Logotipo

Descripción generada automáticamente

Universidad Austral

Maestría en Economía Aplicada

**Trabajo Práctico N°1**

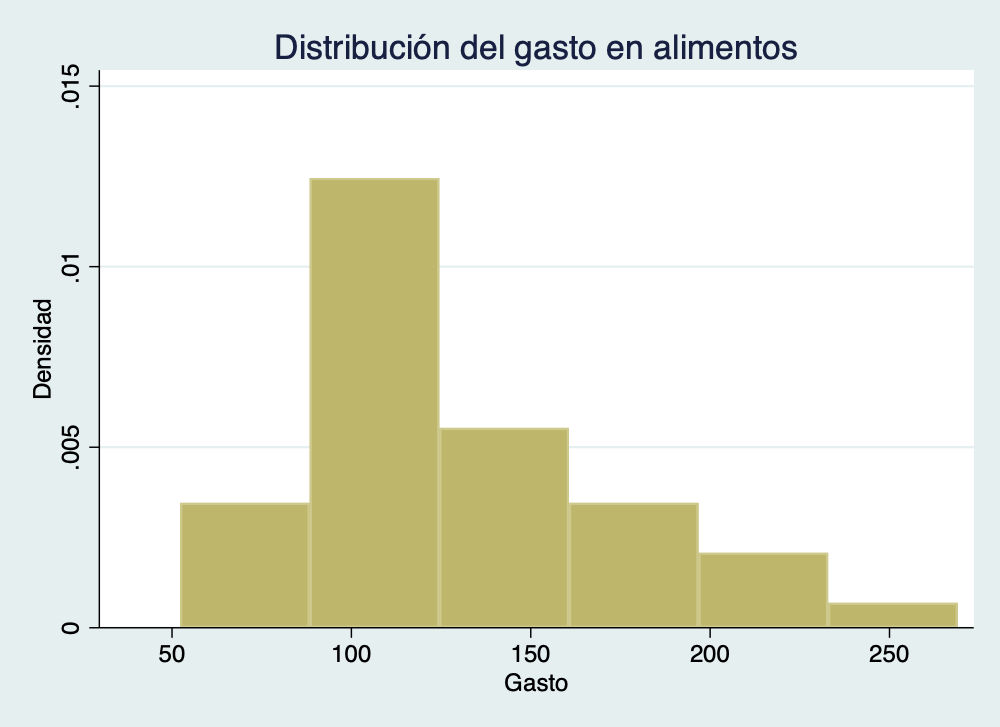
**Econometría**

**Autor:**

Bocco, Alessio

1. **Grafique la distribución de la variable y utilizando un histograma. Calcule medidas de dispersión, asimetría y curtosis. Interprete.**

La Figura 1 muestra un histograma con la distribución de la variable gastos en alimentos. Se puede observar que la distribución está centrada en valores cercanos a U$S 130 pero no es simétrica ya que los datos no se distribuyen de manera uniforme alrededor de la masa de la distribución. Las medidas de posición, media y mediana, toman valores de U$S 130 y U$S 120, respectivamente, lo que indica que la distribución es asimétrica a derecha (mediana < media). No obstante esto deberá se comprobado utilizando medidas de dispersión. El rango de la variable se encuentra entre aproximadamente U$S 50 y U$S 260.

Figura 1 Distribución de la variable gasto en alimentos (U$S). 

La Tabla 1 muestra las medidas de dispersión y forma principales para la distribución de la variable.

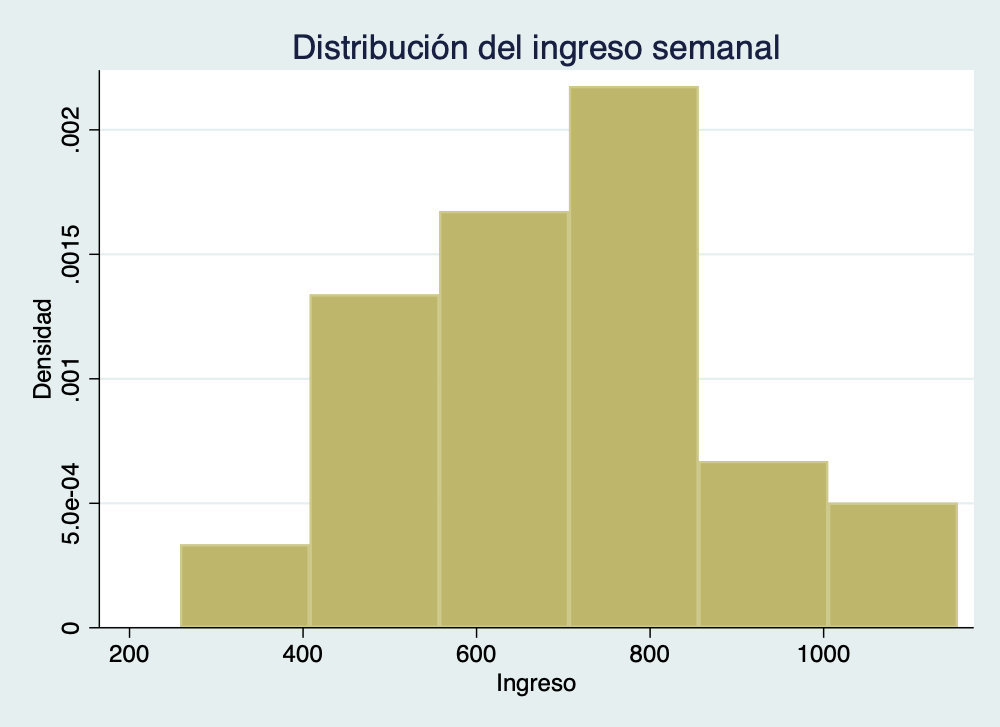
Tabla 1 Medidas de dispersión y forma de la distribución de la variable gastoalimentos

|  |  |
| --- | --- |
| **Métrica** | **Valor** |
| Varianza | 2039.296 |
| Desvío estándar | 45.158 |
| Coef Asimetría | 1.034 |
| Coef Curtosis | 4.129 |

Las medidas de variabilidad se utilizan para describir la dispersión o concentración, es decir, la cercanía de un conjunto de datos. La varianza corresponde al promedio del cuadrado de la distancia entre cada observación con la media, por lo tanto, no se encuentra en las mismas unidades que la variable original y debe interpretarse con cautela. El desvío estándar, en cambio, corresponde a la raíz cuadrada positiva de la varianza por lo que recupera la unidad original de la variable. Un desvío estándar de 45 indica que el gasto en alimentos se desvía en promedio U$S 45 de la media de la distribución. En la Tabla 1 también se incluyen medidas de forma, asimetría y curtosis. La asimetría refiere a la deformación horizontal del conjunto de datos, muy relacionada a lo mencionado sobre las medidas de posición, mientras que la deformación vertical está explicada por la curtosis. Estos momentos centrados están elevados a la tercer y cuarta potencia, respectivamente, por lo que sus unidades poco tienen que ver con las unidades de la variable original. Sin embargo, es posible calcular coeficientes al dividirlos por el desvío estándar elevado a sendas potencias. Esto permite recuperar la unidad original de las variables y dotar a los coeficientes de interpretabilidad. Dado que el coeficiente de asimetría es mayor que 0, la distribución es asimétrica a derecha como sugerían las medidas de posición. Que sea asimétrica a derecha quiere decir que la mayor cantidad de observaciones se encuentra hacia la izquierda de la distribución. Con respecto a la deformación vertical, un valor mayor a 3 indica una distribución leptocúrtica (puntiaguda), es decir, con datos más concentrados en la masa de la distribución y colas “más livianas”.

1. **Rehaga el punto 1) anterior pero con la variable x. Compare con la distribución de y.**

La Figura 2 muestra la distribución del ingreso semanal en U$S. *A priori* parecería que la distribución del ingreso es asimétrica a izquierda dado que la media (US$ 698) es menor que la mediana (U$S 712). El rango de la variable se encuentra entre U$S 258 y U$S 1154.

Figura 2 Distribución de la variable ingreso semanal (U$S). 

La Tabla 2 muestra las medidas de dispersión y forma principales para la distribución de la variable.

Tabla 2 Medidas de dispersión y forma de la distribución de la variable ingresosemanal

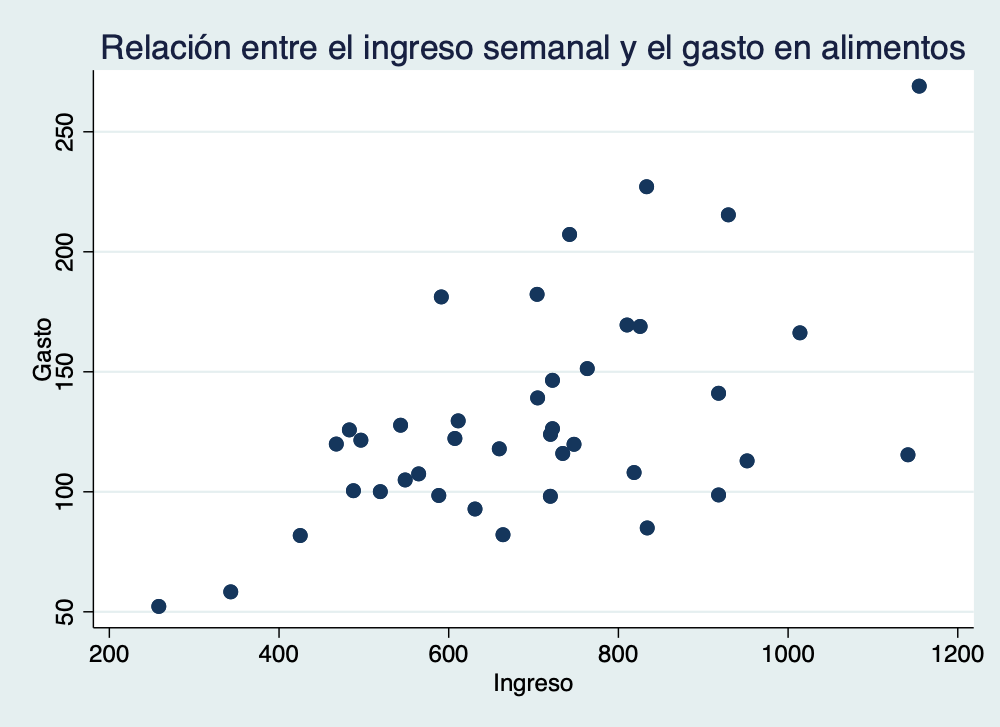
|  |  |
| --- | --- |
| **Métrica** | **Valor** |
| Varianza | 39293.92 |
| Desvío estándar | 198.226 |
| Coef Asimetría | 0.218 |
| Coef Curtosis | 2.985 |

Las medidas de dispersión indican que la variable *ingresosemanal* se desvía en aproximadamente U$S 198 de la media de la distribución. Sin embargo, la deformación horizontal sugerida por las medidas de posición no se confirma por el coeficiente de asimetría ya que éste toma valores superiores a 0 por lo que los valores se concentran hacia la izquierda de la distribución. Es decir, hay una mayor cantidad de observaciones con ingresos menores a la media. Sin embargo, el valor de asimetría es relativamente bajo y cercano a 0 por lo que podría considerarse bastante simétrica. Con respecto a la deformación vertical, un valor de curtosis tan próximo a 3 sugiere una distribución mesocúrtica, es decir, con una distribución pareja entre el centro y las colas de la distribución.

Al comparar las dos variables, *gastoalimentos e ingresosemanal,* se observa que los ingresos son cuatro veces más variables que los gastos en alimentos. A su vez, la distribución de los ingresos es más simétrica horizontal y verticalmente por lo que se intuye que los ingresos semanales están mejor distribuidos en la muestra. En cambio, los gastos en alimentos se concentran en valores más bajos de la distribución lo que indica que existe un determinado nivel de gasto que se debe gastar indefectiblemente e independientemente del nivel de ingreso.

1. **Realice un diagrama de dispersión entre el gasto en alimentos y el ingreso semanal. Comente.**

La Figura 3 muestra la relación entre ambas variables, ubicándose en el eje horizontal el ingreso semanal en dólares y en el vertical, el gasto de alimentos en dólares. Hasta valores de ingreso de aproximadamente U$S 600 se observa una relación bastante lineal entre el ingreso y gasto en alimentos. A partir de este umbral hay una mayor variabilidad ya que para mimos niveles de ingreso hay distintos niveles de gastos en alimentos. La variabilidad aumenta considerablemente a medida que aumentan los ingresos. *A priori* parecería existir una relación entre ingreso y gasto que tiene un comportamiento lineal hasta un determinado valor de ingreso por encima del cual se vuelven independientes.

Figura 3 Diagrama de dispersión entre el ingreso semanal (eje horizontal) y el gasto semanal en alimentos (eje vertical). 

De hecho, si se divide la población en dos submuestras utilizando el umbral mencionado, el coeficiente de correlación de Pearson es el siguiente (Tabla 3).

Tabla 3 Correlación lineal entre las variables para las dos submuestras.

|  |  |
| --- | --- |
| **Muestra** | **Correlación** |
| Muestra 1: ingreso < 600 (n = 13) | 0.753 |
| Muestra 2: ingreso < 600 (n = 27) | 0.395 |

Como se observa en los valores del coeficiente, en ingresos bajos, existe una relación lineal entre ambas variables que se pierde por encima del umbral.

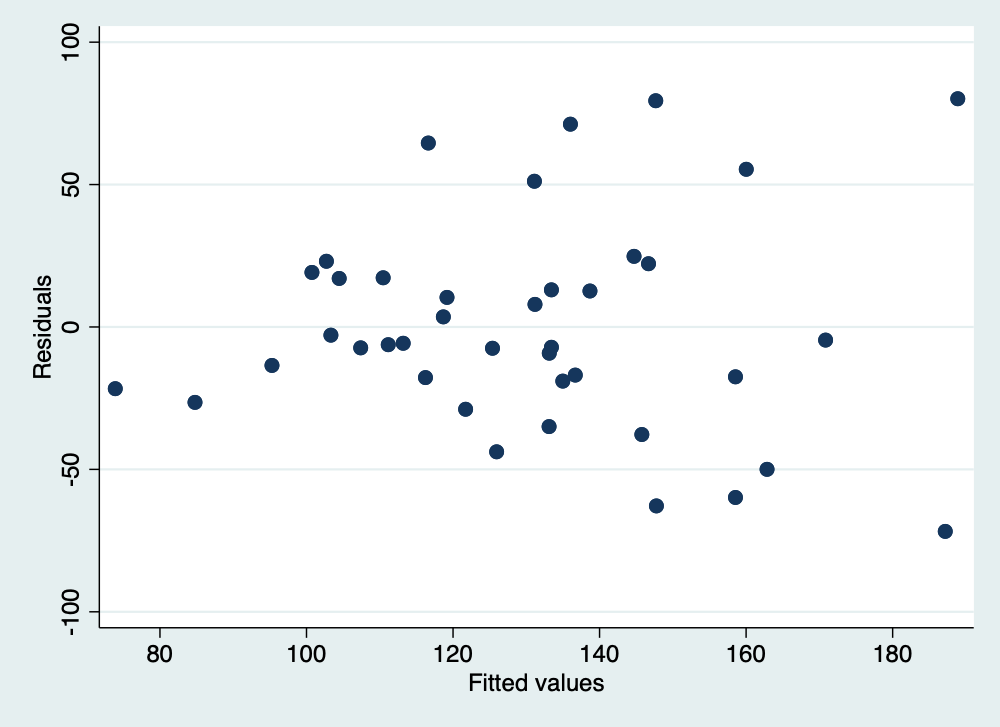
1. **Estime por OLS los valores de b1 y b2 para el gasto de los hogares donde la línea de regresión es: 𝑦̂𝑡 = 𝑏1 + 𝑏2 ∙ 𝑥𝑡 ¿Cuál es la dirección de la relación entre ambas variables? Interprete el valor de los estimadores.**

El modelo lineal ajustado es el siguiente:

Donde, *x* e *y* corresponden a los ingresos y gastos en alimentos, respectivamente. Un valor de β1 de 40.767 indica que la familia tiene un gasto en alimentos fijo de U$S independiente de su nivel de ingreso. Es importante resaltar que pese a que este coeficiente no es estadísticamente significativo para un α de 5%, las familias deben gastar un valor mínimo de subsistencia. El β2 indica cuánto aumentarán los gastos en alimentos por el incremento de una unidad de los ingresos semanales. Es decir, un aumento de un dólar en los ingresos semanales implicará un aumento de 0.12 dólares del gasto en alimentos. Al ser β2 mayor a 0, incrementos en los ingresos se traducen en aumentos de los gastos por la que la relación es positiva.

1. **Analice gráficamente si los residuos se comportan bajo el supuesto de homoscedasticidad**

La Figura 4 muestra el diagrama de dispersión entre los valores ajustados por el modelo (*yhat*) y los residuos del modelo lineal.

Figura 4 Diagrama de dispersión entre los valores ajustados (eje horizontal) y los residuos del modelo lineal (eje vertical). 

Idealmente los puntos deberían distribuirse al azar por encima y por debajo de cero sin mostrar un patrón determinado. Sin embargo, puede verse que los puntos no se disponen al azar sino que tienen un patrón de abanico. Es decir, a medida que los valores ajustados aumentan, aumenta la variabilidad de los puntos. Un comportamiento de este tipo es una indicación de que se está violando uno de los supuestos del modelo lineal, la homoscedasticidad.

1. **Realice un test de hipótesis para testear homoscedasticidad. ¿Qué puede concluir?**

Si bien la prueba gráfica es un indicador de una posible violación de los supuestos, ésta debe confirmarse con un test de hipótesis. Para ello se aplicaron dos pruebas estadísticas. El primero fue un contraste general de heteroscedasticidad como es el test de White, que es una prueba de multiplicadores de Lagrange (LM test). El test de White arrojó un *p-valor* de 0.007 por lo que se rechaza la hipótesis nula de homoscedasticidad. El segundo test fue la prueba de contraste de Breusch-Pagan, que también es un LM test. Para este test la hipótesis nula es una varianza constante en los residuos pero ésta también se rechaza dado que el *p-valor* es de 0.0008. Al ser rechazadas las dos hipótesis nulas de los tests, existe evidencia para confinar el diagnóstico grafico y afirmar que no existe homoscedasticidad.

1. **Analice el resto de los resultados que aparecen en la salida de la regresión en Stata**

* **SS:** suma de cuadrados relacionadas con las tres fuentes de variación: modelos, residuos y total. SSTotal es la variabilidad total alrededor de la media . SSResidual es la suma de los errores cuadrados de la predicción . SSModel es la mejora al utilizar el valor de *yhat* en lugar de la media de *y .* SSModel/SSTotal es igual al valor del R2. Esto se debe a que R2 es la proporción de la varianza explicada por las variables independientes del modelo.
* **df:** son los grados de libertad asociados a cada una de las fuente de varianza. dfTotal, la varianza total es igual a N-1 grados de libertad, 40 observaciones – 1. Los df del modelo corresponden al número de predictores menos uno. Dado que se cuenta con una variable independiente más la constante, dfModel = 2 – 1 = 1. dfResidual surge de restar dfTotal (39) – dfModel (1) = 38.
* **MS:** son la suma de cuadrados que surgen de dividir cada uno de los SS por sus respectivos grados de libertad.
* **Number of obs:** cantidad de datos utilizados para ajustar el modelo
* **F y Prob > F:** El valor de F es la Suma de cuadrados del modelo (MSModel) dividido por la suma de cuadrados de los residuos (MSResiduals). Este valor del estadístico de Fischer tiene asociado un *p-valor*. Si el este valor es menor al valor de α se puede concluir que hay evidencia estadísticamente significativa para afirmar que la variable ingreso predice al gasto semanal en alimentos. En este caso, un valor de 0.0002 los confirma.
* **R2:** es una medida de la proporción de la varianza en la variable dependiente (*gastoalimentos*) que es explicada por la variable independiente (*ingresosemanal*). El valor indica que el 37% de la variabilidad del gasto es predicha por el ingreso.
* **R2 ajustado:** se trata de un indicador más honesto de bondad ya que considera la cantidad de observaciones y parámetros del modelo en su fórmula de cálculo y permite hacer inferencia sobre la población. Pese a que la cantidad de datos es pequeña, el hecho de tener poco parámetro hace que no haya tanta diferencia entre las dos métricas anteriores.
* **Root MSE:** es la desviación estándar del término de error y se calcula como la raíz cuadrado del MSResidual.
* **Coef:** muestra el valor de los coeficientes del modelo lineal: β1 el término constante (\_cons) y β2 (ingresosemanalxt)
  + **Std.** **Err.:** error estándar asociado a la estimación de los coeficientes. Se utilizarán para probar si éstos son distintos de 0. Al dividir el valor del coeficiente por el error estándar se obtiene el estadístico *t.* También se utilizan para el cálculo del intervalo de confianza.
  + **t y P > |t|:** estadístico y *p-valor* de la prueba de hipótesis de dos colas usada para determinar si el coeficiente es distinto de 0. Valores por debajo de α son estadísticamente significativos. En este caso, β2 tuvo un *p-valor* inferior y fue significativo, mientras que β1, no. Sin embargo, como se mencionó anteriormente, el modelo teórico sobre el que se basa el modelo estadístico supone un consumo mínimo de subsistencia, por lo tanto la constante permanece dentro del modelo.
  + **Intervalo de confianza:** corresponde al intervalo de confianza al 95% de los valores de los coeficientes. Se utiliza para entender cuán alto o bajo es el valor del coeficiente con respecto al que podría tener la población. Si el intervalo de confianza abarca a 0, como en el término constante, dicho término no sería significativo. Con respecto a β2, un intervalo de .0664651 - .1901121, indica que una unidad más de ingreso podría implicar un aumento del gasto en alimentos tan bajo como U$S 0.06 o tan alto cómo U$S 0.19. Esta gran variabilidad hace muy riesgoso emitir conclusiones para la población ya que el ingreso podría tener un efecto muy dispar sobre el gasto.

1. **Calcule los valores estimados del gasto en alimentos y los residuos para cada observación y verifique si la suma de los residuos se aproxima a cero.**

Al sumar los residuos el resultado es muy pequeño y próximo a 0. Además, al utilizar el test de Shapiro-Wilk, se comprobó la normalidad de su distribución (*p-valor = 0.22253*). Por lo tanto, los residuos tienen distribución N~(0, 37.805).

1. Estime la elasticidad ingreso de la demanda en los valores medios de x e y. Según el valor de la elasticidad, ¿qué clase de bien son los alimentos?

Al aplicar la formula anterior se obtiene una elasticidad de 0.687. Dado que la elasticidad es mayor a 0, los alimentos son un bien normal. Sin embargo, al ser ésta menor a 1 se los considera bienes necesarios o de primera necesidad.

1. **Estime cuál sería el gasto en alimentos de un individuo que gana $ 750 por semana.**

La estimación del nivel de gasto se obtiene al reemplazar el valor de *x* por 750 en el modelo lineal.