UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Trabajo de Grado
Junio 2019

Medición de los ingresos en Bogotá mediante modelos lineales generalizados con distribución Tweedie para el año 2017

Measurement of income in Bogotá using generalized linear models with Tweedie distribution for the year 2017

Ferney Arturo Simbasica Numpaque[a]
ferneysimbasica@usantotomas.edu.co

Dagoberto Bermúdez Rubio[b]
dagobertobermudez@usantotomas.edu.co

Resumen

En este estudio, se analiza la relación entre los ingresos mensuales de los bogotanos con respecto a la edad, los años de experiencia laboral, el tipo de trabajo, el estrato socio económico, el nivel educativo y la localidad en que vive, aplicando un modelo lineal generalizado con distribución Tweedie dada la ventaja que tiene función de admitir ceros en su variable respuesta de tipo continua no negativa, así como distribuciones que se encuentren sesgadas a la derecha, pudiendo ser utilizada en el análisis de los ingresos mensuales basados en la encuesta multipropósito de Bogotá realizada en el año 2017 a sus 20 localidades, encuestando a más de 77025 hogares entre septiembre de 2017 a febrero de 2018, donde se examina el impacto que tiene las variables explicativas en la variable respuesta ingresos mensuales promedio, de tal forma que se logre identificar las variables más influyentes que proporcionen una visión sobre la situación laboral y esto sirva como apoyo en estrategias que pueda llegar a tomar el distrito en políticas publicas enfocadas a mejorar el bienestar socioeconómico de los bogotanos.

Palabras clave: Tweedie, Modelos generalizados, Ingresos mensuales.

Abstract

This study analyzes the relationship between the monthly income of Bogotanos with respect to age, years of work experience, type of work, socioeconomic stratum, educational level and the locality in which they live, applying a generalized linear model with Tweedie distribution given the advantage that it has the function of admitting zeros in its variable response as well as distributions that are biased to the right, being able to be used in the analysis of the monthly income based on the multipurpose survey of Bogota carried out in 2017 to its 20 localities, by surveying more than 77025 households between September 2017 and February 2018, which examines the impact that explanatory variables have on the average monthly income response variable, in order to identify the most influential variables that provide a vision on the employment situation and this serves as support in strategies that the district may adopt in public policies aimed at improving the socioeconomic well-being of the Bogotanos.

Key words: Tweedie, Generalized models, Monthly income.

[a]Estudiante pregrado en Estadística, U. Santo Tomás, sede Bogotá
[b]Docente Facultad de Estadística, U. Santo Tomás, sede Bogotá

1. Introducción

En Colombia, cada año se discute sobre que incremento debe tener el salario mínimo de acuerdo a líderes de sindicatos, representantes de los sectores de producción y el gobierno, donde se reúnen a discutir sobre el valor que debe tener el salario mínimo pero muchas veces desconocen todos los factores que pueden llegan a influir en dicho salario, motivo por el cual se hace necesario realizar un estudio estadístico que me permita identificar cuáles son las condiciones que más influyen a que una persona que trabaja en la ciudad y que conducen a un bajo o alto ingreso. En este estudio se busca inicialmente analizar la ciudad capital en cada una de sus localidades para determinar el comportamiento de cada una de ellas en cuanto a salarios mensuales de sus habitantes y proponer un modelo lineal generalizado con el fin de determinar los principales factores de crecimiento o decrecimiento en el ingreso mensual de una persona.

El presente trabajo, se enfoca en los ingresos mensuales en Bogotá por medio de los modelos lineales generalizados, utilizando regresiones que permiten identificar los factores de mayor ingreso laboral mensual en Bogotá. Núñez y Bonilla (2001) estimaron la probabilidad que se tiene de estar desempleado y en la cual encontraron que el salario mínimo influía en la tasa de desempleo, debido a que personas que ganan un mńimo tienen una alta probabilidad de perder el empleo.

Hernández y Lasso (2003) estimaron ecuaciones de demanda de trabajo para adultos y jóvenes y para trabajadores calificados y no calificados durante el período 1984- 2000. Encontraron que el salario mínimo no es un determinante de la demanda de trabajo en ningún caso, mientras que el ciclo económico sí lo es. Adicionalmente, encontraron que aumentos en el salario mínimo tendrán un efecto positivo en el empleo de los jóvenes y un efecto negativo en los adultos. Sin embargo, estos resultados no tienen en cuenta el efecto sustitución entre trabajo de jóvenes y adultos y el efecto ingreso ante cambios del salario mínimo, por lo que los autores usaron otras ecuaciones. En este último caso, encontraron que un aumento del salario mínimo de 10 por ciento disminuye la demanda de trabajo de los jóvenes en 1,3 por ciento y la de los adultos en 0,9 por ciento por el efecto sustitución. No obstante, el efecto escala (el efecto positivo del crecimiento económico) compensa con creces el efecto sustitución, haciendo que el efecto total de un aumento del salario mínimo sobre la demanda de trabajo de los jóvenes y de los adultos sea positivo. El objetivo de este trabajo es lograr analizar las variables que presentan un mayor impacto sobre los ingresos laborales mensuales en Bogotá por medio de un modelo lineal generalizado Tweedie, donde también se busca comparar con respecto a otras alternativas de modelos lineales generalizados que admitan ceros en la variable respuesta y sean acordes a distribuciones sesgadas a la derecha. A partir de lo anterior, el documento tiene la siguiente estructura, en la sección 2 se abarca el marco teórico de modelos generalizados, distribución Tweedie, distribución inversa Gaussiana ajustada a cero y la distribución Gamma ajustada a cero. en la sección 3 se especifica el marco metodológico, diseño muestral, base de datos, variables del modelo, en la sección 4 se presentan los análisis de resultados, validación del modelo y por ultimo en la sección 5 se dan las conclusiones y trabajos futuros.

2. Marco Teórico

2.1. Modelo Lineal Generalizado

Los modelos lineales generalizados, son una extensión de distribuciones diferentes a la normal en las variables dependientes, como lo pueden ser todas aquellas distribuciones que hacen parte de la familia exponencial. La familia de distribuciones utilizada para los modelos lineales generalizados se denomina familia de modelos de dispersión exponencial (EDMs), que incluye distribuciones comunes como las distribuciones normal, binomial, Poisson y gamma, entre otras. Dunn, P.K, & Smyth,G.K (2005).

Un modelo generalizado está compuesto por:

- La componente aleatoria: Es la variable respuesta que siguen una distribución de la familia exponencial como inversa gauissiana, gamma, poisson, etc.

- La componente sistemática: Está compuesta por los parámetros desconocidos y las variables independientes utilizadas en el modelo.

- Función de enlace: La función de enlace relaciona el componente aleatorio con el componente sistemático del modelo como se muestra en la siguiente ecuación.

$$g(E(y)) = X\beta + \epsilon$$

2.2. Distribución Tweedie

La distribución Tweedie definida por Maurice Tweedie en el año 1984, tiene un soporte no negativo y se puede utilizar para para modelar respuestas que son una mezcla de ceros y valores de la variable respuesta positivos, motivo por el cual se puede utilizar en modelos lineales generalizados, donde los parámetros de la distribución son:

$$E[y] = \mu$$
$$V[y] = \phi\mu^{\alpha}$$

Donde ϕ es el parámetro de dispersión y α es un parámetro que controla la varianza de la distribución, la familia de distribuciones Tweedie contiene varias distribuciones importantes para modelos lineales generalizados, cuando $\alpha = 0$, la distribución se convierte en una distribución normal, cuando $\alpha = 1$ se convierte en una distribución de Poisson, $\alpha = 2$ se convierte en una distribución gamma, $\alpha = 3$ es una distribución gamma inversa, como se muestra en la siguiente tabla 1.

Tweedie	α	Soporte
Extremo estable	$\alpha < 0$	$\mathbb{R}$
Normal	$\alpha = 0$	$\mathbb{R}$
No existe	$0 < \alpha < 1$	
Poisson	$\alpha = 1$	$y = 0, \phi, 2\phi, ...$
Poisson-gamma	$1 < \alpha < 2$	$\mathbb{R}_0^+$
Gamma	$\alpha = 2$	$\mathbb{R}^+$
Estable positivo	$2 < \alpha < 3$	$\mathbb{R}^+$
Inversa Gaussiana	$\alpha = 3$	$\mathbb{R}^+$
Estable positivo	$\alpha > 3$	$\mathbb{R}^+$

Tabla 1: Distribución Tweedie

Para los demás casos, la distribución no tiene los términos se expresan en términos de series donde se hace necesario el uso de aproximaciones numéricas para evaluar la función de densidad. Dunn, P.K, Smyth,G.K (2005), proponen el uso de una serie finita y proporcionan una fórmula para determinar sus índices inferior y superior con el fin de lograr una mayor precisión. En este caso, la variable aleatoria Y de Tweedie puede generarse a partir de una distribución de Poisson compuesta como:

$$Y = \sum_{i=1}^{T} x_i$$

$$T \approx Poisson(\lambda)$$

$$x_i \approx gamma(\beta, \gamma)$$

dónde T y x_i son estadísticamente independientes, $gamma(\beta, \gamma)$ denota una variable aleatoria gamma que tiene media $\alpha\gamma$ y varianza $\alpha\gamma^2$. Estos parámetros están determinados por la distribución Tweedie de la siguiente manera:

$$\lambda = \frac{\mu^{2-\alpha}}{\phi(2-\alpha)}$$
$$\beta = \frac{2-\alpha}{\alpha-1}$$
$$\gamma = \phi(\alpha-1)\mu^{\alpha-1}$$

Inversamente, dados los parámetros de distribución de Tweedie, los parámetros de la distribución de Poisson compuesto se determinan de la siguiente manera:

$$\mu = \lambda\beta\gamma$$

$$\alpha = \frac{\beta+2}{\beta+1}$$

$$\phi = \frac{\lambda^{1-\alpha}(\beta\gamma)^{2-\alpha}}{2-\alpha}$$

Las distribuciones Tweedie pertenecen a la familia exponencial univariada (FEU).

$$f(y:\mu,\phi,\alpha) = a(y,\phi,\alpha)\exp\left\{\frac{y\theta - b(\theta)}{\phi}\right\} \tag{1}$$

Donde la media $\mu = \mathrm{E}(y) = b'(\theta)$, $\phi > 0$ es el parámetro de dispersión, θ el parámetro canánico y $b(\theta)$ la función cumulante. La función $a(y,\phi,\alpha)$ no se puede escribir en una forma cerrada aparte de los casos de Normal, Poisson, gamma e inversa gaussiana. La varianza de Y, está dada por $\mathrm{Var}(y) = \phi V(\mu)$, donde $V(\mu) = b''(\theta)$ es la función de varianza.

En términos de parametrizaciones de modelos lineales generalizados, el parámetro canónico θ para la densidad de Tweedie se puede expresar como

$$\theta = \begin{cases} \frac{\mu^{1-\alpha}}{1-\alpha}, & \text{si } \alpha \neq 1, \\ log\mu, & \text{si } \alpha = 1. \end{cases} \tag{2}$$

y la función $\beta(\theta)$ es:

$$\beta(\theta) = \begin{cases} \frac{\mu^{2-\alpha}}{2-\alpha}, & \text{si } \alpha \neq 2, \\ log\mu, & \text{si } \alpha = 2. \end{cases} \tag{3}$$

La distribución de Tweedie no se define cuando α está entre 0 y 1. En la práctica, el rango más interesante es de 1 a 2, en el que la distribución de Tweedie pierde gradualmente su masa en 0 a medida que cambia de una distribución de Poisson a una distribución gamma.

Por último en la siguiente tabla se muestra la función de varianza $V(\mu)$, la función $\beta(\theta)$, el parámetro canónico θ, el parámetro de dispersión ϕ y el soporte S.

EDM	$V(\mu)$	$\beta(\theta)$	θ	ϕ	S
Normal	μ^0	$\theta^2/2$	μ	σ^2	$\mathbb{R}$
Poisson	μ^1	$exp(\theta)$	$\log\,\mu$	1	$\mathbb{N}_0$
Gamma	μ^2	$-log(-\theta)$	$-\frac{1}{\mu}$	ϕ	$\mathbb{R}^+$
Inversa Gaussiana	μ^3	$-\sqrt{(-2\theta)}$	$-\frac{1}{2\mu^2}$	ϕ	$\mathbb{R}^+$
Tweedie	μ^α	$\frac{((1-\alpha)\theta)^{(2-\alpha)/(1-\alpha)}}{2-\alpha}$	$\frac{\mu^{1-\alpha}}{1-\alpha}$	ϕ	$1 < \alpha < 2 : \mathbb{R}_0^+$

Tabla 2: Familia de modelos de dispersión exponencial relacionadas a la Tweedie

2.3. Distribución Inversa Gaussiana ajustada a cero.

Comparamos el modelo Tweedie con otro modelo lineal generalizado que también admite valores de cero como lo es el modelo mixto continuo discreto, con una masa de probabilidad en cero y un componente continuo gaussiano inverso, Heller(2006).

Sea y_i = ingreso mensual, $i = 1, ..., n$. Podemos escribir la distribución de y como una función de probabilidad discreta-continua mixta:

$$f(y|\mu,\sigma,\pi) = \begin{cases} \pi & si \;\; y = 0 \\ (1-\pi)\cdot \frac{1}{\sqrt{2\pi y^3}} e^{-\frac{1}{2y}\left(\frac{y-\mu}{\mu\sigma}\right)^2} & si \;\; y > 0 \end{cases} \tag{4}$$

donde f (y) es la densidad de una distribución continua, sesgada a la derecha y es la probabilidad de que la variable respuesta tome valores de cero.

donde:

$$0 < y < \infty \;\; , \;\; 0 < \pi < 1 \;\; , \;\; \mu > 0 \;\; , \;\; \sigma > 0$$

con

$$E(Y) = (1-\pi)\mu$$
$$Var(Y) = (1-\pi)\mu^2(\pi + \mu\sigma^2)$$

2.4. Distribución Gamma ajustada a cero

La función de probablidad de la distribución gamma ajustada a cero definida por Rigby, R.A, (2010).

$$f(y|\mu,\sigma,\pi) = \begin{cases} \pi & si \;\; y = 0 \\ (1-\pi)\cdot \frac{1}{(\sigma^2\mu)^{1/\sigma^2}} \frac{y^{1/\sigma^2-1}e^{-y/\sigma^2\mu}}{\Gamma(1/\sigma^2)} & si \;\; y > 0 \end{cases}$$

Donde $0 \leq y < \infty \;\; , \;\; 0 < \pi < 1 \;\; , \;\; \mu > 0 \;\; , \;\; \sigma > 0$ y el valor esperados y varianza es:

$$E(Y) = (1-\pi)\mu$$
$$V(Y) = (1-\pi)\mu^2(\pi + \sigma^2)$$

El paquete en R *gamlss* por medio de un método semiparamétrico logra modelar la relación entre las variables predictorias y la variable respuesta a distribuciones sesgadas a la derecha. S.N.Wood (2006)

3. Marco Metodológico

3.1. Tipo de estudio

Estudio de tipo aplicado donde el objetivo es caracterizar y medir los ingresos mensuales a partir de variables explicativas contenidas en la encuesta multiproposito de Bogotá del año 2017.

3.2. Diseño muestral

3.2.1. Tipo de muestreo

Se realizó un muestreo multietápico, estratificado y de conglomerados para las diferentes UPZ de Bogotá.

Multietápico, para lograr la selección de las unidades de observación (viviendas, hogares o personas) se seleccionaron secuencialmente las unidades de muestreo (UPM y USM) en dos etapas. En la primera etapa se usa un muestreo proporcional al tamaño sistemático para elegir las UPM. En la segunda etapa, después de una segmentación de la UPM, se hace un muestreo aleatorio simple de conglomerados para elegir la o las USM. De esta forma se asegura que los conglomerados estén distribuidos por todo el estrato muestreo y no se recargue la muestra en espacios geográficos que tienen poca población.

Estratificado, se usó la partición de Bogotá en las unidades de planeación zonal para conformar 90 estratos de muestreo: 73 de estos son UPZ y los otros 17 son agrupaciones de UPZ. En el caso de los municipios de Cundinamarca, cada municipio es un estrato. Se establece esta estratificación para poder mejorar la precisión y generar resultados a nivel de UPZ, grupos de UPZ y de cada municipio.

De conglomerados, para el caso de Bogotá un conglomerado corresponde a un conjunto de predios con chip que contienen viviendas ubicadas dentro de la misma manzana o manzanas cercanas, a este grupo de viviendas se le denomina segmento o Medida de Tamaño (MT). En cada segmento seleccionado, se encuestan todas las viviendas, todos los hogares y todas las personas que los conforman.

Para el caso de los municipios de Cundinamarca los conglomerados están definidos por segmentos o medidas de tamaño de 10 viviendas (contiguas) en promedio donde se encuestan todas las viviendas, hogares y personas que lo conforman.

La encuesta contiene información de 77025 hogares en Bogotá que representan cerca de 281182 personas encuestadas, se realizó un muestreo probabilístico, la encuesta multipropósito con cobertura total de Bogotá y su parte Rural, Bogotá cuenta con 112 UPZ, de las cuales 73 UPZ se realizaron encuesta y las otras 39 UPZ se incluyen dentro de 17 agrupaciones. Las unidades estadísticas de observación están conformadas por las personas y hogares, siendo la unidad primaria las manzanas o unidades de manzanas, conformadas mínimo por 10 viviendas, mientras que la unidad secundaria de muestreo es el conjunto de predios que corresponde en promedio a 10 viviendas ubicadas dentro de la manzana o conjunto de manzanas.

3.2.2. Tamaño de la muestra

El tamaño de muestra se calcula por medio de la siguiente formula:

$$n = \frac{NPQdeff}{N(ESrelP)^2 + PQdeff} \tag{5}$$

Dónde:

n: Total de personas de la muestra
N: Total de personas en cada UPZ
P: Porcentaje de prevalencia a nivel de UPZ
$Q = 1 - P$
$ESrel =$ error esperado de las estimaciones
$deff = \frac{Var(congl.)}{Var(MAS)}$: Efecto de los conglomerados en el diseño

La precisión esperada del error relativo es de 7 % con un nivel de confianza del 95 % para una prevalencia de 10 % y un efecto de diseño de 1.2, donde el tamaño de muestra de cada UPZ (Anexo D) y de cada localidad es la siguiente:

Localidad	Tamaño de muestra
Antonio Nariño	4954
Barrios Unidos	5098
Bosa	10796
Candelaria	1834
Chapinero	3750
Ciudad Bolivar	13793
Engativá	14949
Fontibón	16769
Kennedy	24558
Los mártires	4162
Otra localidad rural	879
Puente Aranda	9249
Rafael Uribe Uribe	8442
San Cristóbal	11026
Santa Fé	7670
Suba	20336
Sumapaz	1025
Teusaquillo	9209
Tunjuelito	4639
Usaquén	10857
Usme	12614

Tabla 3: Tamaño de muestra por localidad

3.3. Base de datos

Para realizar el estudio, se utilizó la encuesta Multipropósito 2017 (EM2017) que contiene información estadística de las condiciones socioeconómicas y del entorno de los hogares y habitantes de Bogotá. La encuesta se realizó entre el 1 de septiembre del 2017 hasta el 30 de noviembre de 2017. La base de datos inicialmente contenía 281182 personas encuestadas en el año 2017 en Bogotá de las cuales se seleccionaran para el estudio aquellas que vivan en una de las 20 localidades de Bogotá, pertenecientes a uno de los seis estratos socieconomicos, de genero hombre o mujer y que a la pregunta ¿En qué actividad ocupó la mayor parte del tiempo la semana pasada?, hayan contestado la opción uno correspondiente a estar trabajando, luego del filtrado por estas condiciones, los datos de la encuesta a analizar serán

93125.

3.4. Variables del modelo

La variables a utilizar en el modelo lineal generalizado son las siguientes:

VARIABLE	CÓDIGO	ESPECIFICACIONES
Ingresos mensuales	NPCKP23 NPCKP36 NPCKP52A	Ingresos mensuales de los bogotanos en pesos.
Edad	NPCEP4	Años cumplidos
Genero	NPCEP5	Hombre = 1 , Mujer = 2
¿Cuantos años lleva trabajando?	NPCKP38A	Años de experiencia laboral
Estrato	NVCBP11AA	Estrato de 1 a 6
Localidad	NPCHP4A	¿En qué localidad está ubicado?
Tipo de trabajador	NPCKP17	En este trabajo es: 1 Obrero o empleado de empresa particular 2 Obrero o empleado del gobierno 3 Empleado doméstico 4 Profesional independiente 5 Trabajador independiente o por cuenta propia 6 Patrón o empleador 7 Trabajador de su propia finca
Último año o grado aprobado	NPCHP4A	nivel educativo: 1 Ninguno 2 Preescolar 3 Básica primaria (1 - 5) 4 Básica secundaria (6 - 9) 5 Media (10 - 13) 6 Técnico 7 Tecnológico 9 Universitaria completa (con título) 11 Especialización completa (con título) 13 Maestría completa (con título) 15 Doctorado completo (con título)

Tabla 4: Variables del modelo lineal

3.4.1. Ingresos mensuales

La variable ingresos mensuales en miles de pesos, corresponde a los ingresos que las personas que se encuentran trabajando reciben antes de descuentos en pensión y salud, incluyendo propinas, comisiones, al igual que ganancia neta o los honorarios en la actividad, negocio o profesión a la que se dedica. a continuación, se muestra un histograma con el comportamiento del salario a nivel descriptivo.

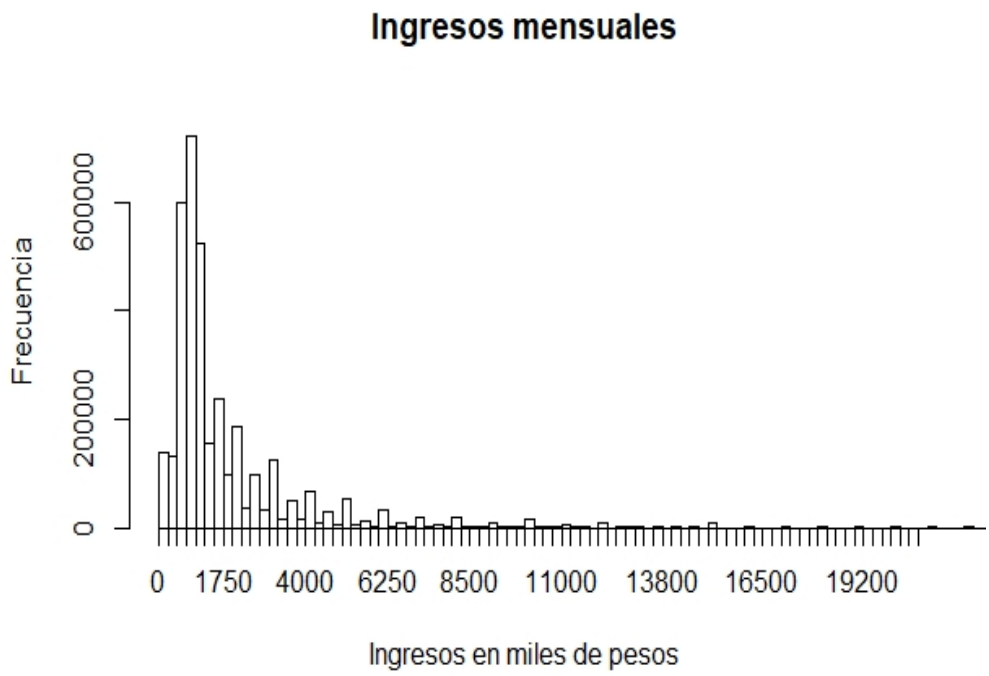


Figura 1: Ingresos mensuales.

En 25 % de los bogotanos, tienen un salario inferior a $ 750.000 pesos, el 50 % de los bogotanos que estan trabajando poseen un ingreso mensual superior a $1.000.000 pesos y el 75 % un ingreso máximo de $2.000.000 pesos, aunque la mayoría de salarios no superan 25 salarios mínimos, existen bogotanos que pueden llegar a ganar más de $100.000.000 pesos mensuales pero se considerán valores muy atípicos de la población, motivo por el cual no se tuvieron en cuenta en el análisis realizado.

	Ingresos mensuales en miles de pesos
Mínimo	$ 0
Primer cuartil	$750
Mediana	$1.000
Tercer cuartil	$2.000
Máximo	$100.000

Tabla 5: Ingresos mensuales por cuartiles

3.4.2. Estrato

De la variable estrato, tenemos que el comportamiento en cuanto a ingresos mensuales promedio es similar en el estrato uno y dos, presentando promedios de $846.194 y $1.093.495 pesos respectivamente. Mientras que los estratos cinco y seis presentan una diferencia considerable con respecto a los demás estratos.

Estrato	Ingreso en miles de pesos
1	$ 859
2	$1082
3	$1768
4	$4092
5	$5693
6	$7593

3.4.3. Edad

La edad(NPCEP4) corresponde a la edad de la persona encuestada en el año 2017, las edades están comprendidas desde los 16 años, a continuación se presenta el histograma de frecuencias para las diferentes edades de los encuestados, donde se presenta un sesgo hacia la derecha.

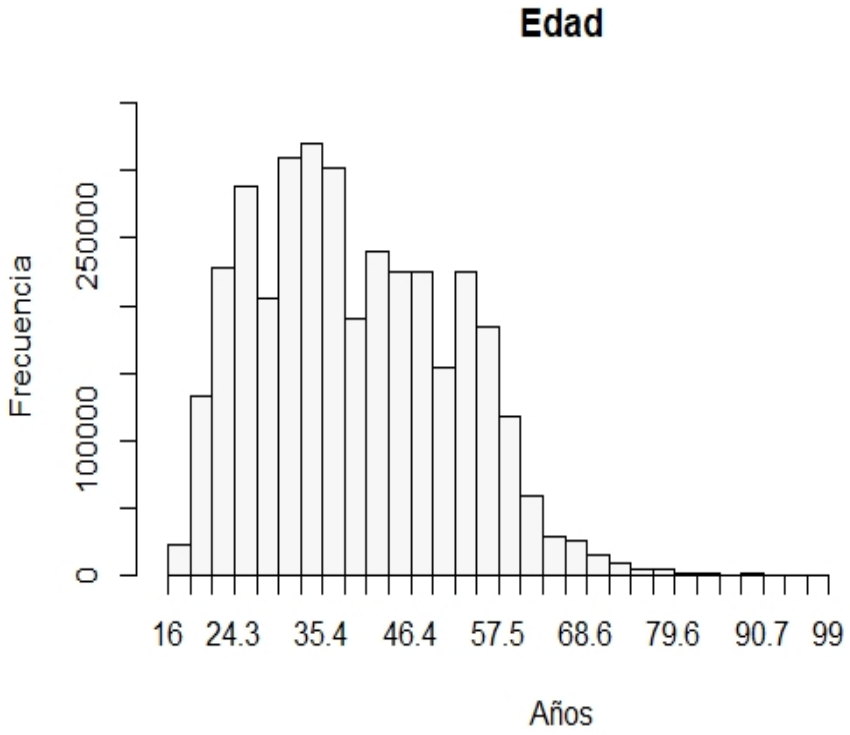


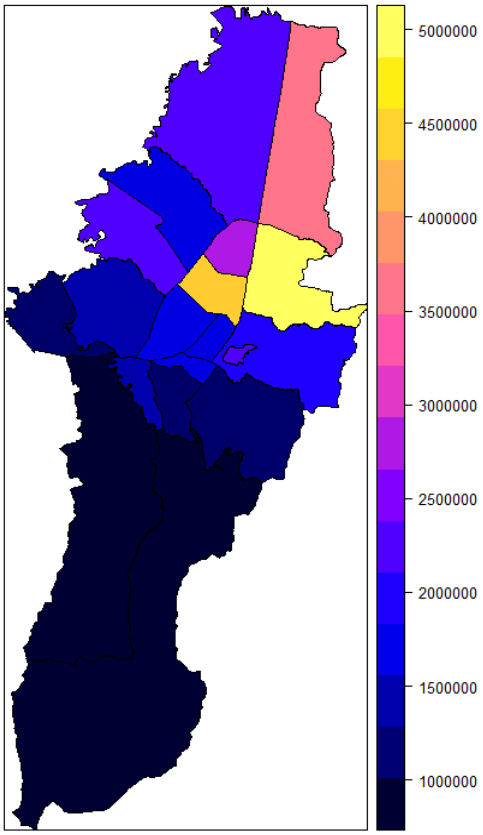
Figura 2: Histograma de frecuencias de la edad, elaboración propia.

3.4.4. Localidad

La localidad corresponde al lugar donde habita la persona encuestada, esta relacionado con la distribución geográfica en Bogotá de las cuales se tienen en cuenta las 20 localidades, a continuación se presenta el ingreso promedio de los ciudadanos dependiendo de la localidad.

Localidad	Ingresos mensuales en miles de pesos
Chapinero	$4.856
Teusaquillo	$4.391
Usaquen	$3.563
Barrios Unidos	$2.746
Suba	$2.257
Candelaria	$2.199
Fontibon	$2.190
Santa fé	$1.883
Engativá	$1.762
Los Mártires	$1.760
Antonio Nariño	$1.677
Puente Aranda	$1.617
Kennedy	$1.421
Tunjuelito	$1.326
Rafael Uribe Uribe	$1.183
San Cristóbal	$1.136
Bosa	$1.025
Usme	$1.004
Ciudad Bolívar	$1.000
Sumapaz	$627

Tabla 6: Ingresos mensuales por localidad



El cuadro especifica los ingresos promedios por localidad, donde se evidencia que las localidades de Chapinero y Teusaquillo son las que mejores ingresos tienen por habitante, mientras que localidades como Usme y Ciudad Bolívar son las de menor ingreso promedio por habitante, a excepción de Sumapaz que su economía es rural y no urbana, por lo tanto esta clasificación por localidades no es homogénea y presenta diferencia en sus ingresos mensuales promedio.

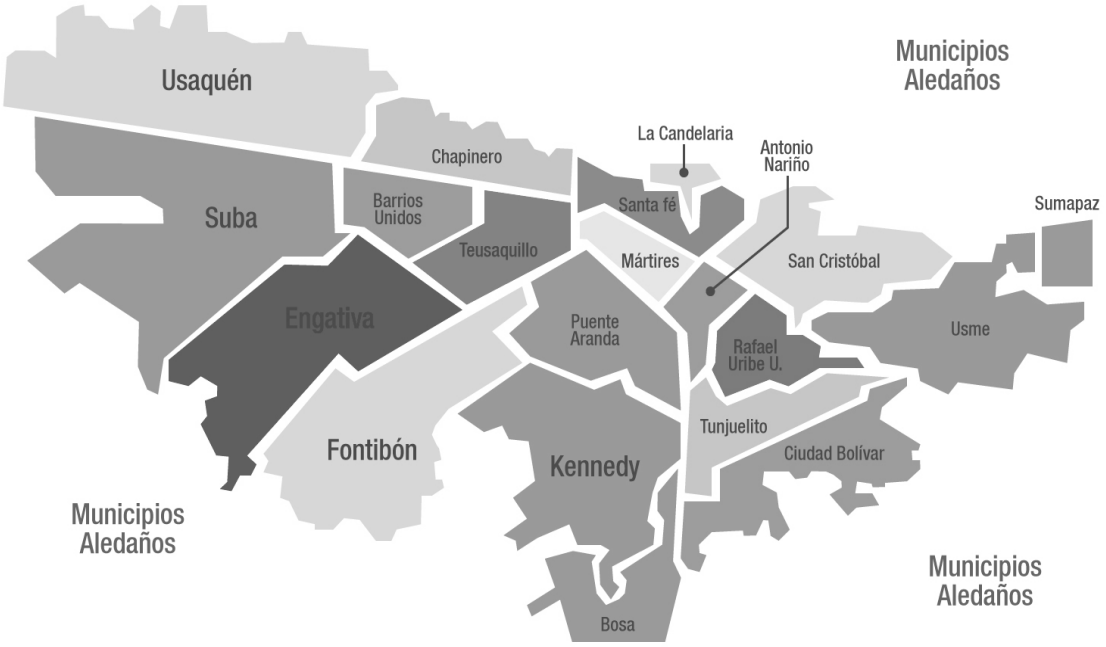


Figura 3: Localidades de Bogotá.

3.4.5. Tipo de trabajador

TIPO DE TRABAJADOR	INGRESO MENSUAL
Empleado de empresa particular	$1.799
Empleado del gobierno	$3.003
Empleado doméstico	$716
Profesional independiente	$2.874
Trabajador independiente	$1.274
Empleador	$3.760
Trabajador de su propia finca	$1.566

Tabla 6: Ingreso mensual en miles de pesos dependiendo del tipo de trabajador

El ingreso mensual promedio más alto, se presenta por el empleador, que tienen un ingreso promedio de $3.760 miles de pesos, seguido por el empleado del gobierno que tiene ingresos mensuales promedio de $3.003. Las personas con menor ingreso mensual son los empleados domésticos que en promedio ganan $716 miles de pesos.

4. Análisis de resultados y validación del modelo

4.1. Imputación por regresión

En la encuesta analizada, se presentaron datos incompletos en la variable NPCHP4 correspondiente al nivel educativo de la persona, para solucionar este problema se utilizó el método propuesto por Buck (1960), donde se emplean modelos de regresión para imputar información en la variable NPCHP4, La metodología inicialmente consiste en crear una nueva base sin los datos faltantes y a partir de ella hacer un modelo de regresión lineal para estimar los valores faltantes, donde las K variables, $X = (X_1, ..., X_K)$, no presenten valores perdidos, para los casos i en el cual no hay un valor de Y_i, este valor faltante es imputado por medio de un modelo lineal generalizado como se muestra a continuación:

$$g\{E(Y)\} = \beta,\ Y \sim F$$

donde g se denomina función enlace y F es la función de distribución. En este caso como la variable nivel educativo es categórica ordinal que toma valores $j = 1, 2, ..., 15$ por lo que se realiza el modelo Logit ordinal acumulado para estimar los datos faltantes.

Así, para una categoría dada j se define la probabilidad acumulada como:

$$P(Y <= j) = \pi_1 + \pi_2 + ... + \pi_{15}$$

Donde las probabilidades acumuladas reflejan el orden entre las categorías:

$$P(Y \leq 1) \leq P(Y \leq 2) \leq \cdots \leq P(Y) = 1$$

El modelo logit acumulado generalizado es:

$$logit\,(P(Y \leq j)|x) = \alpha_j + \sum_{i=1}^{d} \beta_i x_i$$

los datos faltantes que presenta la variable nivel educativo es la siguiente

Nivel educativo	Datos sin imputación	Datos con imputación
1.Ninguno	0.557 %	0.557 %
2.Preescolar	0.049 %	0.049 %
3.Básica primaria (1 - 5)	9.891 %	9.933 %
4.Básica secundaria (6 - 9)	8.374 %	8.374 %
5.Media (10 - 13)	27.046 %	31.318 %
6.Técnico	10.326 %	10.326 %
7.Tecnológico	4.859 %	4.859 %
8.Universitaria incompleta (sin título)	2.632 %	2.632 %
9.Universitaria completa (con título)	19.463 %	21.850 %
10.Especialización incompleta (sin título)	0.239 %	0.239 %
11.Especialización completa (con título)	6.421 %	6.451 %
12.Maestría incompleta (sin título)	0.209 %	0.209 %
13.Maestría completa (con título)	3.039 %	3.045 %
14.Doctorado incompleto (sin título)	0.004 %	0.004 %
15.Doctorado completo (con título)	0.153 %	0.153 %
NA	6.737 %	0 %

Tabla 7: Imputación de datos

El 6.737 % de los datos son imputados, donde el modelo lineal logit acumulado, asigna los valores perdidos a las categorías 5, 9, 11 y 13 que corresponde a educación media, educación universitaria, especialización y maestría respectivamente, quedando la base de datos sin valores perdidos en la variable NPCHP4.

4.2. Coeficiente de correlación de Spearman

Primero se verifica el supuesto de normalidad en las variables cuantitativas para utilizar el coeficiente de correlación de Pearson en el caso de cumplirse el supuesto o Spearman en el caso contrario.

$$H_0 : \text{La variable siguen una distribución normal}$$
$$H_1 : \text{La variable no siguen una distribución normal}$$

En una prueba de normalidad por Shapiro Wilk, utilizando el paquete en R $shapiro.test()$ se obtienen p-valor aproximadamente de cero, a un nivel de significancia del 5 % se rechaza la hipotesís nula, las variables no siguen una distribución normal, se procede a utilizar la prueba de correlaciones por el metodó de Spearman.

Variable	W	P valor
Ingresos	0.61806	0.00000000000000022
Edad	0.97457	0.00000000000000022
Años de experiencia	0.75965	0.00000000000000022

Tabla 8: Prueba Shapiro Wilk

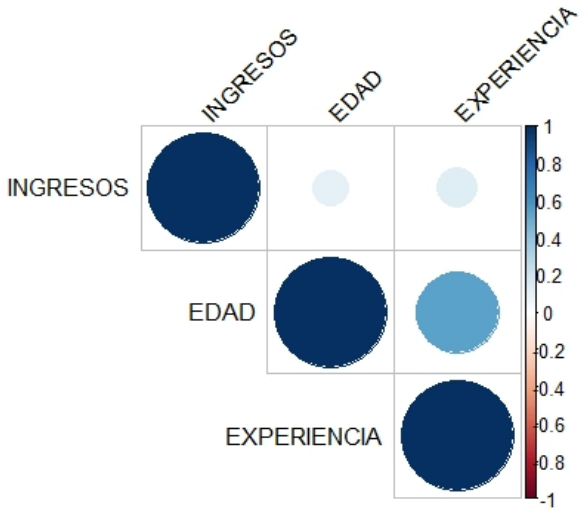


Figura 4: Correlación.

Del paquete de R *cor.test*() utilizando el método de Spearman obtenemos que son significativas las correlaciones entre las variables ingresos, años de experiencia y edad de la persona, siendo factible usarlas en el modelo de regresión.

Variables explicativas cuantitativas	ρ	P-valor
Edad	0.08156792	0.000000007654
Años de experiencia	0.1855532	0.0000000000000002

Tabla 9: Correlación entre las variables explicativas y la variable respuesta

La variable ingresos, la variable edad y la variable años tienen una correlación positiva, a medida que aumenta la edad de la persona aumentan sus ingresos, al igual que entre más experiencia laboral, mayor es el ingreso mensual.

4.3. Estimación del parámetro α

El modelo lineal generalizado basado distribución tweedie se va a utilizar para estimar los ingresos mensuales de los bogotanos, la distribución tweedie pertenece a la familia exponencial de distribuciones con potencia α entre $1 < \alpha < 2$, para ajustar mejor la variabilidad que presentan los datos de ingresos mensuales, esta es una de las grandes ventajas que presenta frente a otros modelos lineales donde la homocedasticidad de la variable mantenerse. De Jong,P.(2008).

Las distribuciones de Tweedie se definen como:

$$var[y] = \phi V(\mu) = \phi \mu^{\alpha}$$

$$log(var[y]) = log(\phi) + \alpha log(\mu) \tag{6}$$

Por este motivo, se puede realizar una estimación aproximada del parámetro por medio de una regresión lineal simple, donde la idea es dividir los datos en pequeños grupos y comparar el logaritmo de las

variaciones de los grupos contra el logaritmo de la medias de los grupos, sin embargo las estimaciones dependen de como se dividan los datos. Dunn & Smyth (2018).

En este caso la variable ingreso se dividió por las 20 localidades que tiene la ciudad de Bogotá y por medio de una regresión lineal se obtuvo un parámetro aproximado de $\alpha = 1.889$ (Anexo B), indicando que el parámetro se encuentra entre $1 < \alpha < 2$.

Como este método lineal es solo una aproximación, se procede a realizar una estimación de α por máxima verosimilitud, se utilizara la función en r tweedie.profile() para calcular la estimación del parámetro α.

Si los datos contienen ceros, es probable que la estimación del parámetro este entre $1 < \alpha < 2$, en este caso se van a hacer divisiones por los estratos sociales, siendo el estrato uno, dos y tres de comportamiento similar, mientras que los estratos altos presentan mayor dispersión, con base a estas divisiones el parámetro estimado por máxima verosimilitud esta en un intervalo de confianza al 95 % entre $1.694871 < \alpha < 1.694876$ con una estimación puntual de $\hat{\alpha} = 1.6948735$.

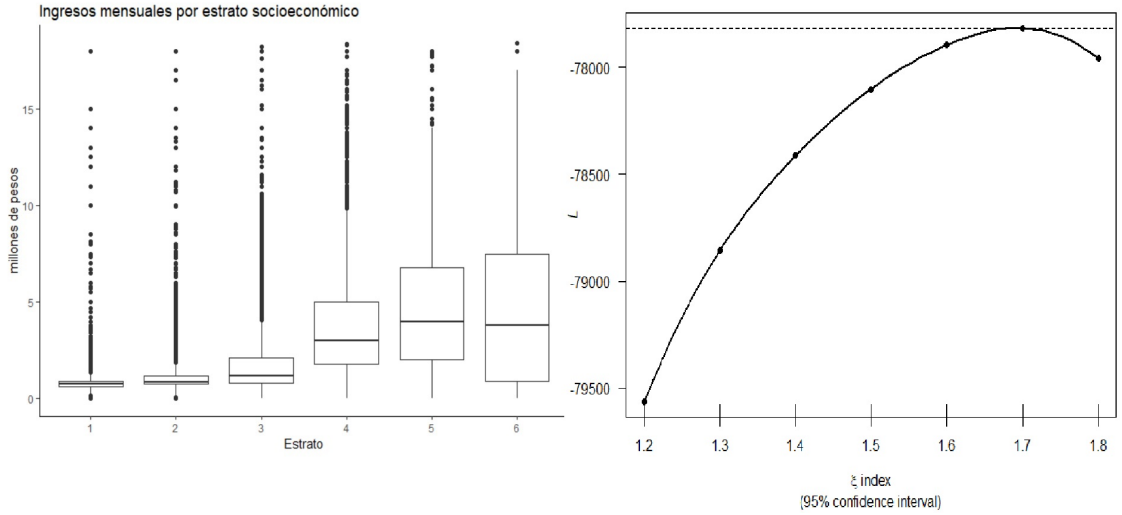


Figura 5: Ingresos por estrato y gráfica de verosimilitud para estimar el parámetro $\hat{\alpha}$.

$$2\left\{\ell(\hat{\alpha}; y; \hat{\phi}; \hat{\mu}) - \ell(\alpha; y; \hat{\phi_\alpha}; \hat{\mu_\alpha})\right\} \sim \chi^2_1 \tag{7}$$

siendo $\ell(\alpha; y; \hat{\phi_\alpha}; \hat{\mu_\alpha})$ la log-verosimilitud del parámetro α y $\ell(\hat{\alpha}; y; \hat{\phi}; \hat{\mu})$ la máxima verosimilitud.

La función de probabilidad para la distribución Tweedie no tiene forma cerrada, excepto en los casos donde el parámetro toma valores de $\alpha = 0$ (Normal), $\alpha = 1$ (Poisson), $\alpha = 2$ (Gamma) y $\alpha = 3$ (Inversa Gaussiana), sin embargo existen métodos numéricos para evaluar la distribución Tweedie, Dunn & Smyth (2018), por lo tanto se utilizara la función R tweedie.profile(), Dunn, P.K(2017), para estimar el parametro α.

El cálculo a veces es lento, para este caso en particular el proceso duró 5531 segundos, siendo aproximadamente hora y media en el proceso computacional de estimación del parámetro cuando se tienen más de noventa mil datos de la variable ingresos mensuales. La estimación por máxima verosimilitud del parámetro en un intervalo de confianza al 95 % está entre $1.684871 < \alpha < 1.684876$ con una estimación puntual de $\hat{\alpha} = 1.689796$, gráficamente se ve los valores obtenidos para la estimación por máxima

verosimilitud. Anexo C.

4.3.1. Bondad de ajuste del modelo

La bondad de ajuste permite determinar, qué modelo lineal generalizado es el que presenta mayor explicación y precisión de acuerdo a la base de datos analizada. Para este análisis del modelo lineal Tweedie los criterios a utilizar sera la desviación (Deviance) y el criterio de información de Akaike(AIC). Se realiza una prueba con diferentes valores para α, valores que están entre 1 y 2 que es donde la distribución tweedie permite ceros en la respuesta. El deviance es la diferencia entre el modelo analizado De Jong, P., Heller, G. Z. (2008).

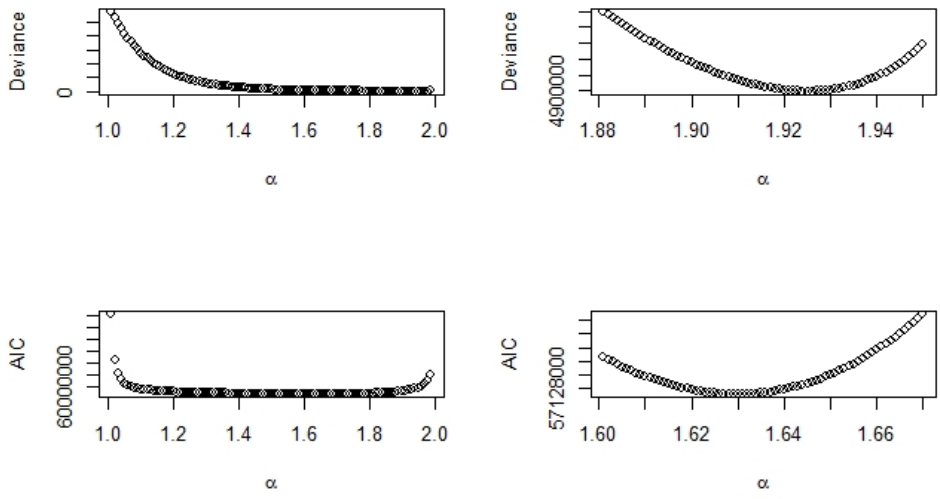


Figura 6: Deviance residual y AIC según $1 < \alpha < 2$

De acuerdo a las anteriores gráficas, estimamos un modelo generalizado tweedie con un $\hat{\alpha} = 1.93$ que es en el que se logra el menor Deviance de 4896839, utilizando el paquete R *tweedie*(*var.power*, *link.power*), mientras que en el caso del AIC se logra el mínimo cuando $\hat{\alpha}$ es 1.63 con AIC de 57127137, escogiendo finalmente el modelo con $\alpha = 1.93$ que presenta el menor Deviance.

```
modelo.tweedie.1.93 <- glm(INGRESOS     ~
                       NPCEP4           + #Edad
                       NPCKP38A         + #Experiencia laboral
                       factor(NPCEP5)   + #Genero
                       factor(NVCBP11AA)+ #Estrato
                       factor(LOCALIDAD)+ #Localidad
                       factor(EDUCACION)+ #Nivel educativo
                       factor(NPCKP17)  , #Tipo de trabajador
                       weights = FEX_C  ,
                       family = tweedie(var.power=1.93, link.power=0),
                       data = BASE)
```

Del modelo obtenemos los siguientes betas estimados (ver Anexo A), al dejar link.power = 0 se obtiene una Tweedie con enlace de logaritmo, motivo por el cual se transforma a e^{β_i} para su correcta interpretación, a continuación mostrados algunos de los coeficientes betas estimados del modelo.

Nombre	Variable	β_i	e^{β_i}
Edad	NPCEP4	0.0044	1.0044
Años de experiencia	NPCKP38A	0.0100	1.0101
Genero mujer	factor(NPCEP5)2	-0.2276	0.7964
Estrato 1	factor(NVCBP11AA)1	-0.3121	0.7319
Estrato 2	factor(NVCBP11AA)2	-0.1665	0.8466
Estrato 4	factor(NVCBP11AA)4	0.4274	1.5333
Estrato 5	factor(NVCBP11AA)5	0.6270	1.8720
Estrato 6	factor(NVCBP11AA)6	0.7874	2.1977
Localidad	factor(LOCALIDAD)BOSA	-0.0368	0.9639
Localidad	factor(LOCALIDAD)CANDELARIA	0.1524	1.1646
Localidad	factor(LOCALIDAD)CHAPINERO	0.1234	1.1313
Localidad	factor(LOCALIDAD)FONTIBON	-0.0366	0.9641
Localidad	factor(LOCALIDAD)SUMAPAZ	-0.3585	0.6987
Localidad	factor(LOCALIDAD)TEUSAQUILLO	0.0628	1.0648
Localidad	factor(LOCALIDAD)USAQUEN	0.0630	1.0650
Básica Primaria	factor(EDUCACION)3	-0.2173	0.8047
Técnico	factor(EDUCACION)6	0.1381	1.1481
Universitaria	factor(EDUCACION)9	0.6128	1.8456
Especialización	factor(EDUCACION)11	1.0478	2.8514
Maestría	factor(EDUCACION)13	1.0285	2.7969
Doctorado	factor(EDUCACION)15	1.5290	4.6136
Empleado del gobierno	factor(NPCKP17)2	0.1663	1.1809
Empleado domestico	factor(NPCKP17)3	-0.2750	0.7596
Profesional independiente	factor(NPCKP17)4	0.0638	1.0659
Trabajador independiente	factor(NPCKP17)5	-0.0665	0.9357
Empleador	factor(NPCKP17)6	0.3239	1.3825

Tabla 10: Algunos coeficientes beta estimados del modelo

De la tabla anterior, que contiene los valores estimados de los coeficientes beta de la distribución Tweedie, podemos decir que por cada año cumplido de mas que tienen la persona, su ingreso mensual aumenta en promedio 0.44 %, con respecto a una persona que tenga igualdad en condiciones pero siendo un año menor, en cuanto a la experiencia laboral por cada año de experiencia aumenta un 1 % su ingreso mensual, mientras que si es mujer su ingreso disminuye un 20.36 % con respecto a un hombre, manteniendo constantes las demás variables de experiencia laboral, edad, estrato, tipo de empleo y nivel educativo. Tomando como referencia el estrato tres, una persona de estrato dos tiene un ingreso mensual 15.34 % menor, si es de estrato cuatro sus ingresos aumentan en un 53.33 % y de estrato cinco los ingresos mensuales son 87.2 % más altos, bajo las mismas condiciones de edad, género, localidad, nivel de educación y tipo de trabajo. En cuanto al lugar donde viven, se toma como referencia la localidad de Suba y comparándola con las otras localidades por medio del modelo lineal Tweedie analizamos que una persona que viva en chapinero tiene un 13.13 % más alto sus ingresos mensuales, si vive en la localidad de Usaquen un 6.5 % más alto y si vive en las localidades de Antonio Nariño, Barrios unidos, Bosa, Ciudad Bolívar, Engativá, Fontibón, Kennedy, Los Mártires, San Cristóbal,

Teusaquillo y Usme, son estadísticamente similares en cuanto a ingresos mensuales para el modelo lineal tweedie.

Con respecto al nivel de educación observamos que es el factor que más influye en los ingresos mensuales de una persona, en el modelo Tweedie se tomó como referencia las personas que tienen educación media, por consiguiente una persona que tiene un nivel técnico tiene un ingreso mensual 14.81 % más alto, un universitario 1.84 veces más alto, en tanto las personas que posee un posgrado su nivel de ingreso mejora significativamente con respecto a una persona que solo tiene educación media, siendo la especialización 2.85, la maestría un 2.79 y un doctorado un 4.61 veces más altos, manteniendo las demás variables explicativas constantes, por ultimo tomando como referencia las personas que trabajan en una empresa privada, tenemos que aquellas que trabajan con el gobierno tienen un 18.09 % más en sus ingresos mensuales y los profesionales que trabajan de forma independiente un 6.59 % más en sus ingresos.

Por medio de la función en R *predict*() se realiza una estimación de los ingresos mensuales con modelo Tweedie como muestra la siguiente gráfica.

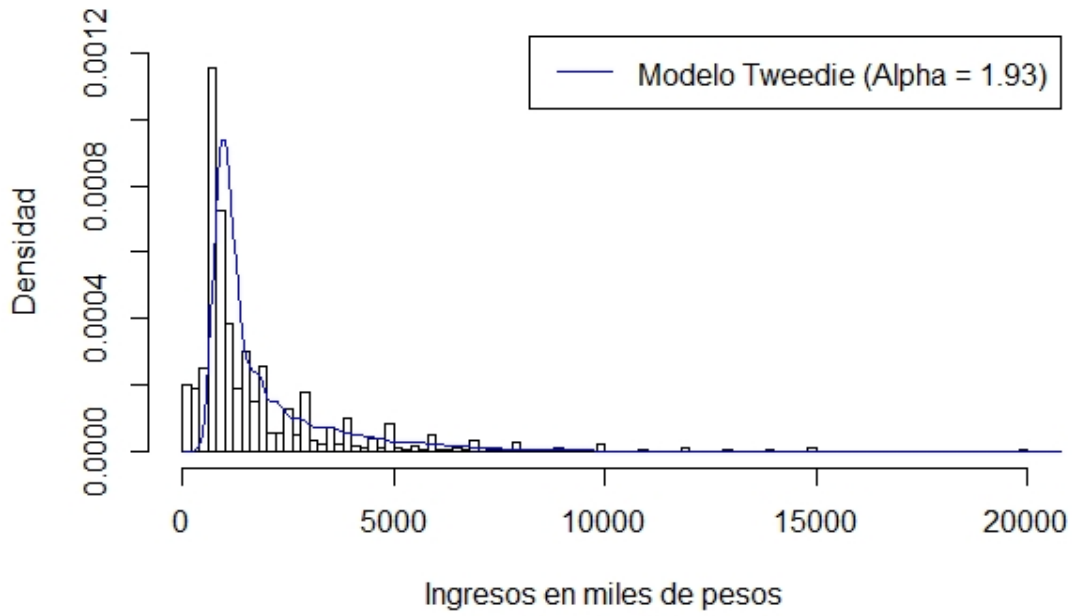


Figura 7: Histograma de ingresos mensuales

El modelo lineal generalizado Tweedie se ajusta relativamente bien al comportamiento de los ingresos mensuales dado que son datos que se encuentran sesgados a la derecha, otras distribuciones que también funcionan en este caso, son la distribución inversa gaussiana y la distribución gamma ajustadas a cero.

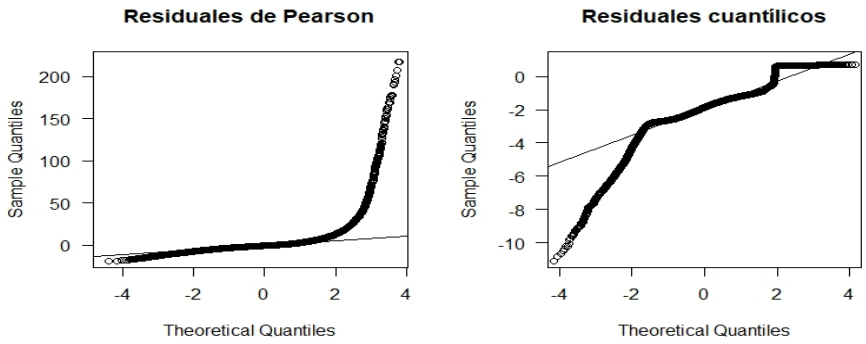


Figura 8: Residuales del modelo lineal Tweedie

Se compara los residuales de Pearson y los residuales cuantílicos, como se ve en los gráficos, en los valores inferiores que es donde se presentan los valores de ingreso mensual cero y bajos ingresos, es difícil para la distribución Tweedie ajustarse a dichos valores pero a medida que aumentan los ingresos, se logra ajustar de una mejor manera, en este caso los residuos cuantílicos se basan en el concepto de invertir la función de distribución estimada para cada observación con el fin de obtener exactamente residuos normales estándar, para el caso de distribuciones discretas Poisson se introduce cierta aleatorización para producir residuos normales continuos. Los residuos cuantílicos son los residuos de elección para los modelos lineales generalizados en situaciones de gran dispersión cuando la desviación y los residuos de Pearson pueden ser muy poco normales. Dunn, K. P., and Smyth, G. K. (1996).

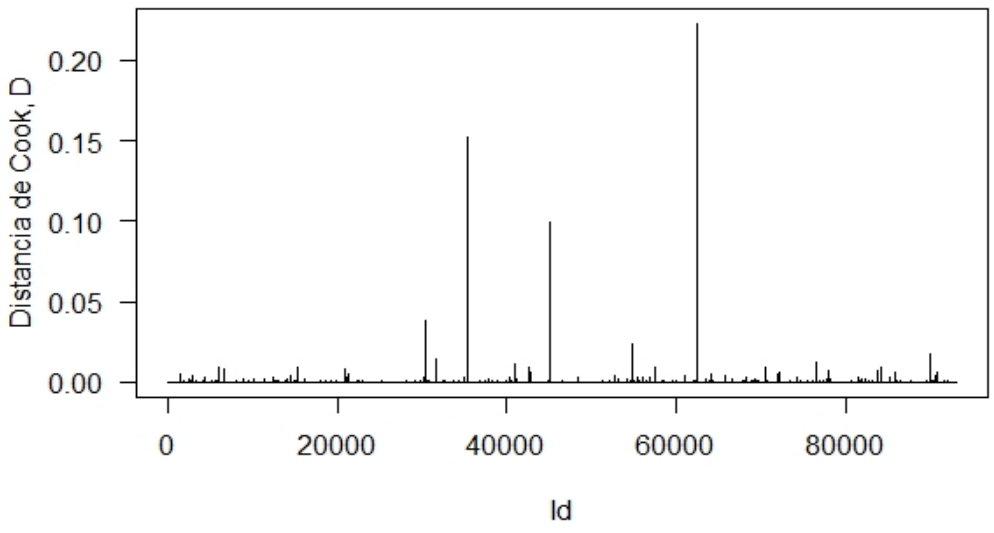


Figura 9: Distancia de cook

Utilizando la función en R *cooks.distance*() podemos ver en la gráfica que existen varios valores de la

encuesta que se encuentran muy distantes, siendo valores atípicos en el modelo de regresión, de los cuales se muestran a continuación en la tabla con los cinco valores con mayor distancia de Cook, donde el valor más atípico corresponde a una persona que tiene ingresos mensuales de cien millones de pesos, sin ningún nivel de educación, con dos años de experiencia en el trabajo que realiza y que trabaja de forma independiente, un caso similar sucede con la persona id 62419, que al no tener ningún nivel educativo y solo tener 26 años de edad logra tener un ingreso mensual de cincuenta y ocho millones de pesos, resultando incoherente con respecto a otros datos atípicos de la encuesta como por ejemplo el id 47456 que corresponde a un hombre de 53 años con 20 años de experiencia , es el patrón en su trabajo y posiblemente bajo estas condiciones sea coherente el ingreso mensual de $60 millones de pesos dados en la encuesta multipropósito de Bogotá. La encuesta fue realizada por la Secretaría Distrital de Planeación de Bogotá (SDP) y el Departamento Administrativo Nacional de Estadística (DANE) quedando a consideración que dichos valores atípicos de la encuesta corresponden al porcentaje de errores que puede llegar a tener una encuesta de esta magnitud.

Id	Ingresos en miles de pesos	Genero	Edad	Experiencia	Nivel de educación	Tipo de trabajo
62412	$100.000	Hombre	49	2	Ninguno	Independiente
35321	$100.000	Hombre	24	1	Primaria	Independiente
44997	$95.000	Hombre	28	15	Media	Independiente
62419	$58.000	Hombre	26	3	Ninguno	Independiente
47456	$60.000	Hombre	53	20	Primaria	Patrón o empleador

Tabla 11: Valores mas atípicos dados por la distancia de Cook

4.4. Comparación con otros modelos

Otros modelos con los cuales se puede utilizar una variable respuesta con ceros es el modelo ZAIG (Zero Adjusted Inverse Gaussian Distribution) y el modelo ZAGA (Zero adjusted Gamma distribution), dado que son modelos adecuados cuando la variable respuesta presenta sesgo extremo a la derecha. Rigby, R.A, Stasinopoulos D.M (2005).

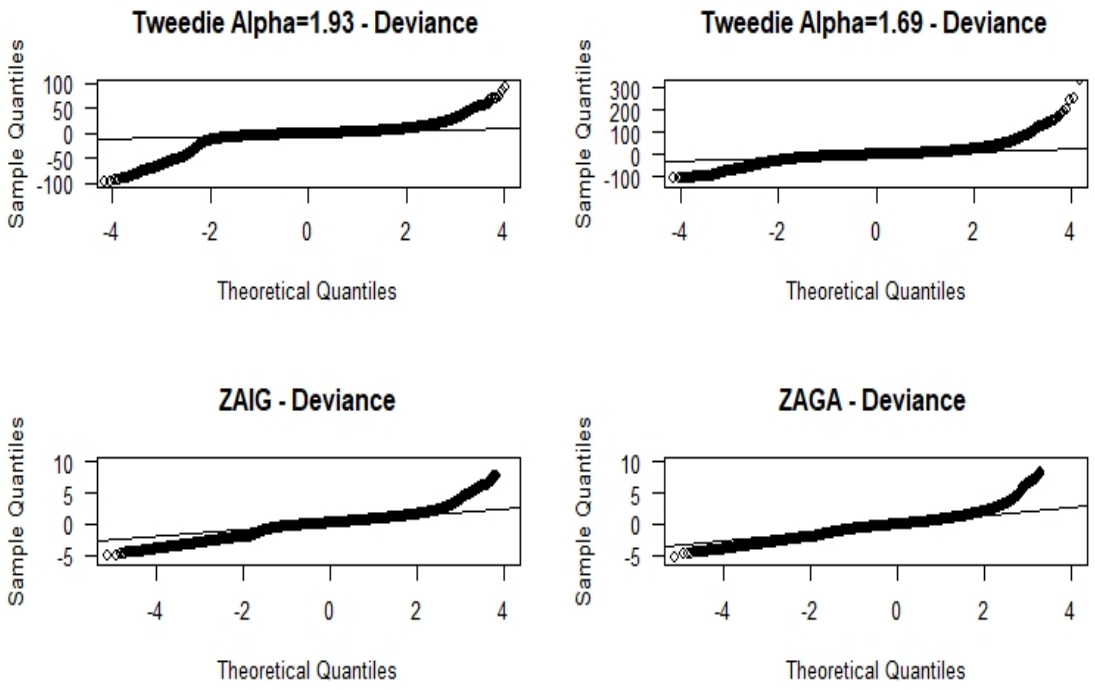


Figura 10: Residuales Deviance modelos Tweedie, ZAIG y ZAGA

De los modelos expuestos el de menor Deviance es el modelo tweedie que presenta un valor de 4905250, mientras que si elegimos el modelo lineal según el criterio de akaike, el modelo ZAGA resulta ser el mas adecuado. Aunque en modelos lineales generalizados la decisión se toma de acuerdo al menor Deviance y en este caso se elige el modelo lineal generalizado tweedie con parámetro $\alpha = 1.93$

MLG	AIC	Deviance
Tweedie $\alpha = 1.63$	57127137	24971921
Tweedie $\alpha = 1.69$	57154455	16457855
Tweedie $\alpha = 1.93$	59027624	4905250
ZAIG	58005821	58005719
ZAGA	56358233	56358131

Tabla 12: AIC - Deviance

5. Conclusiones y trabajos futuros

- La distribución Tweedie, sirve para modelar el ingreso mensual de las personas en una ciudad, teniendo validez similar a lo que se logra con otras distribuciones como la inversa gaussiana ajustada a cero y la distribución gamma ajustada a cero, dado que se presentan casos donde los ingresos son cero, estas tres distribuciones son aptas por permitir valores nulos en sus funciones. Sin embargo en la práctica es habitual que utilicen para el conteo de ceros el modelo generalizado Poisson y para comportamiento de ingresos mayores a cero el modelo generalizado gamma, motivo por el cual se planteó utilizar la distribución tweedie en vista que combina ambas distribuciones en una sola cuando la potencia α de la distribución Tweedie cumple $1 < \alpha < 2$, a diferencia de los modelos mixtos el modelo lineal generalizado Tweedie presenta el mismo parámetro de dispersión, entre tanto los otros métodos asignan parámetros a cada uno de los grupos, en consecuencia el modelo generalizado Tweedie es una buena alternativa a los métodos usados habitualmente para distribuciones sesgadas a la derecha que presenten ceros en su variable respuesta.
- Del análisis descriptivo se concluye que el 50 % de la población bogotana que se encuentra trabajando presentan ingresos menores a un millón de pesos y tan solo un 25 % logran superar los dos millones pesos en sus ingresos mensuales, también cabe resaltar que los empleados que mejor estabilidad económica poseen son los que trabajan con el gobierno dado que sus ingresos mensuales de $ 3.003.406 de pesos son en promedio superiores a los ingresos de personas que trabajan con empresas particulares o de forma independiente.
- En las localidades de Bogotá se tiene que Ciudad Bolívar siguen siendo una de las localidades dentro del perímetro urbano que menor ingreso promedio por persona presenta siendo de tan solo $ 1.000.344 pesos, al igual que Usme con un $ 1.003.814, aunque la localidad de Sumapaz presenta ingresos menores se debe tener en cuenta que el comportamiento y características de dicha localidad son de carácter rural y no presenta los mismo comportamientos urbanos de las otras localidades.
- Finalmente, concluimos que el estudio proporciona una visión sobre las variables que más influyen en los ingresos mensuales y esto en trabajos futuros puede llegar a servir como parte de estrategias del distrito en políticas publicas enfocadas a mejorar el bienestar socioeconómico de los bogotanos. Teniendo en cuenta que de todas las variables explicativas usadas en el modelo, el nivel de educación es el que tiene un mayor peso a la hora de estimar los ingresos, siendo significativo el aumento en los ingresos mensuales cuando la persona realiza un posgrado.

Anexos

A. Modelo Tweedie

```
> summary(modelo.tweedie.1.93)

Call:
glm(formula = INGRESOS ~ NPCEP4 + NPCKP38A + factor(NPCEP5) +
    factor(NVCBP11AA) + factor(LOCALIDAD) + factor(EDUCACION) +
    factor(NPCKP17), family = tweedie(var.power = 1.93, link.power = 0),
    data = BASE, weights = FEX_C)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-100.224    -2.953    -0.906     0.960   150.792

Coefficients:
                                      Estimate Std. Error  Pr(>|t|)
(Intercept)                          6.9484556  0.0178892  0.0000 ***
NPCEP4                               0.0044245  0.0003619  0.0000 ***
NPCKP38A                             0.0099701  0.0005198  0.0000 ***
factor(NPCEP5)2                     -0.2275586  0.0072860  0.0000 ***
factor(NVCBP11AA)1                  -0.3121319  0.0187493  0.0000 ***
factor(NVCBP11AA)2                  -0.1665226  0.0100214  0.0000 ***
factor(NVCBP11AA)4                   0.4273529  0.0148649  0.0000 ***
factor(NVCBP11AA)5                   0.6270453  0.0221932  0.0000 ***
factor(NVCBP11AA)6                   0.7874312  0.0338684  0.0000 ***
factor(LOCALIDAD)ANTONIO NARI O      0.0108610  0.0311335  0.7272
factor(LOCALIDAD)BARRIOS UNIDOS      0.0176658  0.0208142  0.3960
factor(LOCALIDAD)BOSA               -0.0368391  0.0158714  0.0202 *
factor(LOCALIDAD)CANDELARIA          0.1523503  0.0652931  0.0196 *
factor(LOCALIDAD)CHAPINERO           0.1233781  0.0294344  0.0000 ***
factor(LOCALIDAD)CIUDAD BOLIVAR      0.0216028  0.0177535  0.2236
factor(LOCALIDAD)ENGATIVA            0.0149868  0.0143218  0.2953
factor(LOCALIDAD)FONTIBON           -0.0365713  0.0178564  0.0405 *
factor(LOCALIDAD)KENNEDY             0.0030428  0.0128701  0.8131
factor(LOCALIDAD)LOS MARTIRES       -0.0026652  0.0323153  0.9342
factor(LOCALIDAD)PUENTE ARANDA      -0.0242716  0.0232618  0.2967
factor(LOCALIDAD)RAFAEL URIBE URIBE -0.0457782  0.0194978  0.0188 *
factor(LOCALIDAD)SAN CRISTOBAL      -0.0014528  0.0191730  0.9396
factor(LOCALIDAD)SANTA FE            0.0140673  0.0345618  0.6839
factor(LOCALIDAD)SUMAPAZ            -0.3584582  0.3757276  0.3400
factor(LOCALIDAD)TEUSAQUILLO         0.0627509  0.0270334  0.0202 *
factor(LOCALIDAD)TUNJUELITO         -0.0277615  0.0258886  0.2835
factor(LOCALIDAD)USAQUEN             0.0629672  0.0172899  0.0002 ***
factor(LOCALIDAD)USME                0.0218894  0.0225756  0.3322
factor(EDUCACION)1                   0.0015768  0.0503545  0.9750
factor(EDUCACION)2                  -0.1995175  0.1749081  0.2539
factor(EDUCACION)3                  -0.2157654  0.0137588  0.0000 ***
factor(EDUCACION)4                  -0.1584256  0.0138704  0.0000 ***
factor(EDUCACION)6                   0.1380860  0.0125619  0.0000 ***
```

```
factor(EDUCACION)7                    0.2524360   0.0170582   0.0000 ***
factor(EDUCACION)8                    0.2053568   0.0228444   0.0000 ***
factor(EDUCACION)9                    0.6127791   0.0112866   0.0000 ***
factor(EDUCACION)10                   0.7867583   0.0768685   0.0000 ***
factor(EDUCACION)11                   1.0477538   0.0174546   0.0000 ***
factor(EDUCACION)12                   1.1426932   0.0789654   0.0000 ***
factor(EDUCACION)13                   1.0284595   0.0237194   0.0000 ***
factor(EDUCACION)14                   1.9562865   0.5730419   0.0006 ***
factor(EDUCACION)15                   1.5289892   0.0894499   0.0000 ***
factor(NPCKP17)2                      0.1663312   0.0161731   0.0000 ***
factor(NPCKP17)3                     -0.2749943   0.0296696   0.0000 ***
factor(NPCKP17)4                      0.0638273   0.0142148   0.0000 ***
factor(NPCKP17)5                     -0.0665153   0.0093532   0.0000 ***
factor(NPCKP17)6                      0.3239058   0.0243375   0.0000 ***
factor(NPCKP17)7                     -0.0594646   0.1014797   0.5578
factor(NPCKP17)11                    -0.2832830   0.0757084   0.0001 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
        1

(Dispersion parameter for Tweedie family taken to be 72.27468)

    Null deviance: 7579644  on 93124  degrees of freedom
Residual deviance: 4905250  on 93076  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 7
```

B. Estimación del parámetro α utilizando en R tweedie.profile()

```
out3 <- tweedie.profile( BASE1$INGRESOS ~ BASE1$NVCBP11AA,
     weights = BASE1$FEX_C, do.plot=TRUE, data=BASE)

xi.est3 <- out3$xi.max
c("CI for xi" = out3$ci )
CI for xi1 CI for xi2
  1.694871   1.694876
c("MLE of phi"=out3$phi.max)
MLE of phi
  5.893099
> xi.est3
[1] 1.6948735
```

C. Estimación del parámetro α utilizando regresión lineal

```
mn <- with( BASE, tapply(BASE$INGRESOS, factor(BASE$NVCBP11AA), "mean"))
vr <- with( BASE, tapply(BASE$INGRESOS, factor(BASE$NVCBP11AA), "var"))
coef( lm( log(vr) ~ log(mn) ) )
(Intercept)      log(mn)
  0.9982243   1.8890502
```

D. Tamaño de la muestra en cada UPZ

UPZ	Tamaño de muestra	UPZ	Tamaño de muestra
20 de Julio	2776	Las Ferias	1588
Américas	2295	Las Margaritas	1967
Apogeo	2675	Los Alcazares	1688
Arborizadora	1905	Los Andes	1426
Parque Salitre + Doce de Octubre	1984	Los Cedros	2337
Bolivia	2356	Los Libertadores	1750
Bosa Central	2390	Lourdes	2123
Bosa Occidental	2385	Lucero	2177
Boyacá Real	2581	Marruecos	2066
Britalia	2216	Minuto de Dios	2326
Calandaima	1989	Modelia	3041
Carvajal	2350	Muzu	2487
Casa Blanca Suba	2493	Niza	1910
CHAPINERO: Chico Lago + Refugio	1104	Patio Bonito	2635
CHAPINERO: Pardo Rubio + Chapinero	976	Zona Industrial + Puente Aranda	1896
Monteblanco + Tesoro + Mochuelo	1954	Quinta Paredes	1710
Ciudad Jardín	2547	Quiroga	2487
Ciudad Montes	2320	Marco Fidel Suarez + San José	1761
Ciudad Salitre Occidental	1887	Restrepo	2407
Ciudad Salitre Oriental	1987	San Blas	2083
Comuneros	1994	San Cristobal Norte	2477
Corabastos	2277	San Francisco	2145
Diana Turbay	2128	San Isidro-Patios	1670
El Prado	2403	San Rafael	2546
El Provenir	1858	SANTA FE: Nieves + Sagrado Corazón	1095
El Rincon	2403	Santa Isabel	2183
Engativa	2106	Sosiego	2271
Santa Cecilia + Alamos + J. Botánico	1983	Suba	1934
Fontibón	2516	La Academia+ Guaymaral + San José	2094
Fontibón San Pablo	2677	SUBA: La floresta + La Alhambra	1442
Aeropuerto Eldorado + Capellanía	1900	Teusaquillo	2105
Galerías	1870	Parque Simón Bolivar + La Esmeralda	1537
Garces Navas	2009	Tibabuyes	2058
Gran Yomasa	2266	Tintal Norte	1507
Ismael Perdomo	2204	Toberin	2439
Kennedy Central	2554	Country Club + Santa Bárbara	1560
Castilla + Bavaria	2080	Verbenal + Paseo Los Libertadores	2044
La Candelaria	1834	USME: Alfonso Lopez + Ciudad Usme	1693
La Flora	2192	USME: Parque Entrenubes + Danubio	2052
La Gloria	2146	Venecia	2354
La Macarena	2182	Zona Franca	2473
La Saban	1979	NA	91411
Las Cruces	2270	Total	281182

E. Modelo Logit acumulado

```
> summary(nivel_mod)

Call:
vglm(formula = NPCHP4 ~ factor(NPCEP5) + NPCEP4 + INGRESOS +
    factor(NVCBP11AA) + factor(LOCALIDAD), family = cumulative(parallel = T),
    data = BASE2, weights = FEX_C)

Pearson residuals:
                              Min       1Q   Median       3Q      Max
logitlink(P[Y<=1])      -2480.1 -0.29548 -0.16066 -0.07935  1222828
logitlink(P[Y<=2])     -69968.8 -0.34337 -0.18159 -0.08839   891338
logitlink(P[Y<=3])    -124905.8 -1.40300 -0.64051 -0.21954  2287455
logitlink(P[Y<=4])   -6147420.7 -2.52410 -0.73652 -0.15383    13772
logitlink(P[Y<=5])    -150250.3 -1.56682 -0.10089  2.91056 36050242
logitlink(P[Y<=6])    -428258.4 -1.12079  0.62904  1.75780 10754628
logitlink(P[Y<=7])   -1153084.3 -0.99185  0.60731  1.42143  6518244
logitlink(P[Y<=8])   -4475605.0 -0.99123  0.59613  1.31867  5101298
logitlink(P[Y<=9])   -1842589.3  0.20002  0.40629  0.95352 35744154
logitlink(P[Y<=10]) -3477847.4  0.19587  0.39427  0.85550 28628367
logitlink(P[Y<=11])      -542.8  0.11553  0.20855  0.39912 21758743
logitlink(P[Y<=12])      -500.8  0.10998  0.19719  0.37668 15404026
logitlink(P[Y<=13])      -726.7  0.02287  0.04019  0.07745   315072
logitlink(P[Y<=14])      -321.7  0.02185  0.03837  0.07348   306856

Coefficients:
                                        Estimate    Std. Error  Pr(>|z|)
(Intercept):1                     -5.7305358959  0.0128464334 0.000000 ***
(Intercept):2                     -5.6505976666  0.0126717303 0.000000 ***
(Intercept):3                     -2.4134859946  0.0104218381 0.000000 ***
(Intercept):4                     -1.5620044138  0.0103507077 0.000000 ***
(Intercept):5                      0.2860510311  0.0103078001 0.000000 ***
(Intercept):6                      0.9761688115  0.0103211748 0.000000 ***
(Intercept):7                      1.3405402908  0.0103349242 0.000000 ***
(Intercept):8                      1.5471687050  0.0103450363 0.000000 ***
(Intercept):9                      3.6197390024  0.0105998854 0.000000 ***
(Intercept):10                     3.6564294573  0.0106073789 0.000000 ***
(Intercept):11                     5.2170330190  0.0111497673 0.000000 ***
(Intercept):12                     5.2974997441  0.0111927005 0.000000 ***
(Intercept):13                     9.1144748278  0.0184539726 0.000000 ***
(Intercept):14                     9.1399368272  0.0185446124 0.000000 ***
factor(NPCEP5)2                   -0.3886336781  0.0020480361 0.000000 ***
NPCEP4                             0.0453127754  0.0000872739 0.000000 ***
INGRESOS                          -0.0003439633  0.0000006187 0.000000 ***
factor(NVCBP11AA)2                -0.5747366258  0.0046372495 0.000000 ***
factor(NVCBP11AA)3                -1.7565698703  0.0052175699 0.000000 ***
factor(NVCBP11AA)4                -3.2110999081  0.0066299834 0.000000 ***
factor(NVCBP11AA)5                -3.3722941366  0.0083463828 0.000000 ***
factor(NVCBP11AA)6                -3.3350750481  0.0115302460 0.000000 ***
factor(LOCALIDAD)BARRIOS UNIDOS   -0.2413630836  0.0098957360 0.000000 ***
```

```
factor(LOCALIDAD)BOSA                0.0950487221  0.0092516749 0.000000 ***
factor(LOCALIDAD)CANDELARIA         -1.3858111403  0.0204275491 0.000000 ***
factor(LOCALIDAD)CHAPINERO          -0.3351014304  0.0119442114 0.000000 ***
factor(LOCALIDAD)CIUDAD BOLIVAR      0.0763151871  0.0094860731 0.000000 ***
factor(LOCALIDAD)ENGATIVA           -0.3559870252  0.0088134236 0.000000 ***
factor(LOCALIDAD)FONTIBON           -0.0291443437  0.0094396702 0.002019 **
factor(LOCALIDAD)KENNEDY            -0.0073163602  0.0087284675 0.401908
factor(LOCALIDAD)LOS MARTIRES        0.2467867454  0.0119384428 0.000000 ***
factor(LOCALIDAD)PUENTE ARANDA       0.0665123400  0.0101345620 0.000000 ***
factor(LOCALIDAD)RAFAEL URIBE URIBE  0.2747474621  0.0096415681 0.000000 ***
factor(LOCALIDAD)SAN CRISTOBAL       0.0805163170  0.0096963508 0.000000 ***
factor(LOCALIDAD)SANTA FE           -0.1747404928  0.0126733761 0.000000 ***
factor(LOCALIDAD)SUBA               -0.1065536329  0.0087630134 0.000000 ***
factor(LOCALIDAD)SUMAPAZ             0.3854990609  0.1069520670 0.000313 ***
factor(LOCALIDAD)TEUSAQUILLO        -0.3599752200  0.0114743859 0.000000 ***
factor(LOCALIDAD)TUNJUELITO          0.0328212185  0.0108353271 0.002453 **
factor(LOCALIDAD)USAQUEN             0.0171281324  0.0094754036 0.070662 .
factor(LOCALIDAD)USME                0.1288090738  0.0102673732 0.000000 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1
       1

Number of linear predictors:  14

Residual deviance: 11195655 on 1216853 degrees of freedom

Log-likelihood: -5597827 on 1216853 degrees of freedom

Number of Fisher scoring iterations: 9

Warning: Hauck-Donner effect detected in the following estimate(s):
'(Intercept):1', '(Intercept):13', '(Intercept):14'


Exponentiated coefficients:
                    factor(NPCEP5)2                                NPCEP4
                         0.67798258                         1.04635508
                            INGRESOS               factor(NVCBP11AA)2
                         0.99965610                         0.56285309
                  factor(NVCBP11AA)3               factor(NVCBP11AA)4
                         0.17263601                         0.04031225
                  factor(NVCBP11AA)5               factor(NVCBP11AA)6
                         0.03431083                         0.03561191
    factor(LOCALIDAD)BARRIOS UNIDOS          factor(LOCALIDAD)BOSA
                         0.78555635                         1.09971243
        factor(LOCALIDAD)CANDELARIA       factor(LOCALIDAD)CHAPINERO
                         0.25012083                         0.71526553
    factor(LOCALIDAD)CIUDAD BOLIVAR        factor(LOCALIDAD)ENGATIVA
                         1.07930270                         0.70048171
          factor(LOCALIDAD)FONTIBON         factor(LOCALIDAD)KENNEDY
                         0.97127626                         0.99271034
      factor(LOCALIDAD)LOS MARTIRES    factor(LOCALIDAD)PUENTE ARANDA
                         1.27990614                         1.06877415
```

```
factor(LOCALIDAD)RAFAEL URIBE URIBE
                        1.31619824
          factor(LOCALIDAD)SANTA FE
                        0.83967489
           factor(LOCALIDAD)SUMAPAZ
                        1.47034793
        factor(LOCALIDAD)TUNJUELITO
                        1.03336578
              factor(LOCALIDAD)USME
                        1.13747293
```

```
factor(LOCALIDAD)SAN CRISTOBAL
                        1.08384653
          factor(LOCALIDAD)SUBA
                        0.89892683
  factor(LOCALIDAD)TEUSAQUILLO
                        0.69769361
       factor(LOCALIDAD)USAQUEN
                        1.01727566
```

Referencias

[1] Buck, S.F, & Smyth, G.K (1960). *A Method of Estimation of Missing Values in Multivariate Data* , Vol. 22, No. 2 (1960), pp. 302-306

[2] De Jong, P., Heller, G. Z. (2008). *Generalized linear models for insurance data*,vol.10. Cambridge: Cambridge University Press.

[3] Dunn, P.K (2017). *Tweedie: Tweedie exponential family models*,URL. `https://CRAN.R-project.org/package=tweedie`. R package version 2.3.2

[4] Dunn, P.K, & Smyth, G.K (1996). *Randomized quantile residuals*, Journal of Computational and Graphical Statistics

[5] Dunn, P.K, & Smyth, G.K (2005). *Series Evaluation of Tweedie Exponential Dispersion Model Densities*, Statistics and Computing

[6] Dunn, P.K, & Smyth, G.K (2018). *Generalized linear models with examples in R*, New York.

[7] Heller, G. Stasinopoulos M and Rigby R.A. (2006). *The zero-adjusted Inverse Gaussian distribution as a model for insurance claims.*, pp 226-233, Galway, Ireland.

[8] Hernández, G.; Lasso, F. (2003). *Estimación de la relación entre salario mínimo y empleo en Colombia: 1984-2000*, De Revista de Economía del Rosario, vol. 6, núm. 2, pp. 11-17.

[9] Núñez, J.; Bonilla, J (2001). *¿Quiénes se perjudican con el salario mínimo en Colombia?*,Coyuntura Social, núm. 24, pp. 87-110

[10] Rigby, R. A, Stasinopoulos D. M. (2005). *Generalized additive models for location, scale and shape*, pp 507-554.

[11] Rigby, R. A, Stasinopoulos, D. M. (2010). *A flexible regression approach using GAMLSS in R*,London Metropolitan University.

[12] Rivas, Luis Arturo. (2017). *Elaboración de tesis.*, Estrutura y metodología. Ed Trillas, Mexico.

[13] S.N. Wood (2006). *Generalized additive models: an introduction with R*, New York.

[14] Secretaría Distrital de Planeación de la Alcaldía Mayor de Bogotá, el Departamento Administrativo Nacional de Estadística (DANE) y la Gobernación de Cundinamarca. (2018). Encuesta multiproposito 2017, Bogotá. Recuperado de http://www.sdp.gov.co/gestion-estudios-estrategicos/estudios-macro/encuesta-multiproposito/encuesta-multiproposito-2017