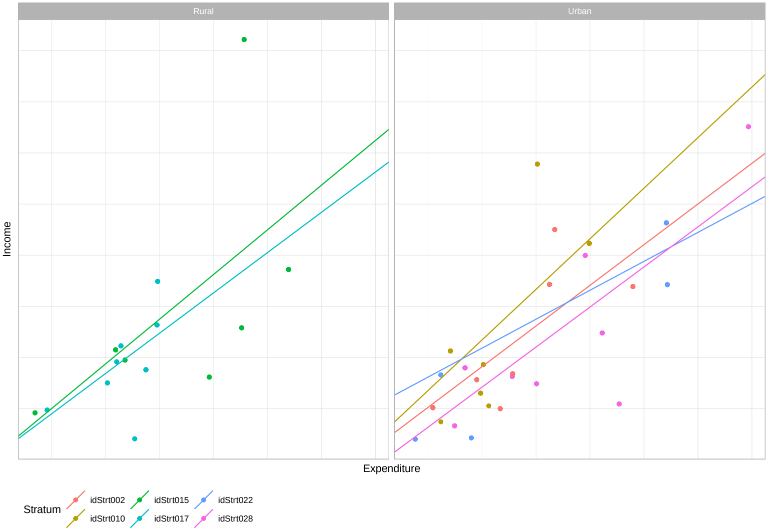

```
(Coef_Estimado %<>%  mutate(B0 = ifelse(
Zone == "Urban", `(Intercept)` + mu.y * mu.x + ZoneUr
                 `(Intercept)` + mu.y * mu.x)) %>%
   select(Stratum, Zone, B0, Expenditure))
```

Stratum	Zone	B0	Expenditure
idStrt002	Urban	45.08	1.592
idStrt010	Urban	72.24	1.980
idStrt015	Rural	25.26	1.749
idStrt017	Rural	21.99	1.577
idStrt022	Urban	209.51	1.133
idStrt028	Urban	-30.34	1.568

Modelo con intercepto y pendiente aleatoria


```
ggplot(data = encuesta_plot,
           aes(y = Income, x = Expenditure,
                 colour = Stratum)) +
   geom_jitter() +
   theme(legend.position = "none",
             plot.title = element_text(hjust = 0.5)) +
   facet_grid( ~ Zone) +
   geom_abline(
     data = Coef_Estimado,
     mapping = aes(
       slope = Expenditure,
       intercept = B0,
       colour = Stratum
     )
   ) +
   theme_cepal()
```

Modelo con intercepto y pendiente aleatoria



Introducción a los modelos logístico multinivel.

Sea la variable $y_{ij} = 1$ si el individuo i en el estrato j esta por enciam de la linea de pobreza y $y_{ij} = 0$ en caso contrario, la variable y_{ij} se puede modelar mediante el modelo logístico:

$$Pr\left(y_{ij}\right) = Pr\left(y_{ij} = 1 \mid x_i : \boldsymbol{\beta}\right) = \frac{1}{1 + \exp\left(-\boldsymbol{\beta}_j \boldsymbol{x}_{ij}\right)}$$

ó

$$\log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = \boldsymbol{\beta}_j \boldsymbol{x}_{ij}$$

donde $\pi_{ij} = Pr\left(y_{ij} = 1 \mid x_i : \boldsymbol{\beta}\right)$.

Ejemplos de modelo logit


```
encuesta_plot <- encuesta %>%
  dplyr::select(Stratum,Expenditure) %>% unique() %>%
  group_by(Stratum)  %>%
  summarise(sd = sd(Expenditure)) %>%
  arrange(desc(sd)) %>% dplyr::select(-sd) %>%
  slice(1:20L) %>%
  inner_join(encuesta) %>%
  dplyr::select(Poverty, Expenditure, Stratum,
         Sex, Region, Zone)
encuesta_plot %>% slice(1:15L)
```

Ejemplos de modelo logit

Poverty	Expenditure	Stratum	Sex	Region	Zone
NotPoor	3367.5	idStrt039	Male	Sur	Urban
NotPoor	3367.5	idStrt039	Female	Sur	Urban
NotPoor	3367.5	idStrt039	Male	Sur	Urban
NotPoor	312.1	idStrt039	Female	Sur	Urban
NotPoor	312.1	idStrt039	Female	Sur	Urban
NotPoor	312.1	idStrt039	Female	Sur	Urban
NotPoor	312.1	idStrt039	Male	Sur	Urban
NotPoor	226.5	idStrt039	Male	Sur	Urban
NotPoor	226.5	idStrt039	Female	Sur	Urban
NotPoor	616.3	idStrt047	Female	Sur	Urban
NotPoor	616.3	idStrt047	Female	Sur	Urban
NotPoor	616.3	idStrt047	Female	Sur	Urban
NotPoor	1385.7	idStrt047	Male	Sur	Urban
NotPoor	1385.7	idStrt047	Female	Sur	Urban
NotPoor	1385.7	idStrt047	Female	Sur	Urban

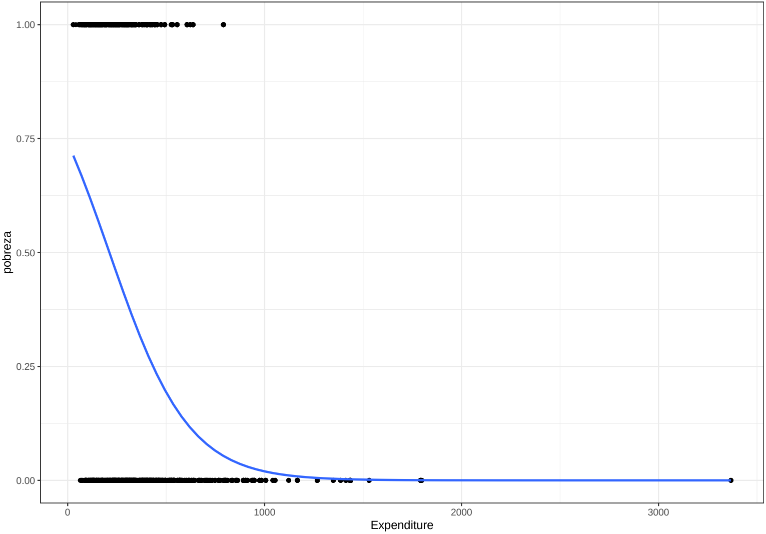
Ejemplos de modelo logit


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

ggplot(data = encuesta,
       aes(y = pobreza, x = Expenditure)) +
  geom_point() +
  geom_smooth(
    formula = y~x, method = "glm",
    se=FALSE,
    method.args = list(family=binomial(link = "logit"
  theme_bw()
```

Ejemplos de modelo logit

Análisis de encuestas de hogares con R



Ejemplos de modelo logit

```r
auxLogit <- function(x,b0,b1){
  1/(1+exp(-(b0+b1*x)))
}
B0 = coef(glm(pobreza~1,data = encuesta_plot,
     family=binomial(link = "logit")))

(coef_Mod <- encuesta_plot %>% group_by(Stratum) %>%
  summarise(B1 = coef(glm(pobreza ~  -1 + Expenditure
                family=binomial(link = "logit")))) %>%
mutate(B0 = B0)) %>% slice(1:6L)
```

Stratum	B1	B0
idStrt007	-0.0189	-0.8782
idStrt020	-0.0010	-0.8782
idStrt022	-0.0057	-0.8782
idStrt024	-0.0020	-0.8782

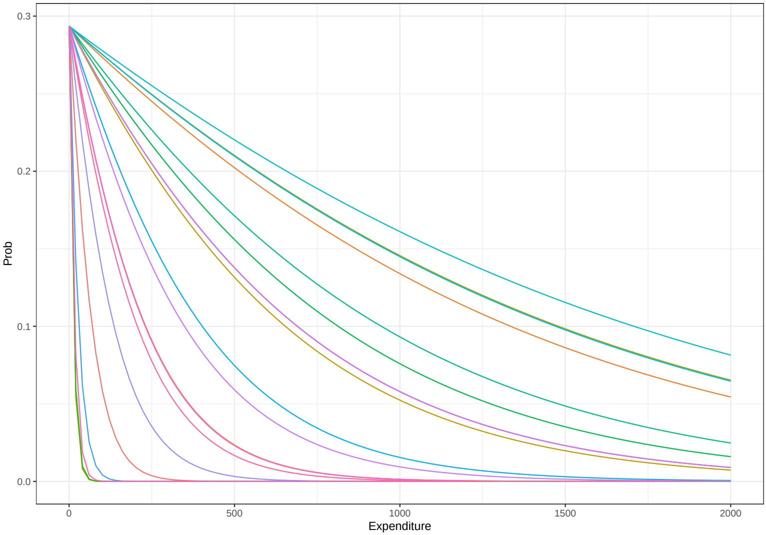
Ejemplos de modelo logit


```
# Creando las variables respuesta
pred_logit <- coef_Mod %>%
  mutate(Expenditure = list(seq(0,2000, length =100))
    tidyr::unnest_legacy()
pred_logit %<>% mutate(Prob = auxLogit(Expenditure,B0

ggplot(data = pred_logit,
       aes(y = Prob, x = Expenditure, colour = Stratu
  geom_line() +
   theme_bw() +
  theme(legend.position = "none")
```

Ejemplos de modelo logit

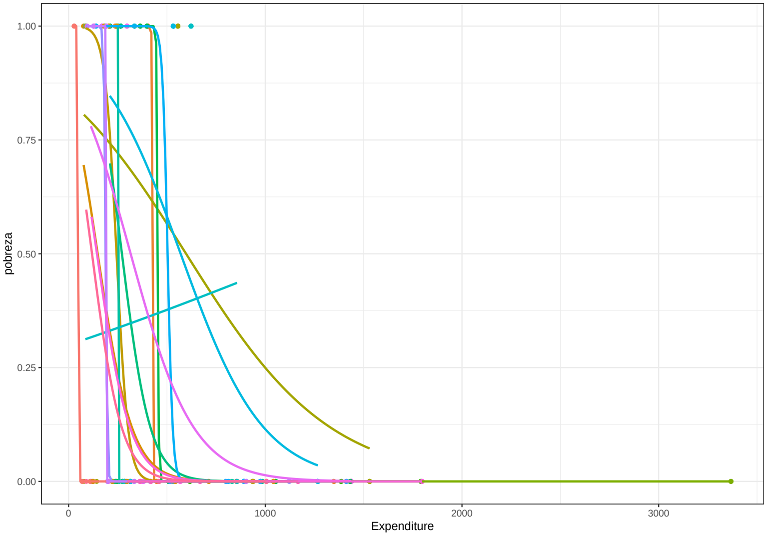
Análisis de encuestas de hogares con R



Ejemplos de modelo logit


```
ggplot(data = encuesta_plot,
       aes(y = pobreza, x = Expenditure, colour = Str
  geom_point() +
  geom_smooth(
    formula = y~x, method = "glm",
    se=FALSE,
    method.args = list(family=binomial(link = "logit"
  theme_bw() +
  theme(legend.position = "none")
```

Ejemplos de modelo logit



Un modelo básico es:

$$logit(\pi_{ij}) = \beta_{0j} + \epsilon_{ij}$$

$$\beta_{0j} = \gamma_{00} + \tau_{0j}$$

- $\pi_{ij} = Pr\left(y_{ij} = 1 \mid x_i : \boldsymbol{\beta}\right)$.

donde, $\tau_{0j} \sim N\left(0, \sigma_{\tau}^2\right)$ y $\epsilon_{ij} \sim N\left(0, \sigma_{\epsilon}^2\right)$.

La correlación intra clásica esta dada por:

$$\rho = \frac{\sigma_{\tau}^2}{\sigma_{\tau}^2 + \sigma_{\epsilon}^2}$$

Un modelo básico es:

$$logit(\pi_{ij}) = \beta_{0j} + \epsilon_{ij}$$

$$\beta_{0j} = \gamma_{00} + \tau_{0j}$$

- $\pi_{ij} = Pr\left(y_{ij} = 1 \mid x_i : \boldsymbol{\beta}\right)$.
- $\beta_{0j} =$ El intercepto en el estrato j.

donde, $\tau_{0j} \sim N\left(0, \sigma_\tau^2\right)$ y $\epsilon_{ij} \sim N\left(0, \sigma_\epsilon^2\right)$.

La correlación intra clásica esta dada por:

$$\rho = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\epsilon^2}$$

Un modelo básico es:

$$logit(\pi_{ij}) = \beta_{0j} + \epsilon_{ij}$$

$$\beta_{0j} = \gamma_{00} + \tau_{0j}$$

- $\pi_{ij} = Pr\left(y_{ij} = 1 \mid x_i : \boldsymbol{\beta}\right)$.
- $\beta_{0j} =$ El intercepto en el estrato j.
- ϵ_{ij} El residual de la persona i en el estrato j.

donde, $\tau_{0j} \sim N\left(0, \sigma_{\tau}^2\right)$ y $\epsilon_{ij} \sim N\left(0, \sigma_{\epsilon}^2\right)$.

La correlación intra clásica esta dada por:

$$\rho = \frac{\sigma_{\tau}^2}{\sigma_{\tau}^2 + \sigma_{\epsilon}^2}$$

Un modelo básico es:

$$logit(\pi_{ij}) = \beta_{0j} + \epsilon_{ij}$$

$$\beta_{0j} = \gamma_{00} + \tau_{0j}$$

- $\pi_{ij} = Pr\left(y_{ij} = 1 \mid x_i : \boldsymbol{\beta}\right)$.
- $\beta_{0j} =$ El intercepto en el estrato j.
- ϵ_{ij} El residual de la persona i en el estrato j.
- $\gamma_{00} =$ El intercepto en general.

donde, $\tau_{0j} \sim N\left(0, \sigma^2_{\tau}\right)$ y $\epsilon_{ij} \sim N\left(0, \sigma^2_{\epsilon}\right)$.

La correlación intra clásica esta dada por:

$$\rho = \frac{\sigma^2_{\tau}}{\sigma^2_{\tau} + \sigma^2_{\epsilon}}$$

Un modelo básico es:

$$logit(\pi_{ij}) = \beta_{0j} + \epsilon_{ij}$$

$$\beta_{0j} = \gamma_{00} + \tau_{0j}$$

- $\pi_{ij} = Pr\left(y_{ij} = 1 \mid x_i : \boldsymbol{\beta}\right)$.
- $\beta_{0j} =$ El intercepto en el estrato j.
- ϵ_{ij} El residual de la persona i en el estrato j.
- $\gamma_{00} =$ El intercepto en general.
- $\tau_{0j} =$ Efecto aleatorio para el intercepto.

donde, $\tau_{0j} \sim N\left(0, \sigma_{\tau}^2\right)$ y $\epsilon_{ij} \sim N\left(0, \sigma_{\epsilon}^2\right)$.

La correlación intra clásica esta dada por:

$$\rho = \frac{\sigma_{\tau}^2}{\sigma_{\tau}^2 + \sigma_{\epsilon}^2}$$

Modelo Nulo


```
library(lme4)
mod_logist_null  <- glmer( pobreza  ~ ( 1   |  Stratum
                  data   =  encuesta,
            weights   =   wk2,
            family = binomial(link = "logit") )


coef( mod_logist_null )$Stratum %>% slice(1:12)
```

Modelo Nulo

	(Intercept)
idStrt001	-0.8334
idStrt002	-0.0133
idStrt003	-2.6023
idStrt004	-2.7770
idStrt005	-1.0268
idStrt006	1.0100
idStrt007	-1.0134
idStrt008	0.2035
idStrt009	2.1966
idStrt010	-0.5948
idStrt011	-1.2986
idStrt012	0.2825

Modelo Nulo


```
library(sjstats)
mod_logist_null
```

```
## Generalized linear mixed model fit by maximum like
##   Approximation) [glmerMod]
##  Family: binomial  ( logit )
## Formula: pobreza ~ (1 | Stratum)
##    Data: encuesta
## Weights: wk2
##      AIC      BIC   logLik deviance df.resid
##     2966     2978    -1481     2962     2603
## Random effects:
##  Groups  Name        Std.Dev.
##  Stratum (Intercept) 1.29
## Number of obs: 2605, groups:  Stratum, 119
## Fixed Effects:
```

Modelo Nulo


```
sjstats::icc(mod_logist_null)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.334
##    Unadjusted ICC: 0.334
```

```
(tab_pred <- data.frame(
  Pred = predict(mod_logist_null, type = "response"),
  pobreza = encuesta$pobreza,
  Stratum = encuesta$Stratum)) %>% distinct() %>%
  slice(1:6L) # Son las pendientes aleatorias
```

	Pred	pobreza	Stratum
1	0.3029	0	idStrt001
10	0.3029	1	idStrt001
28	0.4967	1	idStrt002
36	0.4967	0	idStrt002
61	0.0690	0	idStrt003

Estimación de la propoción para y y $\hat{y}$


```
weighted.mean(encuesta$pobreza, encuesta$wk2)
```

```
## [1] 0.3859
```

```
weighted.mean(tab_pred$Pred, encuesta$wk2)
```

```
## [1] 0.385
```

Modelo con intercepto aleatoria

$$logit(\pi_{ij}) = \beta_0 + \beta_{1j} Gasto_{ij} + \epsilon_{ij}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Stratum_j + \tau_{1j}$$

```
mod_logit_Int_Aleatorio <- glmer(
  pobreza ~ Expenditure  + (1 | Stratum),
  data = encuesta, family = binomial(link = "logit"),
  weights  =  wk2)

sjstats::icc(mod_logit_Int_Aleatorio)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.315
##    Unadjusted ICC: 0.187
```

Modelo con intercepto aleatoria


```
coef(mod_logit_Int_Aleatorio)$Stratum %>% slice(1:10L
```

	(Intercept)	Expenditure
idStrt001	0.9889	-0.0066
idStrt002	1.8837	-0.0066
idStrt003	-0.7463	-0.0066
idStrt004	-0.1484	-0.0066
idStrt005	1.7155	-0.0066
idStrt006	3.2456	-0.0066
idStrt007	0.5601	-0.0066
idStrt008	1.6848	-0.0066
idStrt009	3.9332	-0.0066
idStrt010	1.1207	-0.0066

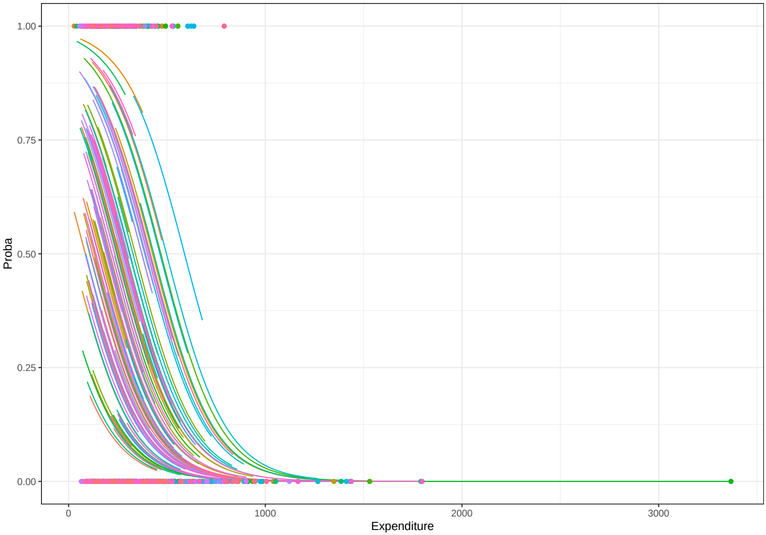
Modelo con intercepto aleatoria

```
dat_pred <- encuesta %>% group_by(Stratum) %>%
  summarise(
    Expenditure = list(seq(min(Expenditure),
                           max(Expenditure), len = 100))) %>%
  tidyr::unnest_legacy()

dat_pred <- mutate(dat_pred,
       Proba = predict(mod_logit_Int_Aleatorio,
                       newdata = dat_pred , type = "response"))

ggplot(data = dat_pred,
       aes(y = Proba, x = Expenditure,
           colour = Stratum)) +
   geom_line()+   theme_bw() +
  geom_point(data = encuesta, aes(y = pobreza, x = Expenditure))+
  theme(legend.position = "none",
        plot.title = element_text(hjust = 0.5))
```

Modelo con intercepto aleatoria



Predicción del modelo


```
(tab_pred <- data.frame(
  Pred = predict(mod_logit_Int_Aleatorio,
                 type = "response"),
           pobreza = encuesta$pobreza,
           Stratum = encuesta$Stratum,
           wk2 = encuesta$wk2)) %>% distinct() %>%
  slice(1:6L) # Son las pendientes aleatorias
```

Pred	pobreza	Stratum	wk2
0.2149	0	idStrt001	0.7770
0.2149	0	idStrt001	0.7501
0.2149	0	idStrt001	0.7463
0.2149	0	idStrt001	0.7717
0.2149	0	idStrt001	0.7438
0.1682	0	idStrt001	0.7507

Estimación de la propoción para y y $\hat{y}$


```
tab_pred %>%
  summarise(Pred = weighted.mean(Pred, wk2),
            pobreza = weighted.mean(pobreza,wk2))
```

Pred	pobreza
0.3855	0.3859

Modelo con intercepto y pendiente aleatoria

$$logit(\pi_{ij}) = \beta_{0j} + \beta_{1j}Gasto_{ij} + \epsilon_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Stratum_{j} + \tau_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}Stratum_{j} + \tau_{1j}$$

```
mod_logit_Pen_Aleatorio <- glmer(
  pobreza ~ Expenditure  + (1 + Expenditure| Stratum),
  data = encuesta, weights  =  wk2,
  binomial(link = "logit"))

sjstats::icc(mod_logit_Pen_Aleatorio)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.886
```

Modelo con intercepto y pendiente aleatoria


```
coef(mod_logit_Pen_Aleatorio)$Stratum %>% slice(1:10L
```

	(Intercept)	Expenditure
idStrt001	5.244	-0.0271
idStrt002	11.059	-0.0394
idStrt003	-1.614	-0.0060
idStrt004	1.655	-0.0153
idStrt005	9.055	-0.0289
idStrt006	-1.354	0.0100
idStrt007	1.035	-0.0136
idStrt008	1.473	-0.0056
idStrt009	4.050	-0.0048
idStrt010	4.310	-0.0214

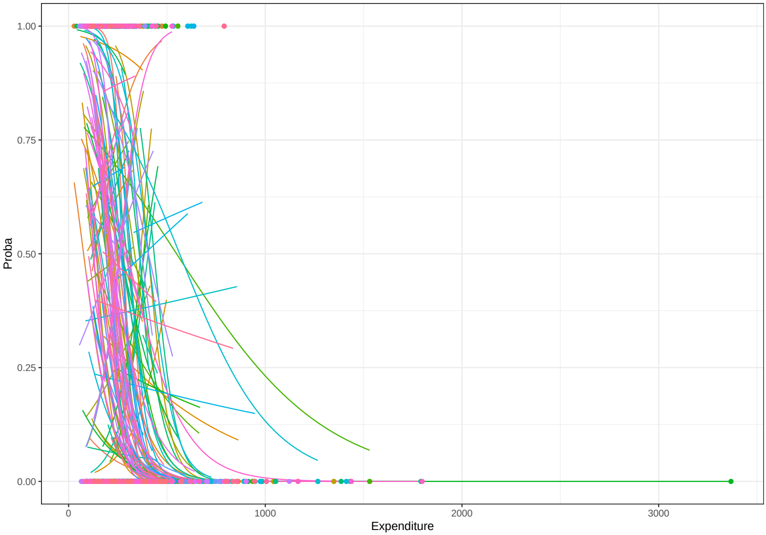
Modelo con intercepto y pendiente aleatoria


```
dat_pred <- encuesta %>% group_by(Stratum) %>%
  summarise(
    Expenditure = list(seq(min(Expenditure),
                             max(Expenditure), len = 100))) %>%
  tidyr::unnest_legacy()

dat_pred <- mutate(dat_pred,
       Proba = predict(mod_logit_Pen_Aleatorio,
                       newdata = dat_pred , type = "response"))

ggplot(data = dat_pred,
       aes(y = Proba, x = Expenditure,
           colour = Stratum)) +
   geom_line()+   theme_bw() +
  geom_point(data = encuesta, aes(y = pobreza, x = Expenditure))+
  theme(legend.position = "none",
        plot.title = element_text(hjust = 0.5))
```

Modelo con intercepto y pendiente aleatoria



Predicción del modelo


```
(tab_pred <- data.frame(
  Pred = predict(mod_logit_Pen_Aleatorio,
                 type = "response"),
            pobreza = encuesta$pobreza,
            Stratum = encuesta$Stratum,
            wk2 = encuesta$wk2)) %>% distinct() %>%
  slice(1:6L)
```

Pred	pobreza	Stratum	wk2
0.0154	0	idStrt001	0.7770
0.0154	0	idStrt001	0.7501
0.0154	0	idStrt001	0.7463
0.0154	0	idStrt001	0.7717
0.0154	0	idStrt001	0.7438
0.0045	0	idStrt001	0.7507

Estimación de la propoción para y y $\hat{y}$


```
tab_pred %>%
  summarise(Pred = weighted.mean(Pred, wk2),
            pobreza = weighted.mean(pobreza,wk2))
```

Pred	pobreza
0.3845	0.3859

Modelo con intercepto y pendiente aleatoria

$$logit(\pi_{ij}) = \beta_{0j} + \beta_{1j} Gasto_{ij} + \beta_{2j} Zona_{ij} + \epsilon_{ij}$$

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Stratum_j + \gamma_{02}\mu_j + \tau_{0j}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Stratum_j + \gamma_{12}\mu_j + \tau_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21} Stratum_j + \gamma_{12}\mu_j + \tau_{2j}$$

donde μ_j es el gasto medio en el estrato j.

Modelo con intercepto y pendiente aleatoria


```
mod_logit_Pen_Aleatorio2 <- glmer(
  pobreza ~ 1 + Expenditure + Zone + mu +
    (1 + Expenditure + Zone + mu | Stratum ),
    data = encuesta, weights  =  wk2,
  binomial(link = "logit"))
sjstats::icc(mod_logit_Pen_Aleatorio2)
```

```
## # Intraclass Correlation Coefficient
##
##      Adjusted ICC: 0.851
##    Unadjusted ICC: 0.594
```

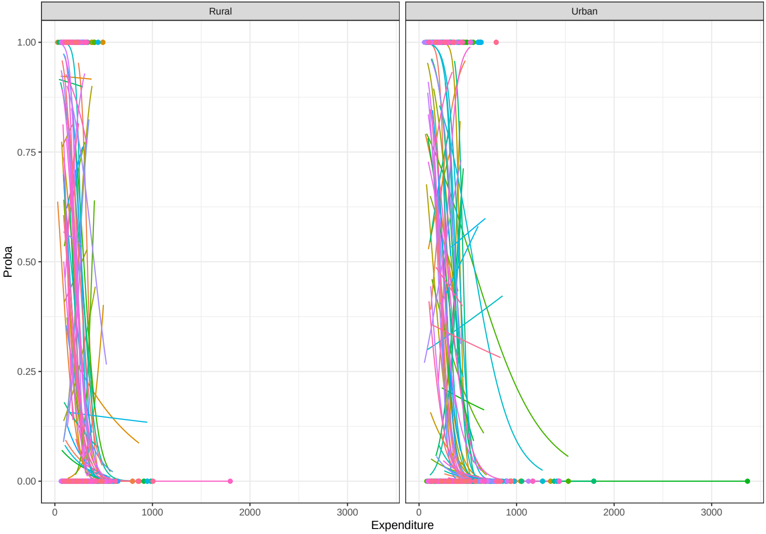
Gráfica del modelo obtenido


```
dat_pred <- encuesta %>% group_by(Stratum, Zone, mu) %>%
  summarise(
    Expenditure = list(seq(min(Expenditure),
                           max(Expenditure), len = 100))) %>%
  tidyr::unnest_legacy()

dat_pred$Proba = predict(mod_logit_Pen_Aleatorio2,
                      newdata = dat_pred , type = "response")

ggplot(data = dat_pred,
       aes(y = Proba, x = Expenditure,
           colour = Stratum)) +
   geom_line()+   theme_bw() +facet_grid(.~Zone)+
  geom_point(data = encuesta, aes(y = pobreza, x = Expenditure))+
  theme(legend.position = "none",
        plot.title = element_text(hjust = 0.5))
```

Modelo con intercepto y pendiente aleatoria



Predicción del modelo


```
(tab_pred <- data.frame(
  Pred = predict(mod_logit_Pen_Aleatorio2,
                 type = "response"),
           pobreza = encuesta$pobreza,
           Stratum = encuesta$Stratum,
           Zone = encuesta$Zone,
           wk2 = encuesta$wk2)) %>% distinct() %>%
  slice(1:6L)
```

Pred	pobreza	Stratum	Zone	wk2
0.0175	0	idStrt001	Rural	0.7770
0.0175	0	idStrt001	Rural	0.7501
0.0175	0	idStrt001	Rural	0.7463
0.0175	0	idStrt001	Rural	0.7717
0.0175	0	idStrt001	Rural	0.7438
0.0053	0	idStrt001	Rural	0.7507

Estimación de la propoción para y y $\hat{y}$

```
tab_pred %>% group_by(Zone) %>%
  summarise(Pred = weighted.mean(Pred, wk2),
            pobreza = weighted.mean(pobreza,wk2))
```

Zone	Pred	pobreza
Rural	0.4309	0.4298
Urban	0.3385	0.3437

¡Gracias!

Email: andres.gutierrez@cepal.org