Universidad La Salle Oaxaca
Escuela de Ingenierías y Arquitectura

Ingeniería de Software y Sistemas Computacionales
Tercer Semestre

Análisis estadístico del Módulo de Condiciones Socioeconómicas 2015

Proyecto final

Curso: Probabilidad y Estadística

Presentan:

Cruz Aguilar Juan Daniel

Nieblas Hernández Glenn

Velasco Chincoya Karen Itzel

Profesor:

Juliho David Castillo Colmenares

Santa Cruz Xoxocotlán, Oaxaca. 05 de diciembre de 2017.

1. Planteamiento del problema

El Módulo de Condiciones Socioeconómicas de la Encuesta Nacional de Ingresos y Gastos de los Hogares (MCS) es un proyecto realizado por el Instituto Nacional de Estadística y Geografía (INEGI), tiene la finalidad de captar la información sociodemográfica, de vivienda, ocupación e ingresos de la Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH), además de los temas específicos para la medición multidimensional de la pobreza.

El documento "Descripción de la base de datos" provee una guía a los diversos usuarios con la información necesaria para el análisis de los microdatos. De acuerdo con el objetivo del Módulo de Condiciones socioeconómicas, la información estadística que produce el INEGI y que pone a disposición del estado y la sociedad, contribuye al desarrollo del país, ya que permite que las autoridades y representantes de los diversos sectores tengan un mejor conocimiento de la realidad para fundamentar sus decisiones, así como para evaluar los resultados de su desempeño. Además, es un insumo fundamental para las investigaciones académicas que coadyuvan a la comprensión del progreso del país y nuestro entorno.

En este contexto, la base de datos del Módulo de Condiciones Socioeconómicas de la ENIGH nos proporciona datos específicos del monto, la estructura y la distribución de los ingresos de los hogares. También otorga información sobre la composición familiar de los hogares, salud, educación, seguridad social, calidad y espacios de la vivienda, servicios básicos, alimentación y cohesión social, así como de la actividad económica de cada uno de sus integrantes.

Con base en el modelo entidad-relación y los atributos de cada variable, elegimos 26 elementos para medir el grado de relación que tienen con el ingreso trimestral por vivienda. Este dato se obtuvo a partir de la suma de los ingresos individuales trimestrales por vivienda; el objetivo es predecir el ingreso trimestral por vivienda a partir de algunas variables presentadas por el modelo relacional.

2. Marco teórico.

Conformación de la base de datos.

La base de datos del módulo de condiciones socioeconómicas está conformada de nueve tablas normalizadas y relacionadas, en ellas se encuentran las respuestas de las encuestas realizadas en el año 2015; estas tablas guardan relaciones uno a uno y uno a muchos según sea el caso de los datos, por ejemplo una hogar corresponde a una vivienda, en cambio, una vivienda puede tener uno o más habitantes.

Tablas de la base de datos.

1) Viviendas. En esta tabla se encuentran contenidas las características de las viviendas que habitan los integrantes de los hogares encuestados, tales como material de pared, piso y techo.

2) Hogares. En esta tabla se encuentran las características de los hogares que habitan los integrantes de los mismos, como el tipo de alimentación, muebles, entre otros.

3) Población. Esta tabla proviene de la tabla Hogares. Identifica las características sociodemográficas de los integrantes del hogar; y el acceso a las instituciones de salud.

4) Gastos hogar. Esta tabla contiene las estimaciones del alquiler del hogar.

5) Ingresos. Esta tabla permite identificar los ingresos y percepciones financieras y de capital de cada uno de los integrantes del hogar, por diversos conceptos.

6) Gastos persona Esta tabla permite identificar los gastos realizados por cada integrante del hogar.

7) Trabajos. Esta tabla muestra la condición de actividad de los integrantes del hogar de 12 o más años y algunas características ocupacionales durante el periodo de referencia.

8) Agro. Esta tabla muestra la información de los trabajadores independientes, mayores de 12 años que tienen en el hogar negocios dedicados a las actividades agrícolas, forestales y de tala, además de actividades de cría, explotación y productos derivados de la pesca y caza.

9) No agro. Esta tabla muestra a los trabajadores independientes mayores de 12 años, que tienen negocios en el hogar dedicados a las actividades industriales, comerciales y de servicios.

10)Concentrado hogar En esta tabla se encuentran las variables construidas a partir de las otras tablas de la base de datos. Registra el resumen concentrado por hogar, de ingresos y gastos en toda modalidad posible. Además todos los ingresos y gastos registrados en esta tabla son trimestrales.

Conceptos estadísticos.

Regresión lineal.

La regresión es una técnica estadística que consiste en calcular la similitud entre dos variables en forma de función matemática.

Regresión lineal múltiple.

La regresión lineal múltiple es una técnica estadística para comprobar hipótesis y relaciones causales.
Estadístico R-cuadrado.
Indica el porcentaje de la varianza de la variable dependiente explicado por el conjunto de variables independientes. Cuanto mayor sea la R-cuadrado en el rango [0,1] , más explicativo y mejor es el modelo causal.

3. Metodología de resolución.

La técnica utilizada para la resolución del problema fue la regresión lineal múltiple, utilizando el paquete statsmodel, perteneciente a pandas y la clase OLS (Ordinary Least Squares), que implementa el método de mínimos cuadrados.

El primer paso para la resolución del problema fue determinar los pesos de 26 variables distintas con ayuda de scikit-learn, los resultados indican cuáles son las variables con mayor peso, es decir, las que se relacionan con la variable de salida. El resultado fueron 2 variables: el número de días por semana que se consume frutas y huevo, con base en este resultado, formulamos la hipótesis nula: *El número de días que se consumen frutas y huevo está relacionado con los ingresos trimestrales por vivienda.*

Con base en los pesos de las 26 variables, se implementó con ayuda de Python el método de regresión lineal, cuyo resultado más significativo fue R-cuadrada, que tomó el valor de 0.016.

4. Documentación del Software (Python).

Python es un lenguaje de programación interpretado cuya filosofía hace hincapié en una sintaxis que favorezca un código legible. Actualmente se utiliza en el análisis de una gran cantidad de datos. En el desarrollo del proyecto se utilizaron dos librerías de Python: Scikit-learn y pandas. Scikit-learn es una librería de código abierto enfocado en el aprendizaje automático (machine learning) y Pandas es una biblioteca, igual de código abierto que proporciona estructuras de datos y herramientas de análisis de datos de alto rendimiento, Pandas permite trabajar con datos en distintos formatos.
Herramientas utilizadas.
Para leer y escribir datos en memoria, se utilizaron archivos CSV y archivos de texto, Microsoft Excel, bases de datos SQL y el formato HDF5.

5. Presentación de resultados.

Una vez determinados los pesos de las 26 variables, obtuvimos dos variables aparentemente relacionadas con los ingresos trimestrales por vivienda: el número de días que se consumen frutas y huevo en una vivienda. Con ayuda de Python se implementó el método de regresión lineal, cuyo resultado fue R-cuadrada = 0.016, debido a que el valor es mucho más cercano a cero que a uno, rechazamos la hipótesis nula. Por lo tanto, el número de días que se consumen frutas y huevo no está relacionado con los ingresos trimestrales de una familia.

Referencias

Módulo de Condiciones Socioeconómicas 2015. Encuesta Nacional de Ingresos y Gastos de los Hogares. Descripción de la base de datos; Módulo de Condiciones Socioeconómicas. Encuesta Nacional de Ingresos y Gastos de los Hogares 2015. Descripción de la base de datos.

Montero, R. (2016). Modelos de regresión lineal múltiple. ESPAÑA: Universidad de Granada.

NUMFocus. (2017). Biblioteca de análisis de datos de Python . diciembre 04, 2017, de pydata Sitio web: https://pandas.pydata.org/