

## Postwork sesión 8

Un centro de salud nutricional está interesado en analizar estadística y probabilísticamente los patrones de gasto en alimentos saludables y no saludables en los hogares mexicanos con base en su nivel socioeconómico, en si el hogar tiene recursos financieros extra al ingreso y en si presenta o no inseguridad alimentaria. Además, está interesado en un modelo que le permita identificar los determinantes socioeconómicos de la inseguridad alimentaria.

La base de datos es un extracto de la Encuesta Nacional de Salud y Nutrición (2012) levantada por el Instituto Nacional de Salud Pública en México. La mayoría de las personas afirman que los hogares con menor nivel socioeconómico tienden a gastar más en productos no saludables que las personas con mayores niveles socioeconómicos y que esto, entre otros determinantes, lleva a que un hogar presente cierta inseguridad alimentaria.

La base de datos contiene las siguientes variables:

- **nse5f** (nivel socioeconómico del hogar): 1 “Bajo”, 2 “Medio bajo”, 3 “Medio”, 4 “Medio alto”, 5 “Alto”
- **area** (zona geográfica): 0 “Zona urbana”, 1 “Zona rural”
- **numpeho** (número de personas en el hogar)
- **refin** (recursos financieros distintos al ingreso laboral): 0 “no”, 1 “sí”
- **edadjef** (edad del jefe/a de familia)
- **sexojef** (sexo del jefe/a de familia): 0 “Hombre”, 1 “Mujer”
- **añosedu** (años de educación del jefe de familia)
- **ln\_als** (logaritmo natural del gasto en alimentos saludables)
- **ln\_alns** (logaritmo natural del gasto en alimentos no saludables)
- **IA** (inseguridad alimentaria en el hogar): 0 “No presenta IA”, 1 “Presenta IA”

```
df <- read.csv("https://raw.githubusercontent.com/beduExpert/Programacion-R-Santander-2022/main/Sesion-8")
```

- 1) Plantea el problema del caso
- 2) Realiza un análisis descriptivo de la información
- 3) Calcula probabilidades que nos permitan entender el problema en México
- 4) Plantea hipótesis estadísticas y concluye sobre ellas para entender el problema en México
- 5) Estima un modelo de regresión, lineal o logístico, para identificar los determinantes de la inseguridad alimentaria en México
- 6) Escribe tu análisis en un archivo README.md y tu código en un script de R y publica ambos en un repositorio de Github.

NOTA: Todo tu planteamiento deberá estar correctamente desarrollado y deberás analizar e interpretar todos tus resultados para poder dar una conclusión final al problema planteado.

### 1) Planteamiento del problema del caso

#### Objetivos

- Analizar patrones de gasto en alimentos saludables y no saludables en familias mexicanas con base en:
  - el nivel socioeconómico
  - disponibilidad de recursos financieros adicionales al ingreso
  - presencia de inseguridad alimentaria
- Elaborar un modelo que prediga la posibilidad de presentar inseguridad alimentaria (IA) con base en las variables disponibles en el extracto de la Encuesta Nacional de Salud y Nutrición 2012 (ENSANUT)

2012)

## 2) Análisis descriptivo de la información

```
install.packages("ggplot2")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'  
## (as 'lib' is unspecified)
```

```
library(ggplot2)
```

Obtención de los datos:

```
url <- "https://raw.githubusercontent.com/beduExpert/Programacion-R-Santander-2022/main/Sesion-08/Postw  
datos <- read.csv(url, encoding = "UTF-8")  
head(datos)
```

```
##      nse5f area numpeho refin edadjef sexojef añosedu IA   ln_als  ln_alns  
## 1      5    0        4     0      43      0      24  0 5.393628      NA  
## 2      5    0        5     1      NA      NA      24  0 7.024649      NA  
## 3      5    0        4     0      46      0      24  0 6.767343 4.605170  
## 4      5    1        1     0      54      0      24  0 3.401197 4.094345  
## 5      5    0        2     1      39      0      24  0 6.115892 5.480639  
## 6      5    0        5     1      NA      NA      24  0 7.514800 5.598422
```

Para obtener un resumen estadístico de los datos, se convierten en factores las variables discretas.

```
datos$nse5f <- factor(datos$nse5f,  
                      levels = 1:5,  
                      labels = c("Bajo", "Medio bajo", "Medio", "Medio alto", "Alto"),  
                      ordered = TRUE)  
datos$area <- factor(datos$area,  
                    levels = 0:1,  
                    labels = c("Zona urbana", "Zona rural"))  
datos$refin <- factor(datos$refin,  
                     levels = 0:1,  
                     labels = c("no", "sí"))  
datos$sexojef <- factor(datos$sexojef,  
                       levels = 0:1,  
                       labels = c("Hombre", "Mujer"))  
datos$IA <- factor(datos$IA,  
                  levels = 0:1,  
                  labels = c("No presenta IA", "Presenta IA"))
```

Y, enseguida, el resumen estadístico.

```
nrow(datos)
```

```
## [1] 40809
```

```
summary(datos)
```

```
##      nse5f      area      numpeho      refin  
## Bajo      :8858  Zona urbana:26591  Min.      : 1.000  no:33046  
## Medio bajo:8560  Zona rural :14218  1st Qu.: 3.000  sí: 7763  
## Medio      :8323                               Median : 4.000  
## Medio alto:7903                               Mean   : 3.941  
## Alto       :7165                               3rd Qu.: 5.000
```

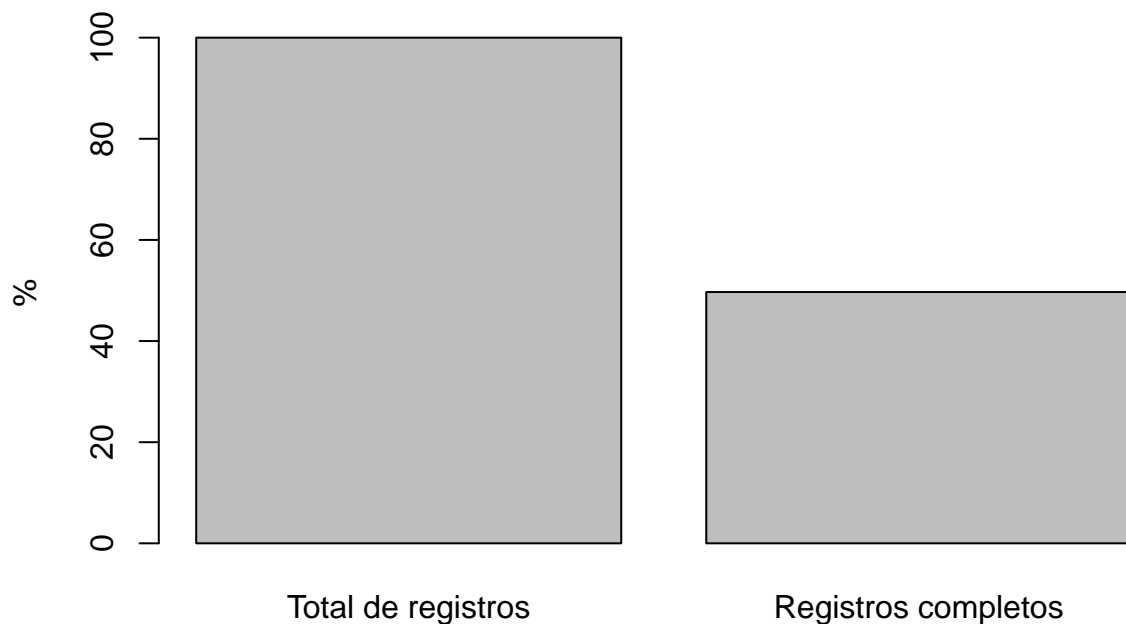
```
##                                     Max.    :19.000
##
##      edadjef      sexojef      añosedu      IA
## Min.    : 18      Hombre:26957   Min.    : 0.00   No presenta IA:10781
## 1st Qu.: 37      Mujer : 8861   1st Qu.: 9.00   Presenta IA   :30028
## Median : 47      NA's  : 4991   Median : 9.00
## Mean    : 49                                     Mean    :10.36
## 3rd Qu.: 60                                     3rd Qu.:12.00
## Max.    :111                                     Max.    :24.00
## NA's    :5017
##      ln_als      ln_alns
## Min.    :0.6931   Min.    :0.000
## 1st Qu.:5.7038   1st Qu.:3.401
## Median :6.1633   Median :4.025
## Mean    :6.0665   Mean    :4.125
## 3rd Qu.:6.5511   3rd Qu.:4.868
## Max.    :8.9699   Max.    :8.403
## NA's    :787     NA's    :17504
```

```
sum(complete.cases(datos))
```

```
## [1] 20280
```

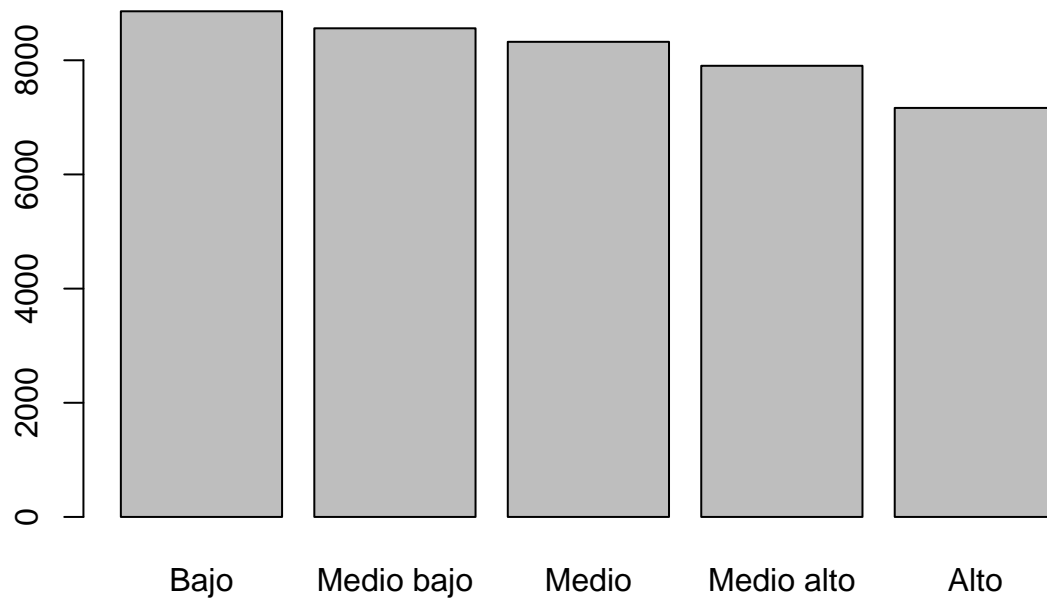
La base de datos tiene 40,809 registros, de los cuales 20,280, poco menos de la mitad, tienen información completa para todos los campos.

```
barplot(c(100, sum(complete.cases(datos))/nrow(datos)*100),
        names.arg = c("Total de registros", "Registros completos"),
        ylab = "%")
```



Con una mezcla más o menos homogénea de niveles socioeconómicos, con una tendencia ligeramente decreciente conforme aumenta el nivel socioeconómico, con entre 7,000 y 9,000 hogares por nivel.

```
plot(datos$nse5f)
```

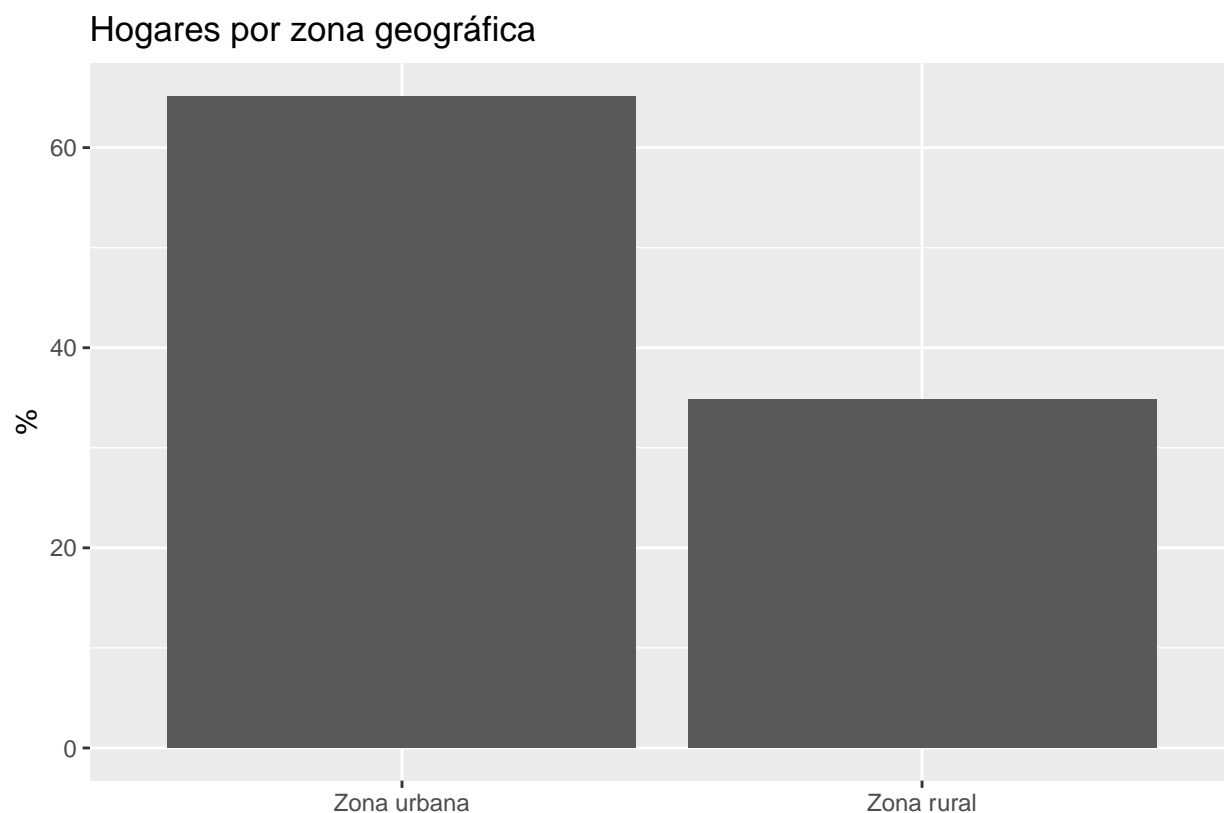


La muestra es mayoritariamente de hogares urbanos, aproximadamente en dos terceras partes.

```
# plot(datos$area)
```

```
ggplot(datos, aes(x = area)) +  
  geom_bar(aes(y = (..count..)/sum(..count..)*100)) +  
  labs(title = "Hogares por zona geográfica",  
        x = "", y = "%")
```

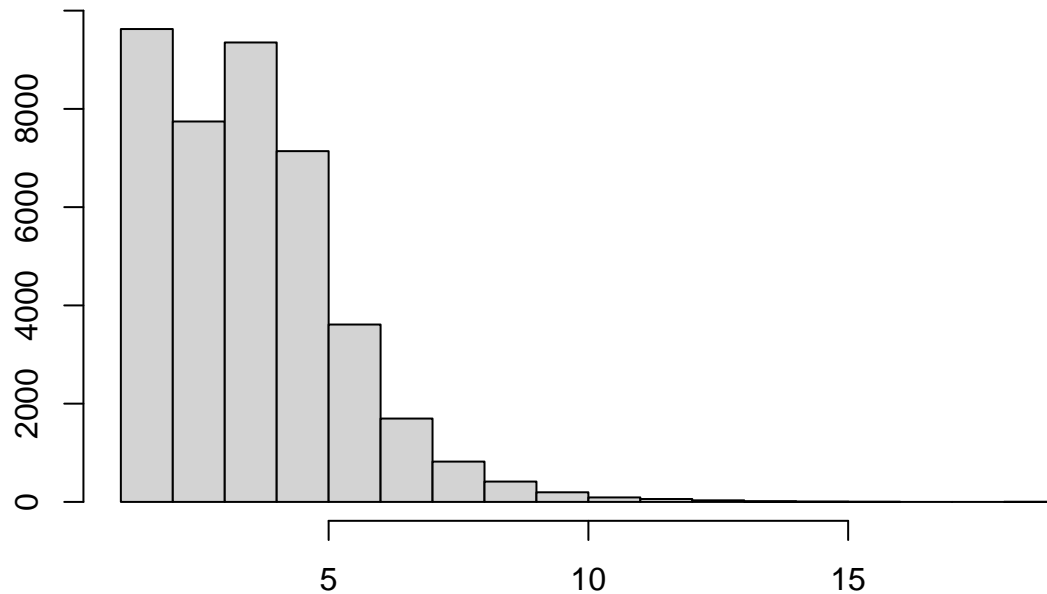
```
## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.  
## i Please use `after_stat(count)` instead.
```



El número de personas por hogar varía entre 1 y 19, con una media de 3.9 y una mediana de 4. El 75% de los hogares es habitado por 5 personas o menos y el 50% por entre 3 y 5 personas. Asimismo, el 95% de los hogares tiene 7 o menos habitantes, siendo la excepción los valores mayores.

```
hist(datos$numpeho,  
      main = "Número de personas en el hogar",  
      xlab = "",  
      ylab = "")
```

## Número de personas en el hogar



```
quantile(datos$numpeho, probs = 0.95)
```

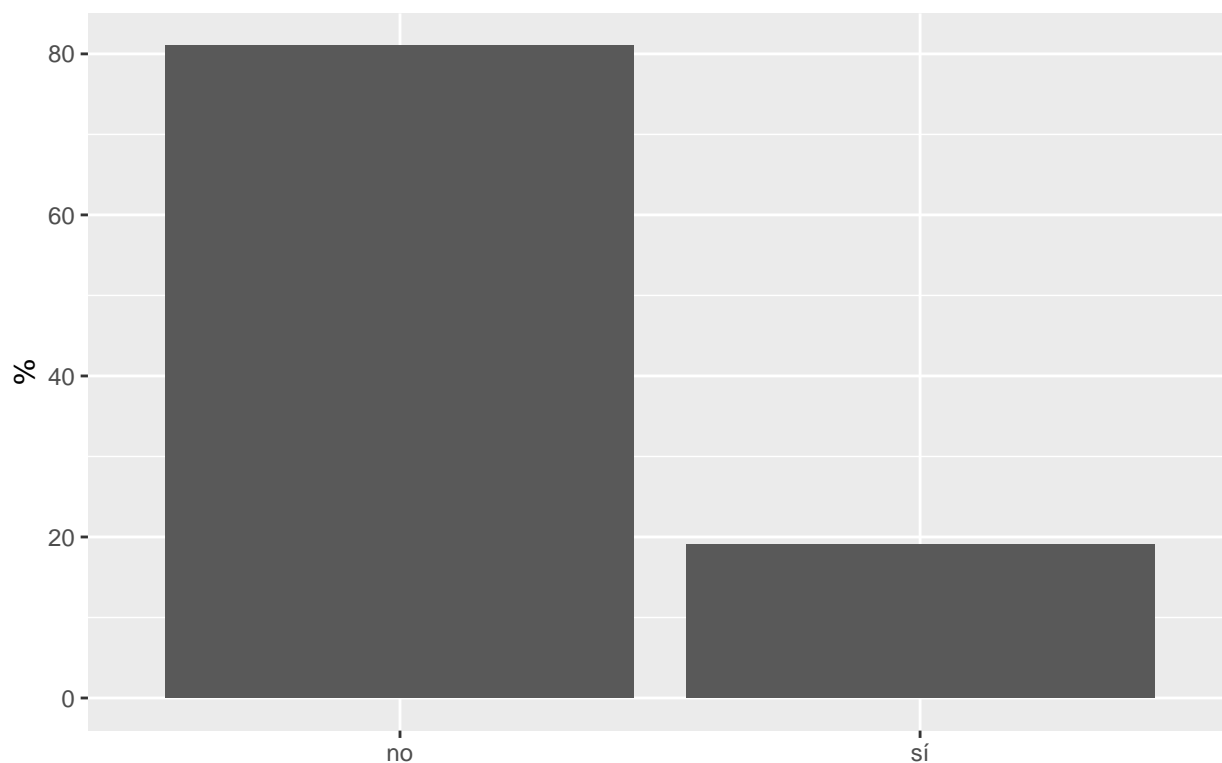
```
## 95%
```

```
## 7
```

En cuanto a recursos financieros adicionales al ingreso, únicamente uno de cada cinco hogares disponen de ellos.

```
#plot(datos$refin,  
#      main = "Disponibilidad de recursos financieros distintos al ingreso")  
  
ggplot(datos, aes(x = refin)) +  
  geom_bar(aes(y = (..count..)/sum(..count..)*100)) +  
  labs(title = "Hogares por disponibilidad de recursos financieros distintos al ingreso",  
        x = "", y = "%")
```

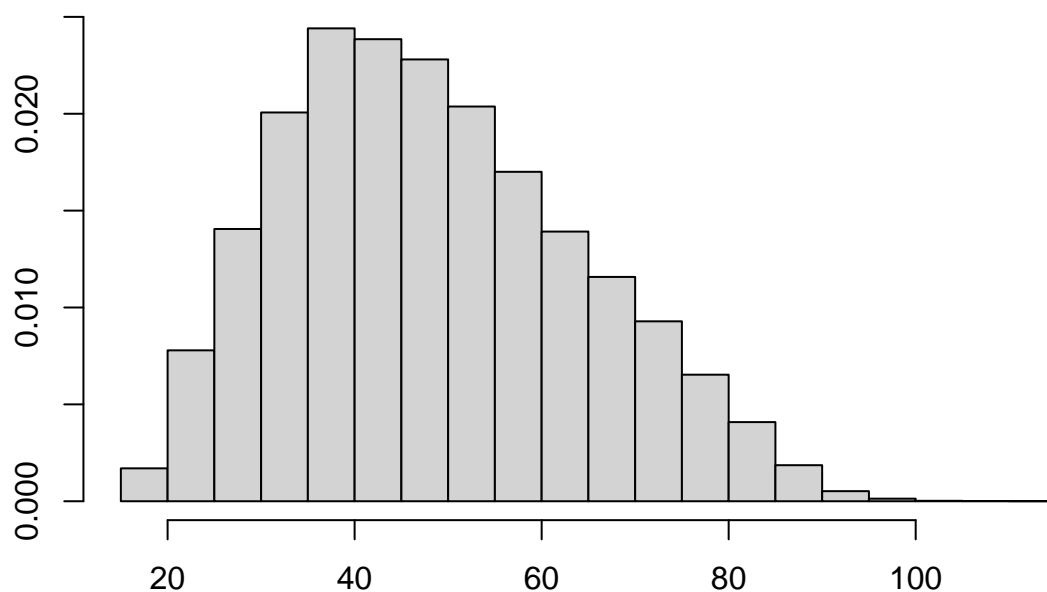
### Hogares por disponibilidad de recursos financieros distintos al ingreso



El promedio de edad del jefe de familia es de 49 años, y la mediana, 47. La edad mínima reportada es de 18 años, y la máxima, de 111 (un posible error de captura). El 50% de las edades de los jefes de familia se ubica entre 37 y 60 años. El 75% de los jefes de familia son menores de 60 años, y el 90%, menores de 72. Para 5,017 de los hogares (12%) no se cuenta con este dato.

```
hist(datos$edadjef,  
      freq = FALSE,  
      main = "Edad del jefe de familia",  
      xlab = "",  
      ylab = "")
```

## Edad del jefe de familia



```
quantile(datos$edadjef, probs = 0.90, na.rm = TRUE)
```

```
## 90%
```

```
## 72
```

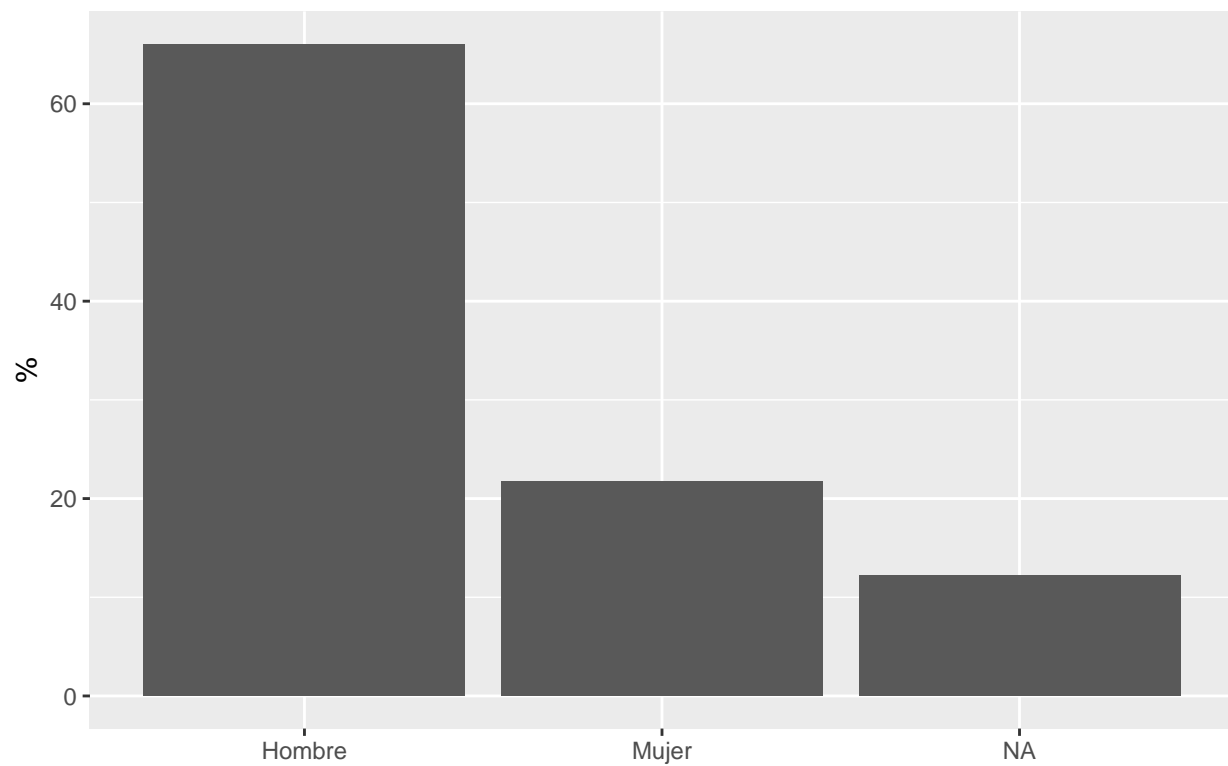
En alrededor del 65% de los hogares, el jefe de familia es hombre, mientras que en poco más de un 20% es mujer. Aproximadamente un 10% no proporcionó este dato.

```
# plot(datos$sexojef)
```

```
ggplot(datos, aes(x = sexojef)) +  
  geom_bar(aes(y = (..count..)/sum(..count..)*100)) +  
  labs(title = "Hogares por sexo del jefe de familia",  
        x = "", y = "%")
```



### Hogares por sexo del jefe de familia

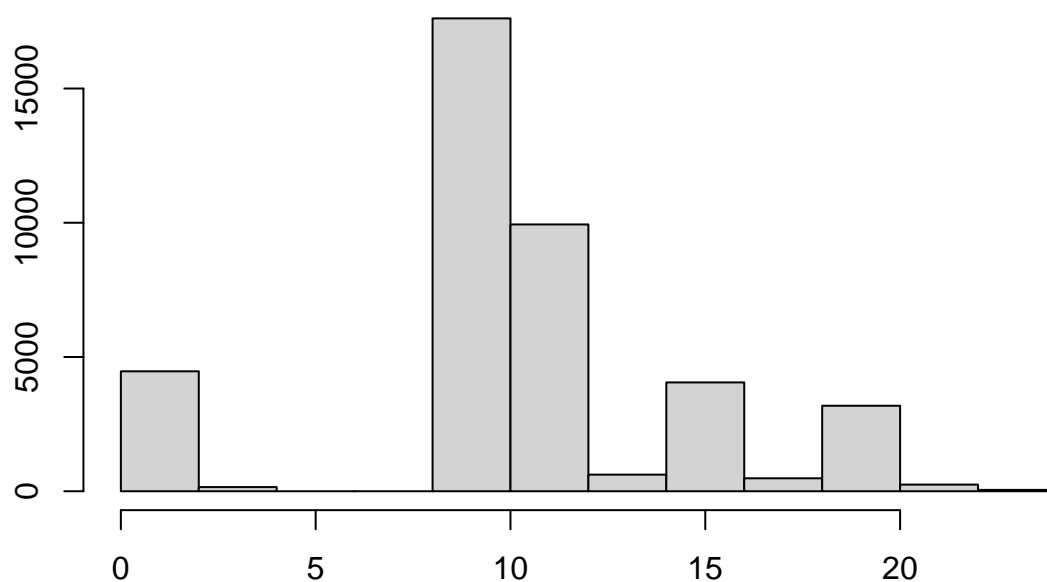


El promedio de años de estudio del jefe del hogar es 10.4 (preparatoria inconclusa), y la mediana, 9 (secundaria completa). Solo el 25% de los jefes de familia cuentan con estudios superiores a la preparatoria. El máximo de años de estudio para el jefe del hogar es de 24.

Es de observarse en el histograma que la mayoría de los datos se ubican entre los 8 y 12 años de estudio, con solo pocos datos fuera de este rango.

```
hist(datos$añosedu,  
      main = "Años de estudio del jefe del hogar",  
      xlab = "",  
      ylab = "")
```

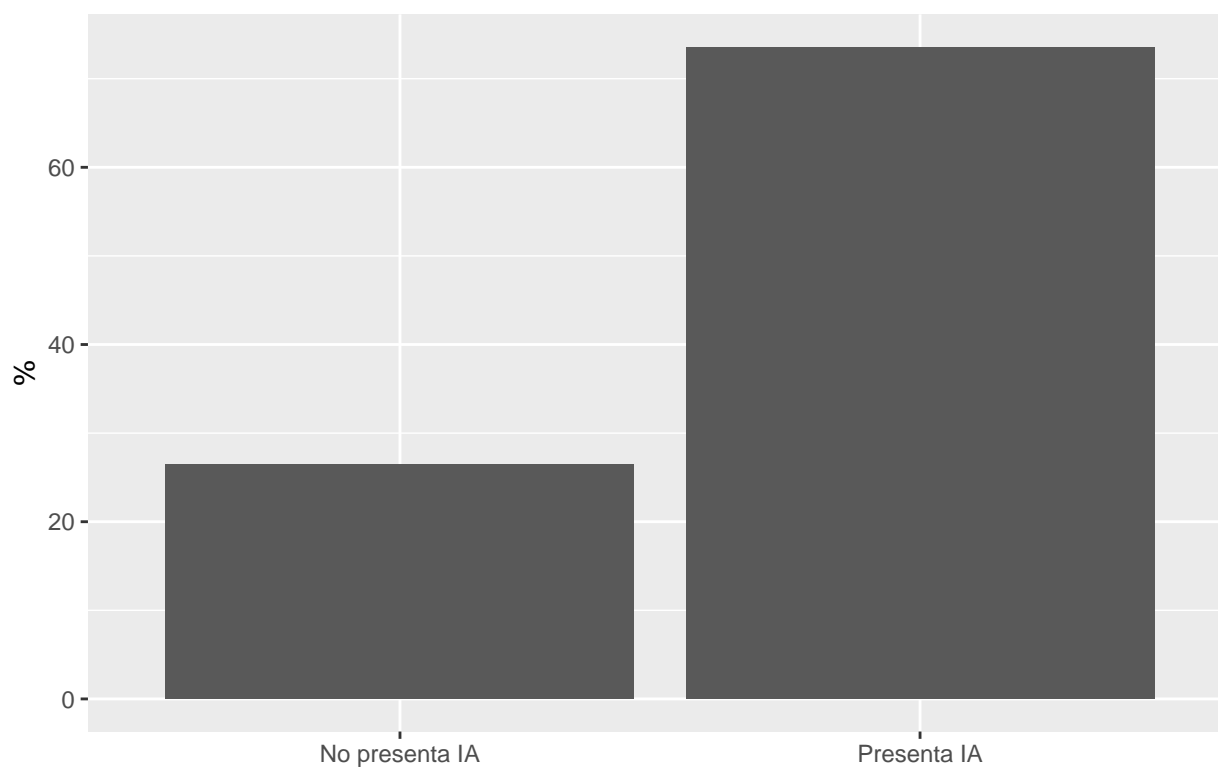
## Años de estudio del jefe del hogar



En lo que respecta a la insuficiencia alimentaria, una cuarta parte de los hogares la presentan.

```
ggplot(datos, aes(x = IA)) +  
  geom_bar(aes(y = (..count..)/sum(..count..)*100)) +  
  labs(title = "Hogares por presencia de insuficiencia alimentaria",  
        x = "", y = "%")
```

## Hogares por presencia de insuficiencia alimentaria



Se observa que los hogares gastan más, en promedio, en alimentos saludables que lo que gastan en alimentos no saludables (en, aproximadamente, casi un orden de magnitud, siete veces más).

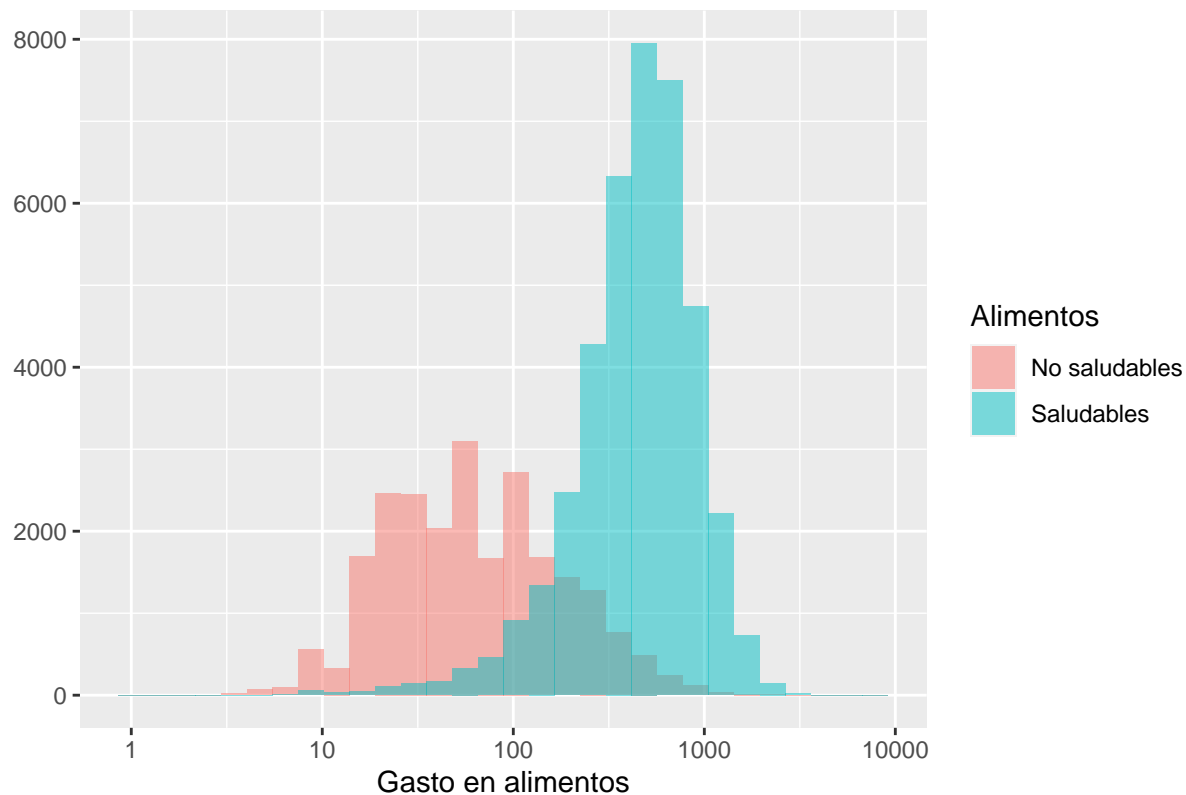
La dispersión del gasto en alimentos saludables es menor que la del gasto en alimentos no saludables. Para los alimentos saludables, el tercer cuartil gasta 2.33 veces más que el primero, mientras que para los alimentos no saludables, el tercer cuartil representa un gasto 38 veces mayor que el primero.

```
als <- data.frame(val = datos$ln_als)
alns <- data.frame(val = datos$ln_alns)
als$Alimentos <- "Saludables"
alns$Alimentos <- "No saludables"
d <- rbind(als, alns)
ggplot(d, aes(exp(val), fill = Alimentos)) +
  geom_histogram(alpha = 0.5, position = "identity") +
  scale_x_log10() +
  labs(title = "Hogares por gasto en alimentos saludables y no saludables",
       x = "Gasto en alimentos",
       y = "")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 18291 rows containing non-finite values (`stat_bin()`).
```

### Hogares por gasto en alimentos saludables y no saludables



```
#hist(datos$ln_als, col = "blue", alpha = 0.4)
#hist(datos$ln_alns, col = "red", alpha = 0.4, add=T)
```

```
exp(mean(datos$ln_als, na.rm = TRUE))
```

```
## [1] 431.1782
```

```
exp(quantile(datos$ln_als, probs = c(0.25, 0.5, 0.75), na.rm = TRUE))
```

```
## 25% 50% 75%  
## 300 475 700
```

```
exp(mean(datos$ln_alns, na.rm = TRUE))
```

```
## [1] 61.86413
```

```
exp(quantile(datos$ln_alns, probs = c(0.25, 0.5, 0.75), na.rm = TRUE))
```

```
## 25% 50% 75%  
## 30 56 130
```

### 3) Algunas probabilidades para un mejor entendimiento del problema

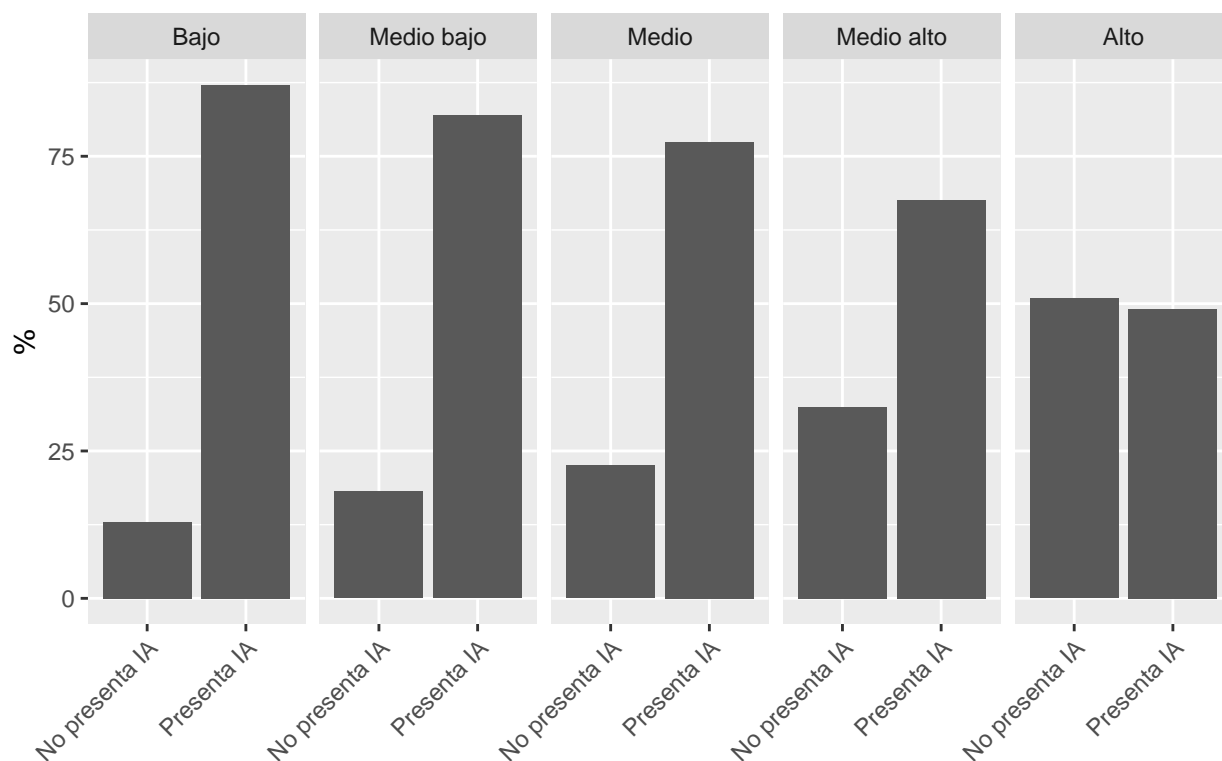
Ya en la sección anterior se hizo una descripción estadística de los datos que, desde el punto de vista frecuentista, es una estimación de las probabilidades asociadas a la población de la que se extrajo la muestra.

La similitud en el número de hogares en cada nivel socioeconómico hace suponer que no se trata de una muestra aleatoria, ya que, de ser ese el caso, se esperaría encontrar una cantidad mucho menor de hogares en los niveles alto y medio alto. Al parecer, se trató de encuestar una cantidad igual o similar de hogares en cada uno de los niveles socioeconómicos, aleatorizando la toma de muestra dentro de cada nivel pero no entre los niveles. Esto se podría corroborar en las notas técnicas de la Encuesta.

Se esperaría que la probabilidad de que un hogar presente inseguridad alimentaria (IA) disminuiría conforme aumente el nivel socioeconómico. Así parece comportarse la muestra pero llama la atención la prevalencia de la IA incluso en el nivel socioeconómico alto, donde la probabilidad de que un hogar presente IA es cercana al 50%.

```
ggplot(datos, aes(x = IA)) +  
  geom_bar(aes(y = (..count..)/tapply(..count.., ..PANEL.., sum)[..PANEL..]*100)) +  
  facet_grid(cols = vars(nse5f)) +  
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +  
  labs(title = "Presencia de inseguridad alimentaria por nivel socioeconómico",  
       x = "", y = "%")
```

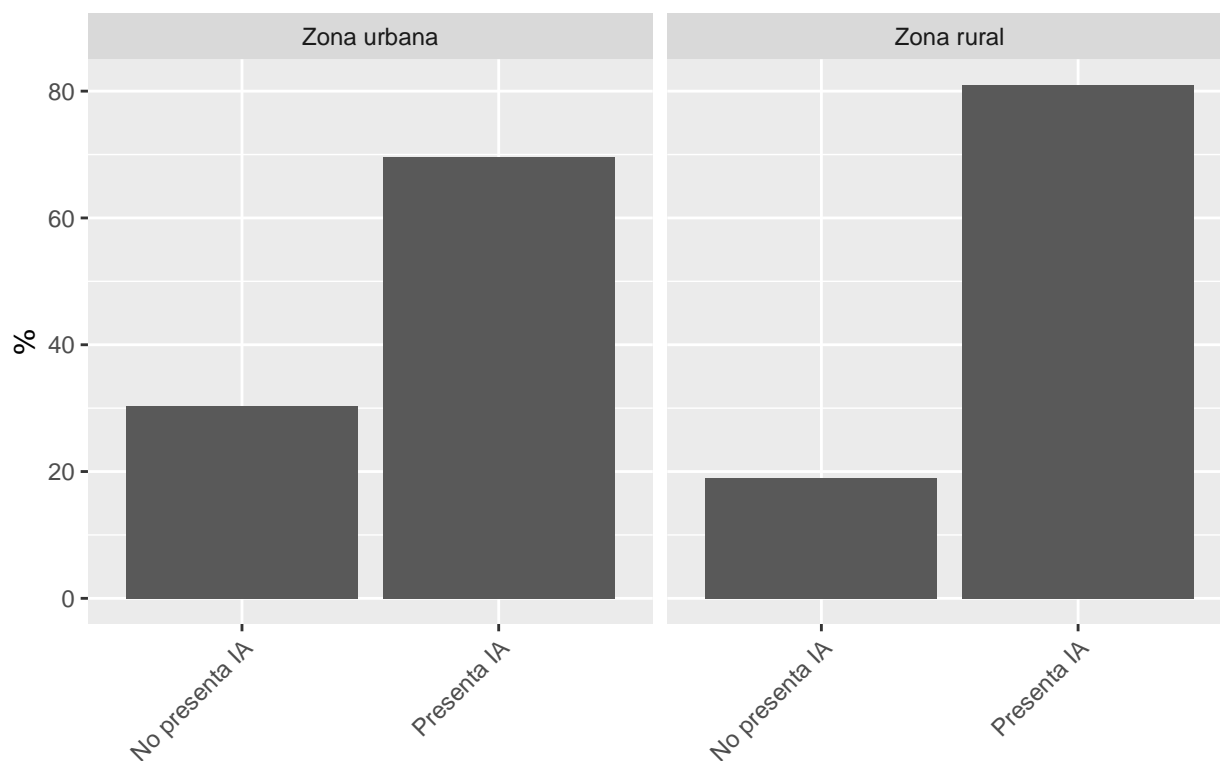
## Presencia de inseguridad alimentaria por nivel socioeconómico



Igualmente, es más probable que un hogar presente IA en las zonas rurales (80%) que en las urbanas (70%).

```
ggplot(datos, aes(x = IA)) +
  geom_bar(aes(y = (..count..)/tapply(..count.., ..PANEL.., sum)[..PANEL..]*100)) +
  facet_grid(cols = vars(area)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Presencia de inseguridad alimentaria por área geográfica",
       x = "", y = "%")
```

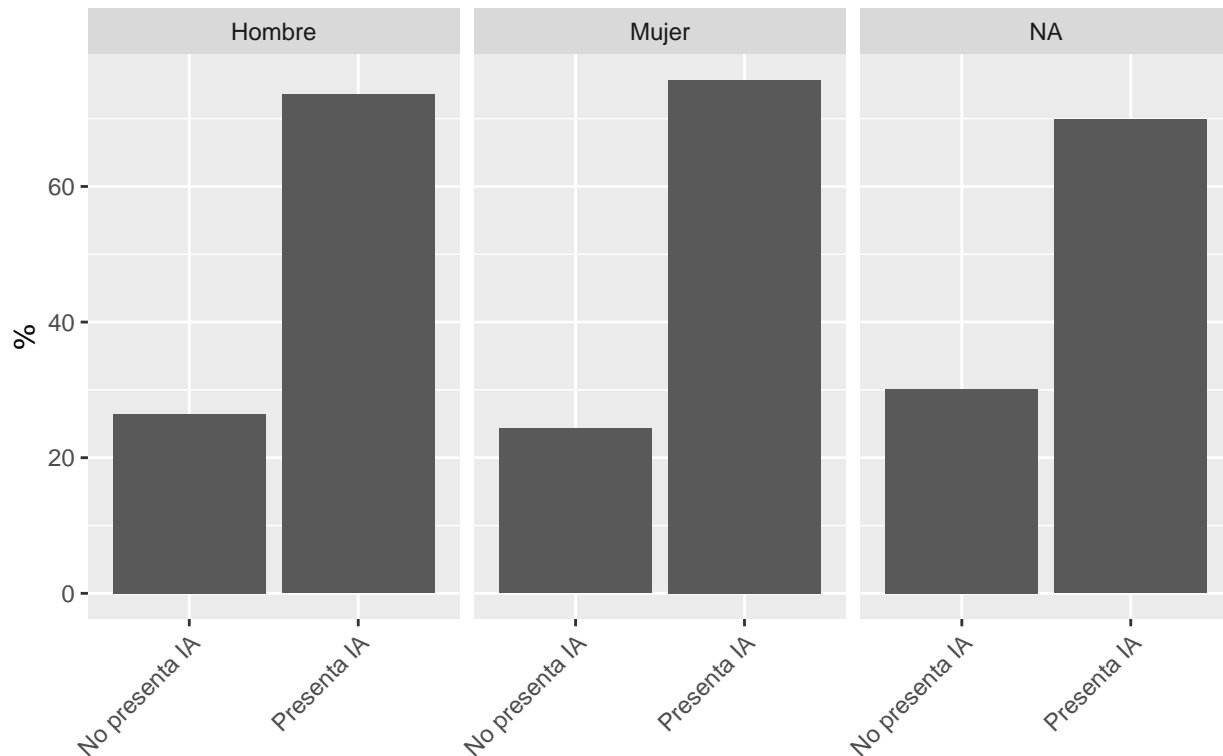
## Presencia de inseguridad alimentaria por área geográfica



La probabilidad de que un hogar presente IA no parece depender de que el jefe de familia sea hombre o mujer.

```
ggplot(datos, aes(x = IA)) +
  geom_bar(aes(y = (..count..)/tapply(..count.., ..PANEL.., sum)[..PANEL..]*100)) +
  facet_grid(cols = vars(sexoief)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(title = "Presencia de inseguridad alimentaria por sexo del jefe de familia",
       x = "", y = "%")
```

## Presencia de inseguridad alimentaria por sexo del jefe de familia



## 5) Modelo de regresión

Dado que el problema es uno de clasificación (presencia o ausencia de inseguridad alimentaria), se utilizará un modelo de regresión logística. Se planea incluir, inicialmente, todas las variables.

En la sección relativa al análisis descriptivo de la información ya se mencionaron los valores faltantes en ciertas variables. La biblioteca **Amelia** permite visualizar gráficamente la densidad de dichos valores faltantes.

```
install.packages("Amelia")
```

```
## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.2'
## (as 'lib' is unspecified)
```

```
library(Amelia)
```

```
## Loading required package: Rcpp
```

```
## ##
```

```
## ## Amelia II: Multiple Imputation
```

```
## ## (Version 1.8.1, built: 2022-11-18)
```

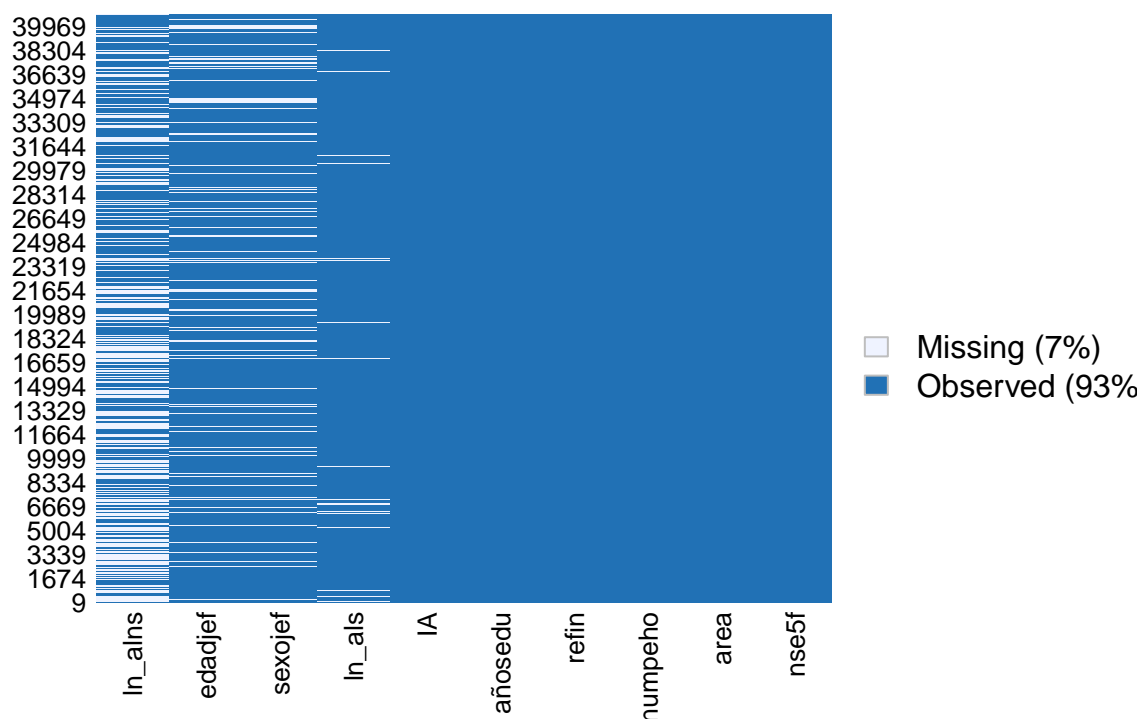
```
## ## Copyright (C) 2005-2022 James Honaker, Gary King and Matthew Blackwell
```

```
## ## Refer to http://gking.harvard.edu/amelia/ for more information
```

```
## ##
```

```
missmap(datos,
  main = "Mapa de valores faltantes (NA)")
```

## Mapa de valores faltantes (NA)



La variable `ln_alns`, que representa el logaritmo natural del gasto en alimentos no saludables, presenta la mayor densidad de datos faltantes (17,504 registros), por lo que es candidata a eliminarla del análisis. Sin embargo, por lo que representa, se considera que puede ser un buen predictor de inseguridad alimentaria, por lo que se decide conservarlo.

El modelo inicial considera todos los factores con sus interacciones de segundo grado.

```
m1 <- glm(IA ~ (nse5f + area + numpeho + refin + edadjef + sexojef + añosedu + ln_als + ln_alns)^2,
          data = datos,
          family = binomial)
summary(m1)
```

```
##
## Call:
## glm(formula = IA ~ (nse5f + area + numpeho + refin + edadjef +
##     sexojef + añosedu + ln_als + ln_alns)^2, family = binomial,
##     data = datos)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.7898  -1.0230   0.5959   0.8108   1.8583
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.6972136   1.2919527   1.314 0.188954
## nse5f.L        -0.3183660   0.5916738  -0.538 0.590524
## nse5f.Q         1.6518814   0.4613047   3.581 0.000342 ***
## nse5f.C         0.1451875   0.4392013   0.331 0.740968
## nse5f^4         0.0799632   0.4462598   0.179 0.857792
## areaZona rural -1.0546597   0.4880544  -2.161 0.030699 *
## numpeho         0.5417532   0.1159581   4.672 2.98e-06 ***
```



## refinsí	0.4578141	0.5372333	0.852	0.394120	
## edadjef	-0.0071308	0.0126042	-0.566	0.571569	
## sexojefMujer	0.9387917	0.4824217	1.946	0.051655	.
## añosedu	-0.1150622	0.0483382	-2.380	0.017296	*
## ln_als	-0.0680134	0.2158779	-0.315	0.752720	
## ln_alns	-0.0977947	0.1801646	-0.543	0.587262	
## nse5f.L:areaZona rural	-0.0124977	0.1164806	-0.107	0.914555	
## nse5f.Q:areaZona rural	0.0666232	0.1007506	0.661	0.508440	
## nse5f.C:areaZona rural	0.0525047	0.0910956	0.576	0.564366	
## nse5f^4:areaZona rural	0.0529577	0.0852440	0.621	0.534436	
## nse5f.L:numpeho	0.0274052	0.0304288	0.901	0.367782	
## nse5f.Q:numpeho	-0.0276508	0.0249315	-1.109	0.267400	
## nse5f.C:numpeho	0.0524637	0.0244943	2.142	0.032204	*
## nse5f^4:numpeho	0.0249756	0.0244252	1.023	0.306529	
## nse5f.L:refinsí	0.1511271	0.1312428	1.152	0.249523	
## nse5f.Q:refinsí	-0.1136760	0.1066436	-1.066	0.286449	
## nse5f.C:refinsí	-0.0393388	0.1016478	-0.387	0.698748	
## nse5f^4:refinsí	0.1047915	0.1023398	1.024	0.305856	
## nse5f.L:edadjef	-0.0169827	0.0036216	-4.689	2.74e-06	***
## nse5f.Q:edadjef	-0.0088625	0.0030040	-2.950	0.003175	**
## nse5f.C:edadjef	-0.0056638	0.0028437	-1.992	0.046405	*
## nse5f^4:edadjef	0.0032353	0.0028289	1.144	0.252766	
## nse5f.L:sexojefMujer	0.0260305	0.1263002	0.206	0.836712	
## nse5f.Q:sexojefMujer	0.0403663	0.1025336	0.394	0.693811	
## nse5f.C:sexojefMujer	0.0820855	0.0973275	0.843	0.399008	
## nse5f^4:sexojefMujer	-0.0181687	0.0936977	-0.194	0.846248	
## nse5f.L:añosedu	-0.0050016	0.0127380	-0.393	0.694576	
## nse5f.Q:añosedu	-0.0189233	0.0111716	-1.694	0.090288	.
## nse5f.C:añosedu	-0.0125503	0.0107862	-1.164	0.244605	
## nse5f^4:añosedu	0.0011633	0.0109061	0.107	0.915055	
## nse5f.L:ln_als	-0.0082607	0.0821837	-0.101	0.919935	
## nse5f.Q:ln_als	-0.1687567	0.0666033	-2.534	0.011285	*
## nse5f.C:ln_als	0.0257662	0.0644167	0.400	0.689162	
## nse5f^4:ln_als	-0.0010046	0.0649899	-0.015	0.987667	
## nse5f.L:ln_alns	0.0088135	0.0517831	0.170	0.864853	
## nse5f.Q:ln_alns	-0.0050211	0.0421020	-0.119	0.905069	
## nse5f.C:ln_alns	-0.0342444	0.0400700	-0.855	0.392764	
## nse5f^4:ln_alns	-0.0772047	0.0392510	-1.967	0.049189	*
## areaZona rural:numpeho	-0.0404855	0.0262376	-1.543	0.122823	
## areaZona rural:refinsí	-0.0191520	0.1111439	-0.172	0.863188	
## areaZona rural:edadjef	0.0065334	0.0030708	2.128	0.033371	*
## areaZona rural:sexojefMujer	0.0036477	0.1078284	0.034	0.973013	
## areaZona rural:añosedu	0.0201767	0.0113925	1.771	0.076552	.
## areaZona rural:ln_als	0.1100942	0.0728684	1.511	0.130823	
## areaZona rural:ln_alns	-0.0233099	0.0445756	-0.523	0.601024	
## numpeho:refinsí	-0.0055732	0.0294025	-0.190	0.849664	
## numpeho:edadjef	-0.0002198	0.0008410	-0.261	0.793812	
## numpeho:sexojefMujer	-0.0120296	0.0259444	-0.464	0.642885	
## numpeho:añosedu	0.0093105	0.0026581	3.503	0.000461	***
## numpeho:ln_als	-0.0615867	0.0155109	-3.971	7.17e-05	***
## numpeho:ln_alns	-0.0116868	0.0109499	-1.067	0.285836	
## refinsí:edadjef	0.0019166	0.0034218	0.560	0.575400	
## refinsí:sexojefMujer	0.0121632	0.1133718	0.107	0.914562	
## refinsí:añosedu	-0.0019808	0.0124443	-0.159	0.873535	

```

## refinsí:ln_als          0.0141131  0.0799644  0.176 0.859908
## refinsí:ln_alns        -0.0468353  0.0461218 -1.015 0.309882
## edadjef:sexojefMujer    -0.0093445  0.0031159 -2.999 0.002709 **
## edadjef:añosedu         0.0012018  0.0002720  4.419 9.93e-06 ***
## edadjef:ln_als         0.0010687  0.0020899  0.511 0.609079
## edadjef:ln_alns        -0.0021093  0.0012869 -1.639 0.101189
## sexojefMujer:añosedu    0.0180239  0.0108268  1.665 0.095963 .
## sexojefMujer:ln_als     -0.0769891  0.0714516 -1.077 0.281257
## sexojefMujer:ln_alns    0.0012746  0.0435174  0.029 0.976634
## añosedu:ln_als         -0.0034459  0.0074780 -0.461 0.644935
## añosedu:ln_alns        -0.0043110  0.0046096 -0.935 0.349670
## ln_als:ln_alns         0.0345713  0.0260568  1.327 0.184586
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
## Null deviance: 24373 on 20279 degrees of freedom
## Residual deviance: 21922 on 20207 degrees of freedom
## (20529 observations deleted due to missingness)
## AIC: 22068
##
## Number of Fisher Scoring iterations: 4

```

Se encuentran los siguientes factores con influencia significativa ( $\alpha = 0.05$ ):

- Nivel socioeconómico
- Área geográfica
- Número de personas en el hogar
- Años de educación del jefe de familia

Así como las siguientes interacciones:

- Nivel socioeconómico con:
  - Número de personas en el hogar
  - Edad del jefe de familia
  - Gasto en alimentos saludables
  - Gasto en alimentos no saludables
- Área geográfica con:
  - Edad del jefe de familia
- Número de personas con:
  - Años de educación del jefe de familia
  - Gasto en alimentos saludables
- Edad del jefe de familia con:
  - Sexo del jefe de familia

Por lo que se desechan las siguientes variables, para las que no se encontraron efectos significativos:

- Disponibilidad de recursos financieros distintos al ingreso

Es de observar que los siguientes factores no presentan un efecto significativo como elementos de primer orden (sin interacción) aunque si lo presentan en el segundo orden (en interacción con otros):

- Edad del jefe de familia
- Sexo del jefe de familia
- Años de educación del jefe de familia
- Gasto en alimentos saludables
- Gasto en alimentos no saludables

Como comparación, se analiza un modelo exclusivamente de primer orden:

```
m0 <- glm(IA ~ nse5f + area + numpeho + refin + edadjef + sexojef + añosedu + ln_als + ln_alns,
          data = datos,
          family = binomial)
summary(m0)

##
## Call:
## glm(formula = IA ~ nse5f + area + numpeho + refin + edadjef +
##      sexojef + añosedu + ln_als + ln_alns, family = binomial,
##      data = datos)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.6972  -1.0549   0.6105   0.8063   1.6641
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    1.740492   0.194519   8.948 < 2e-16 ***
## nse5f.L        -1.135245   0.049522 -22.924 < 2e-16 ***
## nse5f.Q        -0.192516   0.039620  -4.859 1.18e-06 ***
## nse5f.C        -0.097047   0.037984  -2.555 0.010621 *
## nse5f^4         0.017179   0.037188   0.462 0.644107
## areaZona rural -0.089011   0.041162  -2.162 0.030585 *
## numpeho         0.176586   0.010707  16.493 < 2e-16 ***
## refin sí       0.395129   0.044673   8.845 < 2e-16 ***
## edadjef         0.001140   0.001240   0.920 0.357796
## sexojefMujer    0.149462   0.041336   3.616 0.000299 ***
## añosedu        -0.051213   0.004514 -11.345 < 2e-16 ***
## ln_als          -0.091515   0.028626  -3.197 0.001389 **
## ln_alns         -0.097472   0.016921  -5.761 8.39e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 24373  on 20279  degrees of freedom
## Residual deviance: 22117  on 20267  degrees of freedom
## (20529 observations deleted due to missingness)
## AIC: 22143
##
## Number of Fisher Scoring iterations: 4
```

En el modelo de primer orden, la disponibilidad de recursos financieros distintos del ingreso sí es un factor significativo, y el sexo y los años de educación del jefe de familia y el gasto en alimentos saludables y no saludables son significativos por sí mismos. Esto se explica porque la variabilidad observada es mejor explicada por las interacciones de los factores una vez que se toma en cuenta.

Descartando los factores e interacciones que se encontraron no significativas, tenemos el siguiente modelo:

```
m2 <- glm(IA ~ nse5f + area + numpeho + añosedu +
          nse5f:numpeho + nse5f:edadjef + nse5f:ln_als + nse5f:ln_alns +
          area:edadjef +
          numpeho:añosedu + numpeho:ln_als +
          edadjef:sexojef,
```

```

      data = datos,
      family = binomial)
summary(m2)

##
## Call:
## glm(formula = IA ~ nse5f + area + numpeho + añosedu + nse5f:numpeho +
##      nse5f:edadjef + nse5f:ln_als + nse5f:ln_alns + area:edadjef +
##      numpeho:añosedu + numpeho:ln_als + edadjef:sexojef, family = binomial,
##      data = datos)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.7664  -1.0554   0.6061   0.8156   1.7634
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.8219783  0.3688648   2.228 0.025854 *
## nse5f.L          -0.3321930  0.4259754  -0.780 0.435485
## nse5f.Q           1.2918995  0.4172007   3.097 0.001958 **
## nse5f.C          -0.0727779  0.4018773  -0.181 0.856293
## nse5f^4           0.1049406  0.4100867   0.256 0.798029
## areaZona rural   -0.3174145  0.1303123  -2.436 0.014859 *
## numpeho           0.4738410  0.0956439   4.954 7.26e-07 ***
## añosedu          -0.0759381  0.0095680  -7.937 2.08e-15 ***
## nse5f.L:numpeho   0.0439988  0.0265634   1.656 0.097647 .
## nse5f.Q:numpeho  -0.0255851  0.0241963  -1.057 0.290329
## nse5f.C:numpeho   0.0579313  0.0237367   2.441 0.014664 *
## nse5f^4:numpeho   0.0271675  0.0238362   1.140 0.254387
## nse5fBajo:edadjef  0.0070260  0.0035727   1.967 0.049231 *
## nse5fMedio bajo:edadjef 0.0028840  0.0029545   0.976 0.328991
## nse5fMedio:edadjef  0.0057022  0.0027552   2.070 0.038490 *
## nse5fMedio alto:edadjef -0.0017194  0.0024476  -0.702 0.482384
## nse5fAlto:edadjef -0.0116872  0.0024773  -4.718 2.39e-06 ***
## nse5fBajo:ln_als   0.0554527  0.0835163   0.664 0.506706
## nse5fMedio bajo:ln_als 0.1900539  0.0762637   2.492 0.012700 *
## nse5fMedio:ln_als   0.2072624  0.0781700   2.651 0.008015 **
## nse5fMedio alto:ln_als 0.1346266  0.0751376   1.792 0.073175 .
## nse5fAlto:ln_als   0.0029850  0.0781621   0.038 0.969536
## nse5fBajo:ln_alns  -0.0981744  0.0526361  -1.865 0.062160 .
## nse5fMedio bajo:ln_alns -0.0778851  0.0437733  -1.779 0.075193 .
## nse5fMedio:ln_alns  -0.1419623  0.0388588  -3.653 0.000259 ***
## nse5fMedio alto:ln_alns -0.0266347  0.0333335  -0.799 0.424268
## nse5fAlto:ln_alns  -0.1282665  0.0308779  -4.154 3.27e-05 ***
## areaZona rural:edadjef 0.0043590  0.0026721   1.631 0.102831
## numpeho:añosedu     0.0072233  0.0022500   3.210 0.001326 **
## numpeho:ln_als     -0.0592054  0.0144024  -4.111 3.94e-05 ***
## edadjef:sexojefMujer 0.0021714  0.0008149   2.664 0.007712 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 24373  on 20279  degrees of freedom

```

```
## Residual deviance: 22097 on 20249 degrees of freedom
## (20529 observations deleted due to missingness)
## AIC: 22159
##
## Number of Fisher Scoring iterations: 4
```

En este nuevo modelo, pierde significancia el termino de interacción entre el área geográfica y la edad del jefe de familia, por lo que se descarta:

```
m3 <- update(m2, ~ . - area:edadjef)
summary(m3)
```

```
##
## Call:
## glm(formula = IA ~ nse5f + area + numpeho + añosedu + nse5f:numpeho +
##       nse5f:edadjef + nse5f:ln_als + nse5f:ln_alns + numpeho:añosedu +
##       numpeho:ln_als + edadjef:sexojef, family = binomial, data = datos)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.7505  -1.0541   0.6071   0.8155   1.7596
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.7598950  0.3668154   2.072 0.038303 *
## nse5f.L          -0.2324831  0.4216156  -0.551 0.581353
## nse5f.Q           1.2827956  0.4172400   3.074 0.002109 **
## nse5f.C          -0.0708361  0.4019901  -0.176 0.860126
## nse5f^4           0.1066218  0.4101872   0.260 0.794914
## areaZona rural   -0.1156758  0.0413340  -2.799 0.005133 **
## numpeho           0.4714338  0.0956059   4.931 8.18e-07 ***
## añosedu          -0.0761732  0.0095616  -7.967 1.63e-15 ***
## nse5f.L:numpeho   0.0448250  0.0265511   1.688 0.091363 .
## nse5f.Q:numpeho  -0.0259801  0.0241894  -1.074 0.282810
## nse5f.C:numpeho   0.0579391  0.0237322   2.441 0.014632 *
## nse5f^4:numpeho   0.0272747  0.0238303   1.145 0.252399
## nse5fBajo:edadjef  0.0097779  0.0031518   3.102 0.001920 **
## nse5fMedio bajo:edadjef 0.0048760  0.0026904   1.812 0.069932 .
## nse5fMedio:edadjef 0.0069925  0.0026398   2.649 0.008075 **
## nse5fMedio alto:edadjef -0.0008558  0.0023900  -0.358 0.720301
## nse5fAlto:edadjef -0.0112910  0.0024648  -4.581 4.63e-06 ***
## nse5fBajo:ln_als   0.0577546  0.0834191   0.692 0.488722
## nse5fMedio bajo:ln_als 0.1904600  0.0762627   2.497 0.012510 *
## nse5fMedio:ln_als  0.2071275  0.0781794   2.649 0.008064 **
## nse5fMedio alto:ln_als 0.1335127  0.0751363   1.777 0.075578 .
## nse5fAlto:ln_als   0.0012584  0.0781452   0.016 0.987152
## nse5fBajo:ln_alns -0.0976566  0.0526258  -1.856 0.063499 .
## nse5fMedio bajo:ln_alns -0.0763185  0.0437636  -1.744 0.081180 .
## nse5fMedio:ln_alns -0.1409705  0.0388505  -3.629 0.000285 ***
## nse5fMedio alto:ln_alns -0.0258937  0.0333265  -0.777 0.437176
## nse5fAlto:ln_alns -0.1281595  0.0308781  -4.151 3.32e-05 ***
## numpeho:añosedu    0.0072565  0.0022484   3.227 0.001249 **
## numpeho:ln_als     -0.0589928  0.0143998  -4.097 4.19e-05 ***
## edadjef:sexojefMujer 0.0021116  0.0008141   2.594 0.009497 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 24373  on 20279  degrees of freedom
## Residual deviance: 22100  on 20250  degrees of freedom
##    (20529 observations deleted due to missingness)
## AIC: 22160
##
## Number of Fisher Scoring iterations: 4
```

En este último modelo, todos los términos tienen una significancia superior al 0.05.

Es de hacer notar que ciertos niveles socioeconómicos (variable categórica) no inciden sobre la posibilidad de que el hogar presente inseguridad alimentaria.

Los coeficientes estimados representan los momios asociados a las variables (posibilidad de que se presente inseguridad alimentaria dividida entre la posibilidad de que no se presente). Un momio negativo representa una disminución de la probabilidad, y uno positivo, un incremento en la misma.

Así, por ejemplo, vivir en una zona rural disminuye el momio de incidencia de la inseguridad alimentaria en un factor de 0.1157, mientras que por cada persona adicional que viva en un hogar, el momio se incrementa en un factor de 0.47.

## 6) Conclusiones

- Sorpresivamente, el incremento en el nivel socioeconómico del hogar aumenta la probabilidad de que se presente inseguridad alimentaria (momio positivo para el contraste cuadrático de `nse5f`).
- En los hogares rurales es menos probable que se presente inseguridad alimentaria.
- El incremento en el número de personas que viven en un hogar aumenta la probabilidad de que se presente inseguridad alimentaria.
- A mayor grado de estudios del jefe de familia, menor probabilidad de que su hogar presente inseguridad alimentaria.
- El incremento en la probabilidad de inseguridad alimentaria por el número de personas, se potencia con el incremento en el nivel socioeconómico del hogar (interacción cúbica `nse5f.C:numpeho`).
- El incremento en la edad del jefe de familia aumenta la probabilidad de que el hogar presente inseguridad alimentaria en los niveles socioeconómicos bajo y medio, pero la disminuye en el nivel alto.
- Sorpresivamente también, el incremento en el gasto en alimentos *saludables* aumenta la probabilidad de inseguridad alimentaria en hogares de nivel socioeconómico medio bajo y medio, mientras que el aumento en el gasto en alimentos *no saludables* lo disminuye en los niveles socioeconómicos medio y alto.
- No están claros los efectos de la interacción entre el número de personas en el hogar y los años de educación del jefe de familia y el gasto en alimentos saludables.
- A mayor edad del jefe de familia, se incrementa el riesgo de que el hogar presente inseguridad alimentaria cuando el jefe de familia es mujer.