

Trabajo 1

Predicción del número de hijos de los hogares colombianos

Integrantes

- Santiago Franco Valencia
- Stephany Lobo
- Isabela Lujan Jaramillo
- Daniel Alexander Naranjo Ríos
- Ana María Sánchez Henao

Objetivo

El objetivo de este trabajo es predecir el número de hijos que hay en los hogares colombianos a partir de características específicas partiendo de la encuesta de calidad de vida más reciente publicada por el DANE (2019).

Introducción

Las encuestas dirigidas a hogares constituyen una de las principales fuentes de datos socioeconómicos con las que cuentan los países. A partir de la información obtenida de ellas se calculan indicadores para la medición de variados aspectos económicos y sociales. Además, facilitan el conocimiento y explicación de los determinantes o factores causales del comportamiento de dichos aspectos, lo cual es de gran importancia para el diseño, monitoreo y medición de resultados de las políticas públicas.” (DANE, 2018). Por ello son herramientas muy usadas actualmente por el gobierno, aprovechando la alta capacidad que tienen para recoger grandes cantidades de información rápidamente, sobre todo cuando dicha información proviene de la percepción que tiene el pueblo o las condiciones en las que cada ciudadano vive. Para ello, se crearon instituciones como el departamento administrativo nacional de estadísticas (DANE), que se encarga de planear, recopilar, procesar, analizar y difundir las estadísticas oficiales referentes a información de interés referente a varias situaciones actuales del país y su población.

Contextualización de la encuesta

Las encuestas a hogares normalmente se centran en temas específicos como, por ejemplo, la Encuesta de Mercado Laboral (Gran Encuesta Integrada de Hogares) que se aplica en forma regular y continua durante todo el año. Otro ejemplo es la Encuesta Nacional de Presupuestos de los Hogares que se aplica cada diez años. Si bien estas encuestas indagan sobre algunos aspectos que permiten hacer análisis particulares del bienestar, no brindan información que posibilite conocer íntegramente las diferentes variables que determinan las condiciones de vida del hogar en todas sus dimensiones.” (DANE, 2018). Estas encuestas incluso abarcan temas como la calidad de vida de los colombianos, como lo hace la encuesta nacional de calidad de vida, la cual es realizada por el DANE cada año aproximadamente, y que será la base de donde se extraerá la información para este trabajo. Tomando en cuenta cada una de las variables presentadas en la misma y depurando aquellas que no sean de interés para el objetivo final.

Problemas encontrados

Variables a considerar

La encuesta se encuentra conformada por 14 bases de datos en total. Para la selección de variables se realizó un estudio riguroso de las variables a considerar teniendo en cuenta características específicas del hogar y del jefe de hogar. La base de datos tiene 93993 registros y 57 variables. A continuación un listado de las variables seleccionadas por cada base de datos.

Datos de la vivienda

- Alcantarillado: (1. Sí - 2. No)
- Acueducto: (1. Sí - 2. No)
- Estrato para tarifa:
 1. Bajo - Bajo
 2. Bajo
 3. Medio - Bajo
 4. Medio
 5. Medio - Alto
 6. Alto
 7. Planta eléctrica
 8. No conoce el estrato o no cuenta con recibo de pago.
 9. Recibos sin estrato o el servicio es pirata
- Tipo de vivienda:
 1. Casa
 2. Apartamento
 3. Cuarto(s)
 4. Vivienda tradicional indígena
 5. Otro (carpa, contenedor, vagón, embarcación, cueva, refugio natural, etc)
- Región:
 1. Caribe
 2. Oriental
 3. Central
 4. Pacífica(sin valle)
 5. Bogotá
 6. Antioquia
 7. Valle del cauca
 8. San Andrés
 9. Orinoquía - amazonía

Servicios del hogar (PENDIENTE)

- Número de cuartos que dispone el hogar
- Número de dormitorios que dispone el hogar
- Cortes o suspensiones de energía por falta de pago: (1. Sí - 2. No)
- Cambios bruscos del voltaje
- Bajo voltaje
- Iluminación principal
- Servicio de gas natural
- Tipo servicio sanitario
- Ubicación servicio sanitario
- Lavadero

- Lavamanos
- Lavaplatos
- Jabón de manos
- Clasificación basura
- Bombillas bajo consumo
- Plancha
- Reutiliza agua
- Recolecta agua lluvia
- Tanque sanitario bajo consumo
- Obtención de agua para alimentos
- Servicio agua 24 horas
- Ubicación preparación energía
- Tipo de energía
- Cantidad de personas en el hogar #### Características y composición del hogar
- Conyuges:
 0. No tiene
 1. Sí vive en el hogar
 2. No vive en el hogar
- El padre vive en el hogar:
 1. Sí
 2. No
 3. Fallecido
- Educación padre:
 1. Primaria incompleta
 2. Primaria completa
 3. Secundaria incompleta
 4. Secundaria completa
 5. Técnica incompleta
 6. Técnica completa
 7. Universidad incompleta
 8. Universidad completa
 9. Ninguno
 10. No sabe
- La madre vive en el hogar:
 1. Sí
 2. No
 3. Fallecida
- Educación madre:
 1. Primaria incompleta
 2. Primaria completa
 3. Secundaria incompleta
 4. Secundaria completa
 5. Técnica incompleta
 6. Técnica completa
 7. Universidad incompleta
 8. Universidad completa
 9. Ninguno
 10. No sabe
- Cultura

1. Indígena
 2. Gitano (a) (Rom)
 3. Raizal del archipiélago de San Andrés, Providencia y Santa Catalina
 4. Palenquero (a) de San Basilio
 5. Negro (a), mulato (a) (afrodescendiente), afrocolombiano(a)
 6. Ninguno de los anteriores
- Es campesino (1. Sí - 2. No)
 - Condición de vida 0 significa que se siente “totalmente insatisfecho” y 10 significa que se siente “totalmente satisfecho”.

Tenencia y financiación de la vivienda que ocupa el hogar

- La vivienda ocupada por este hogar es:
1. Propia, totalmente pagada
 2. Propia, la están pagando
 3. En arriendo o subarriendo
 4. Con permiso del propietario, sin pago alguno (usufructuario)
 5. Posesión sin título (ocupante de hecho)
 6. Propiedad colectiva
- ¿Cuánto pagan mensualmente por arriendo?

Uso de energéticos del hogar

- Número de computadores en el hogar

Fuerza de trabajo

- Exige mucho esfuerzo físico
- Exige mucho esfuerzo intelectual

¿Cómo se encontró y?

Análisis descriptivo

El primer filtro para descartar alguna de las variables consiste en graficar boxplots para variables categóricas y gráficos de dispersión para las variables numéricas, al ser tantas variables se mostrarán algunos ejemplos de las variables que descartamos y las que por su resultado, se considerarán en el modelo.

Variables categóricas

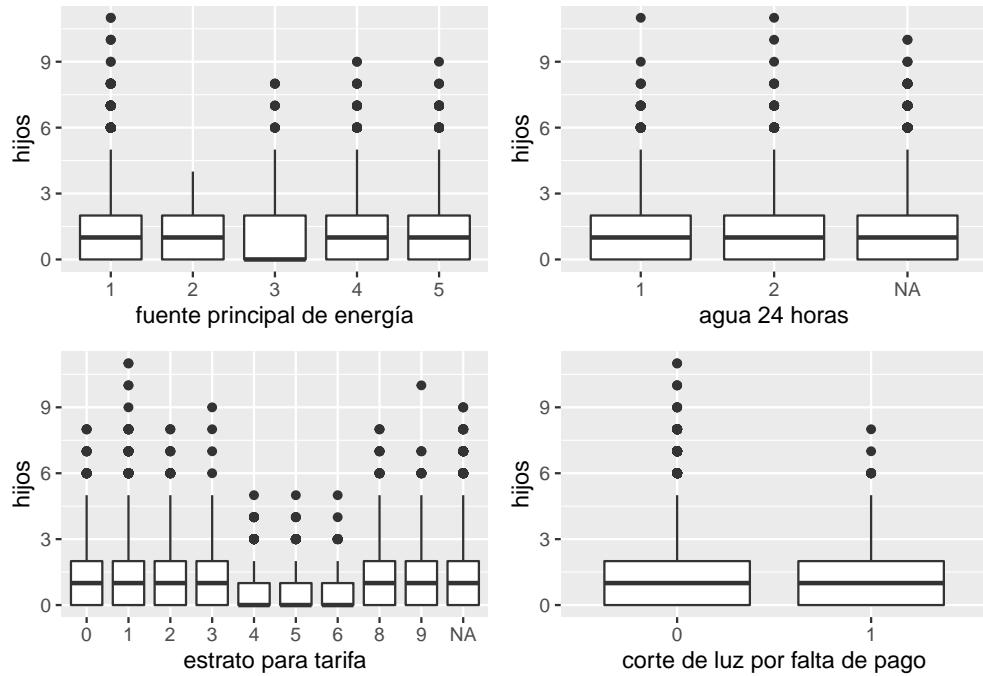
Se encontraron muchas variables que no son significativas a la hora de graficarlas ya que su media permanecía constante al número de hijos y se pudo comprobar que no aportaban mucha información, un ejemplo de estos gráficos se encuentra a continuación:

```
##### gráficos juntos #####
# fuente principal de energía
a <- ggplot(aes(y= hijos, x = ilum_ppal), data=datos) + geom_boxplot() +
  xlab("fuente principal de energía")
# agua 24 horas
b <- ggplot(aes(y= hijos, x = agua_24h), data=datos) + geom_boxplot() +
  xlab("agua 24 horas")
# estrato
c <- ggplot(aes(y= hijos, x = estrato), data=datos) + geom_boxplot() +
```

```

xlab("estrato para tarifa")
# corte luz
d <- ggplot(aes(y= hijos, x = corte_energia_fp), data=datos) + geom_boxplot() + xlab("corte de luz por...")
grid.arrange(a, b,c,d)

```



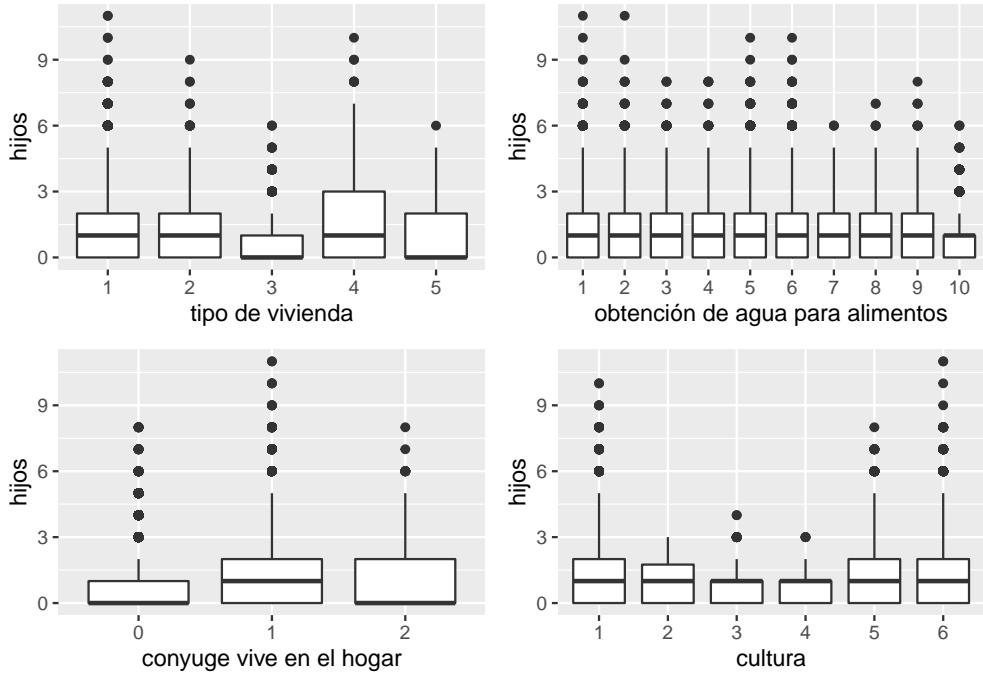
En estos gráficos se observa que para variables como fuente principal de energía, si la vivienda tenía agua las 24 horas del día, estrato por tarifa y algún corte de luz en los últimos 30 días no aportan mucha información para el número de hijos, contrario a lo que se pensaba, pues son variables directamente relacionadas con la calidad de vida, siguiendo este análisis, así fueron removidas algunas variables para implementar un modelo con las variables más cercanas.

Por otro lado, también se encontraron variables que aunque su variación no es mucha, se consideró prudente dejarlas para la implementación del modelo, pues se llegó a la conclusión que eran importantes a la hora de evaluar el número de hijos. Alguna de estas variables son:

```

# boxplot tipo de vivienda
e <- ggplot(aes(y= hijos, x =tipo_vivienda), data=datos) + geom_boxplot() + xlab("tipo de vivienda")
# obtención agua
f <- ggplot(aes(y= hijos, x = obtencion_agua_alimento), data=datos) + geom_boxplot() + xlab("obtención")
# conyugues
g <- ggplot(aes(y= hijos, x = conyuges), data=datos) + geom_boxplot() +
  xlab("conyuge vive en el hogar")
# cultura
h <- ggplot(aes(y= hijos, x =cultura), data=datos) + geom_boxplot()
grid.arrange(e,f,g,h)

```



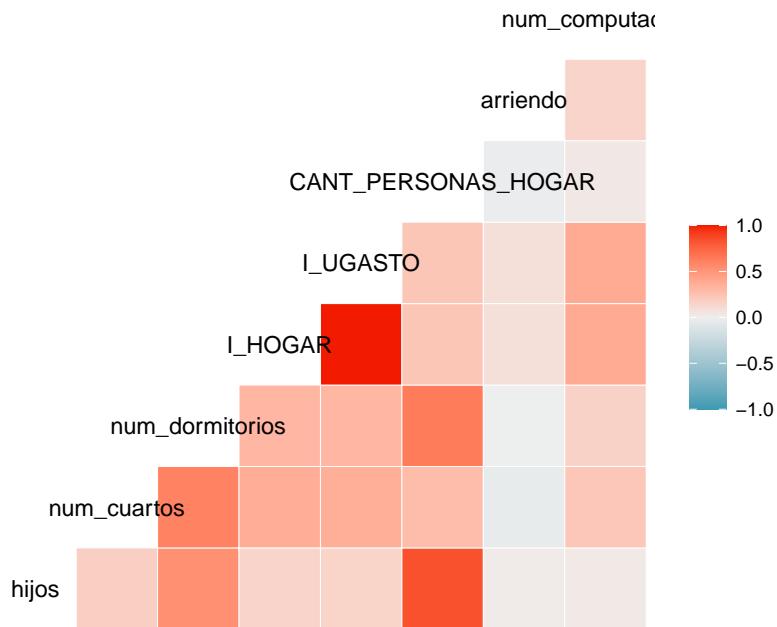
Como grupo, se encontró que se podían pasar otras variables al siguiente filtro, el cual será el modelo en sí, pues aunque se piense que son necesarias solamente el modelo podrá decir qué tan significativas son y si efectivamente se deben retirar.

Variables continuas

Para las variables continuas se utilizó la matriz de correlación.

```
ggcorr(datos[,-2], method = c("pairwise", "spearman"))

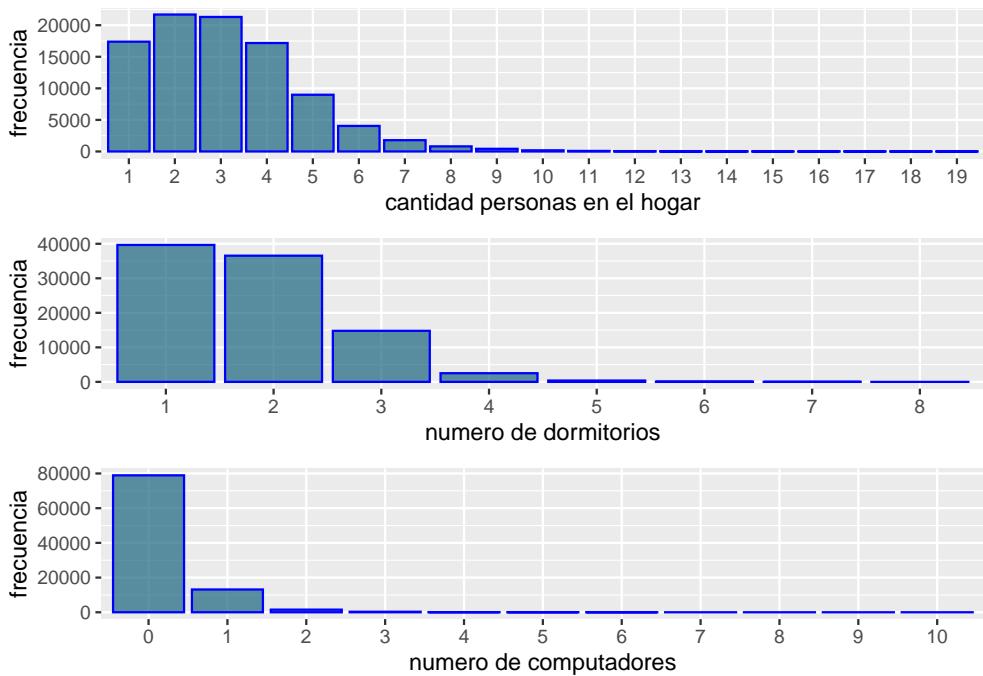
## Warning in ggcrrr(datos[, -2], method = c("pairwise", "spearman")): data in
## column(s) 'indice', 'corte_energia_fp', 'corte_energia_om', 'cambios_voltaje',
## 'bajo_voltaje', 'ninguna', 'ilum_ppal', 'gas_natural_rp', 'tipo_serv_sanitario',
## 'ubic_serv_sanitario', 'serv_sanitario_es', 'lavamanos', 'lavadero',
## 'lavaplatos', 'ninguno', 'jabon_manos', 'basura_hogar', 'clasifica_baura',
## 'bombilla_bajo_consumo', 'apaga_luces', 'planchar', 'desconecta_aparatos',
## 'reutiliza_agua', 'recolecta_agua_lluvia', 'tanque_sanitario_bj',
## 'econom_agua', 'obtencion_agua_alimento', 'agua_24h', 'obtencion_agua_bebida',
## 'ubic_prepar_alimento', 'tipo_energia', 'conyuges', 'padre_hogar',
## 'educacion_padre', 'madre_hogar', 'educacion_madre', 'cultura', 'es_campesino',
## 'condicion_vida', 'region', 'tipo_vivienda', 'energia', 'estrato', 'acueducto',
## 'alcantarillado', 'vivienda_ocupada', 'fisico_t', 'intelectual_t' are not
## numeric and were ignored
```



Como se puede observar en la matriz de correlación, la variable que más se relaciona con número de hijos es la cantidad de personas en el hogar, y muy levemente el número de dormitorios también junto con el ingreso del hogar y el gasto promedio. Aunque las relaciones no son muy fuertes, se dejarán para el siguiente filtro y observar si la poca relación que muestra el gráfico en realidad es significativa.

Otro tipo de gráficos

```
# cantidad personas
i <- ggplot(datos, aes(x=as.factor(CANT_PERSONAS_HOGAR) )) +
  geom_bar(color="blue", fill=rgb(0.1,0.4,0.5,0.7) ) +
  xlab("cantidad personas en el hogar") + ylab("frecuencia")
# número dormitorios
j <- ggplot(datos, aes(x=as.factor(num_dormitorios) )) +
  geom_bar(color="blue", fill=rgb(0.1,0.4,0.5,0.7) ) +
  xlab("numero de dormitorios") + ylab("frecuencia")
# número computadores
k <- ggplot(datos, aes(x=as.factor(num_computadores) )) +
  geom_bar(color="blue", fill=rgb(0.1,0.4,0.5,0.7) )+
  xlab("numero de computadores") + ylab("frecuencia")
grid.arrange(i,j,k)
```

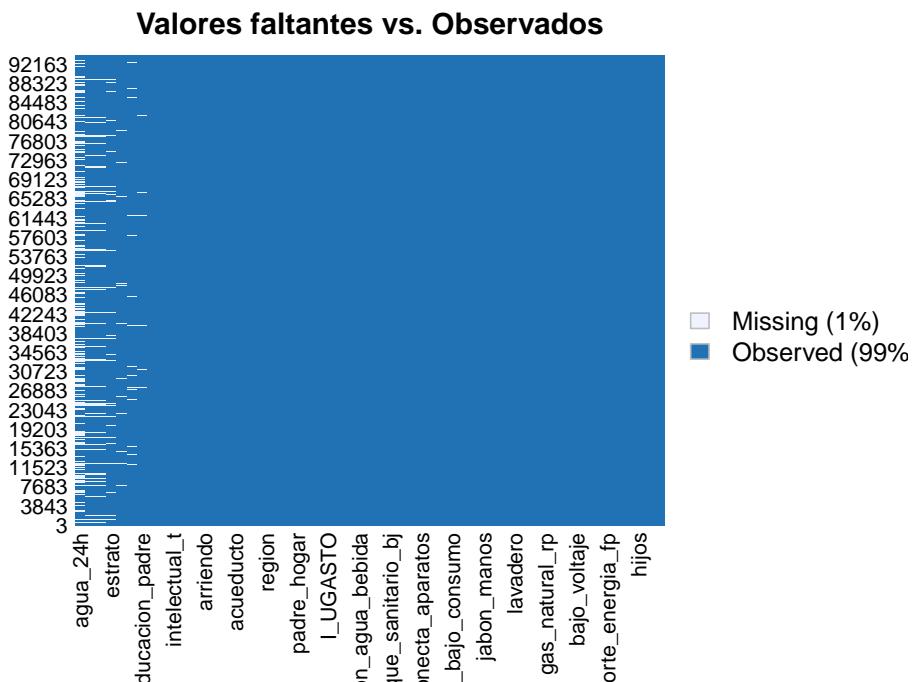


La mayoría de hogares están conformados por entre dos y tres personas, también se encontró que el número de cuartos más recurrente es uno, y al preguntar por cantidad de computadores en el hogar, los gráficos indican que la mayoría respondieron entre no tiene y solo uno en el hogar.

Valores faltantes

Con ayuda del siguiente gráfico se detectaron los valores faltantes que se encontraban en la base de datos.

```
missmap(datos, main = "Valores faltantes vs. Observados")
```



Las variables en las que se encontraron valores faltantes además de las que se pueden observar en el gráfico son:

- ubicación servicio sanitario
- educación madre
- condición de vida
- tipo energía
- estrato
- es campesino

La manera de resolver los datos faltantes consistió en hacer un reemplazo en los valores faltantes con la moda correspondiente de cada columna ya que como se observa en el gráfico, los valores faltantes corresponden solo al 1% de los datos en total.

Definición del usuario

Materiales

Métodos

Resultados

Conclusiones

Bibliografía