Trabajo Práctico 1

Economía Aplicada

Universidad de
SanAndrés

Alumnos:
Juan Diego Barnes
Fransisco Legaspe
Rodrigo Martin
Diego Fasan

Profesor: Martín Rossi
Tutores: Gastón García Zavaleta
Tomás Pacheco

1. Repaso teórico

1. **Bajo ciertos supuestos sobre el método, el modelo de MCO es el mejor estimador lineal e insesgado.**

   Por el Teorema de Gauss-Markov, bajo los supuestos clásicos, el estimador de Mínimos Cuadrados Ordinarios (MCO) es el mejor (de menor varianza) estimador lineal e insesgado.

   Para ser mas precisos, dado un modelo lineal:

   $$Y = X\beta + u_i.$$

   Donde Y es es un vector columna de dimensiones $n1$ que contiene las observaciones de la variable dependiente. X es una matriz de dimensiones $n(k+1)$ que contiene las observaciones de las variables independientes (incluyendo una columna de unos para el intercepto). es un vector columna de dimensiones $(k+1)1$ de coeficientes de regresión, y u es un vector columna de dimensiones $n1$ que contiene los términos de error. Los supuestos clásicos son:

   - Linealidad (en los parámetros): $Y = X\beta + u$
   - Exogeneidad: $E(u|X) = 0$
   - Homocedasticidad: $V(u|X) = \sigma I_n$
   - No correlación Serial: $Cov(u_i, u_j|X) = 0, \ \forall \ i, j$
   - No multicoliliealidad Perfecta: $\rho(X) = K$.

2. **¿Cuál es el supuesto que creen que es necesario hacer para que el estimador de Mínimos Cuadrados Ordinarios sea insesgado? ¿Es el mismo supuesto que necesitamos para garantizar la existencia del estimador?**

   El supuesto necesario para que el estimador de mínimos cuadrados sea instigado es el de exogeneidad, usualmente para su demostración formal se utiliza también el de linealidad, pero este no es necesario (**?**). Sin embargo, el supuesto crucial para que el estimador exista es el de No Multicolinealidad Perfecta, el estimador de MCO se define como $\hat{\beta} = (X'X)^{-1}X'Y$, de modo de que si existe multicolinealidad perfecta, $(X'X)^{-1}$ no esta definido y por lo tanto tampoco el estimador.

3. **¿Cuál(es) es(son) los problemas de levantar supuestos tales como el de homocedasticidad o no correlación serial?**

   Estos no tienen ningún efecto sobre la insesgadez, pero si sobre la eficiencia del estimador. Bajo heteroceidasticidad y/o correlación serial, no se cumple el Teorema de Gauss-Markov, el estimador tiene una mayor varianza por lo que ya no es el mejor (en términos de varianza mínima) estimador lineal e insesgado.

4. **¿Por qué decimos que $\hat{\beta}$ tiene distribución normal cuando suponemos $u_i \sim N(0, \sigma^2)$)?**

   Si el termino de error tiene distribución normal, debido a la linealidad del modelo, Y también es normal al ser una transformación lineal de u_i. Y del mismo modo, $\hat{\beta} = ((X'X)^{-1}X'Y)$ es una combinación lineal de Y por lo tanto hereda también la normalidad de u_i.

5. **Para hacer inferencia siempre se necesita suponer la distribución del término de error.**

   Se puede suponer normalidad para realizar inferencia, pero no es siempre necesario. La propiedad de normalidad asintótica del estimador, a pesar de solo ser demostrable cuando la muestra tiende hacia infinito, proporciona una excelente aproximación cuando el número de observaciones es lo suficientemente grande. Esto hace que el supuesto de normalidad en el error no sea estrictamente necesario, ya que asintóticamente la normalidad del estimador es un resultado que aparece cuando la muestra tiende a infinito, si la necesidad de asumirlo *a priori*.

6. **Un estimador no puede ser consistente y asintóticamente normal al mismo tiempo.**

   Aunque la consistencia y la normalidad asintótica son dos propiedades asintóticas, es decir de muestras grandes (infinitas), estas recaen en dos conceptos matemáticos distintos.

   La consistencia del estimador, nos dice formalmente que $\hat{\beta_n}$ es un estimador consistente de β, sii. $\hat{\beta_n} \xrightarrow{\text{p}} \beta_0$. Este concepto recae en el concepto de convergencia en probabilidad, la cual implica que la sucesión de variables aleatorias, $\{\beta_n\}$ implosiona sobre un punto, el verdadero valor del parámetro, cuando el tamaño de la muestra, n, tiende a infinito.

   En cambio el concepto de normalidad asintótica, nos dice que bajos ciertos supuestos, la distribución del estimador tiende a normal, cuando n tiende a infinito. En particular:

$$\hat{\beta}|X \sim N(\beta, \sigma^2(X'X)^{-1})$$

   Ambos resultados se dan cuando n tiende a infinito, en este sentido suceden al mismo tiempo. Pero estos recaen en conceptos matemáticos diferentes por lo que formalmente no ocurren en simultaneo.

7. **Consistencia e insesgadez son dos propiedades muy similares.**

   La insesgadez es una propiedad de muestras finitas, mientras que la de consistencia es una propiedad asintotica, de muestras grandes (infinitas). La insesgadez es un concepto mas restrictivo, al definirse como una esperanza el cual es un operador lineal, solo puede demostrarse para estimadores lineales. En cambio, la consistencia se basa en el concepto de convergencia en probabilidad el cual puede aplicarse a través de cualquier función continua (lineal o no lineal), por lo que se puede aplicar y demostrar sobre un grupo mas grande de estimadores.

2. Primeros pasos en Stata

Comenzaremos definiendo el *enviroment* y cargando los datos a Stata.

2.1. Limpieza de datos

Comenzamos con la limpieza de los datos, primero corregimos errores de entrada comunes a diferentes variables, como por ejemplo: "one" por "1" o ".a" por ".". Esto lo realizamos mediante un bucle por facilidad. Luego realizamos correcciones mas especificas: Primero, generamos las variables indicadoras para las variables binarias sexo, obesidad y zona geográfica.

Por ultimo corregimos el problema de las variables numericas: *hipsiz*, *totexpr* y *tincm_r*. Todas estas variables tienen el problema de estar definidas como texto, en particular *hipsiz* y *totexpr* tienen un problema de entrada donde antes del numero correspondiente aparece la descripción de la variable. Además, el delimitador de los decimales se encuentra con "," en vez de ".". Por lo que primero extraemos los números del texto y cambiamos el delimitador de decimales para luego definir las variables como numéricas.[1]

2.2. Missing Values

Observamos que 5 variables tienen mas de 5 % de *missing values* sobre total de observaciones de la muestra. Estas son la edad en mesese (*monage*) y la variable indicadora de obesidad (*obese*) con 203 *missing values*, lo que representa un 7,2 % de la muestra. Luego tenemos al gasto real de los hogares (*totexpr*) y al ingreso real de los hogares (*tincm_r*) con 187, lo que equivale a 6,6 %. Y por ultimo, tenemos a la altura auto reportada (*htself*) con 185, un 6,56 % de la muestra.

[1] Alguno comentarios adicionales: Probablemente hubiera sido mucho mejor cambiar el nombre de la variable "sex" a "male", ya que le daría una interpretabilidad inmediata a la variable sin necesidad de definir etiquetas, no lo hicimos por un tema de trazabilidad con las consignas posteriores. Otra es pensar si tiene sentido redefinir las observaciones del ingreso real del hogar que se encontraban como " , " por missing values, o podrían ser estas 0 en su lugar. Se puede pensar que no tiene sentido que un hogar tenga ingreso 0 por lo que podría ser un problema de *missreporting*, por lo que fijar esto a cero nos generaría un problema de selectividad

2.3. Datos irregulares

Estamos buscando datos irregulares par algunas de las variables. Empezamos haciendo un *summarize* para ver si los valores minimos de la distribución de las variables tiene sentido, y encontramos que tanto la variable *totexpr* (El gasto real de los hogares), como *tincm_r* (ingreso real de los hogares) presentan valores menores a 0 lo cual no es feasibel. De modo que decidimos redefinir todo los datos irregulares como *missing values*.

Al hacer esto estamos incluyendo un numero mayor de *missing values* a los expuesto en el punto anterior. A partir de esta modificación, la variable *totexpr* pasa a presentar un 8.84 % de valores perdidos, y *tincm_r* un

2.4. Ordenando los Datos

Para conseguir un mejor orden de los datos que facilite su lectura, los ordenaremos de modo que la primera variable que aparezca en la base debería ser el *id* del individuo, la segunda el sitio (*site*) donde se encuentra y la tercera el sexo (*sex*). Para luego, ordenar las filas de mayor a menor según *totexpr*.

2.5. Descripción de los Datos

Para empezar a explorar los datos, realizamos un análisis descriptivo de las principales variables de interés: sexo, edad en años, satisfacción con la vida, circunferencia de la cintura, circunferencia de la cadera y el gasto real, los resultados se presentan en la siguiente tabla.

	Media	Desv. Estándar	Mínimo	Máximo
Hombres	0.4180	0.4933	0	1
Edad en Años	45.8492	17.7995	18	100
Satisfacción con la vida	2.4494	1.1139	1	5
Tamaño de caderas	101.4756	11.5034	40	180
Gasto Real del Hogar	7656.016	8914.258	147.83	128868.8
Observaciones	2361	2361	2361	2361

El Gasto real esta en ,,,,,,

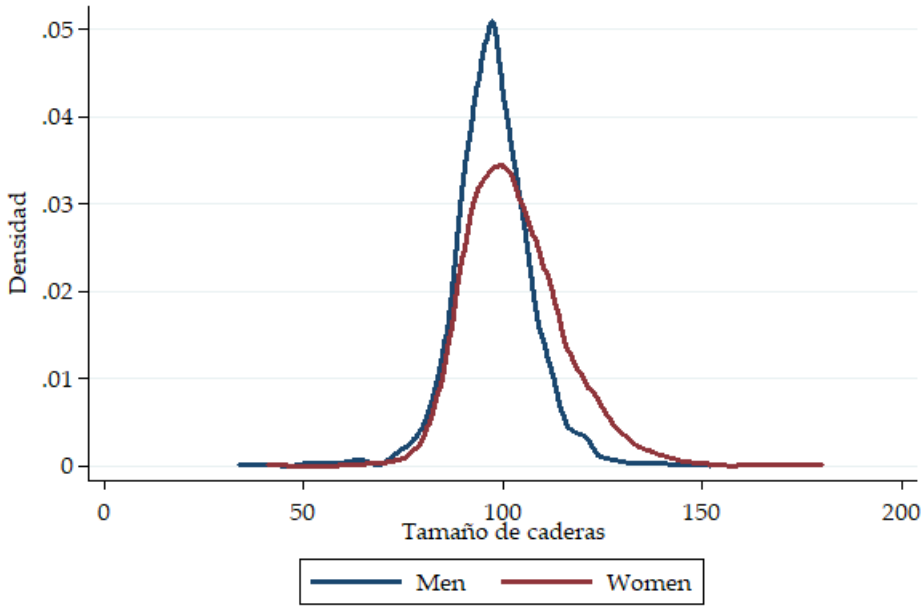
Fuente: Elaboración propia.

El 42 % de las observaciones en nuestros datos corresponden a hombres. La edad promedio de de tanto hombres como mujeres es de casi 46 años, sin embargo hay mucha dispersión dado que el desvío estándar es 17 años. La edad del individuo más grande es de 100 años y se cuenta con todas personas mayores de edad, siendo 18 año la edad más baja que se puede tener. En una escala del 1 al 5, en promedio los individuos se sienten 2.5 puntos de satisfechos con la vida. La media de las caderas de nuestra muestra está en los 101 centímetros pero oscila en promedio entre los 11 centímetros. Si bien no está claro la moneda en la cual se expresa el gasto en hogares, podemos decir algunas cosas al respecto viendo su valor relativo. Hay una gran dispersión en el gasto, su desvío estándar es mayor a su media y, valor máximo y mínimo están muy lejanos al promedio.

2.6. "Hips don't lie" (Shakira, 2005)

Queremos corroborar en nuestros datos, si las caderas de los hombres son mayores que la de las mujeres como sugiere Shakira (2005). Para esto observaremos sus distribuciones, y luego realizaremos un test de medias para comprobarlo.

Figura 1: Distribución del tamaño de las caderas



Fuente: Elaboración propia.

La distribución de los hombre presentan una menor dispersión, la distribución de las mujeres presentan un sesgo hacia la derecha, lo que a primera vista podría pareceiera rechazar la hipótesis de Shakira (2005).

Grupos	Obs	Media	Err. Std.	Dev. Std.	Intervalo de Conf. 95 %
Mujeres	1,637	102.99	0.3139	12.70	102.38 – 103.61
Hombres	1,160	97.68	0.2949	10.04	97.10 – 98.26
Total	2,797	100.79	0.2262	11.96	100.35 – 101.23
Diferencia	–	5.31	0.4480	–	4.43 – 6.19

T-tests

diff = media(Mujeres) - media(Hombres)		gr. de libertad = 2795
Ha: diff < 0	Ha: diff $\neq 0$	Ha: diff > 0
$\Pr(T < t) = 1{,}00$	$\Pr(\lvert T\rvert > \lvert t\rvert) = 0{,}00$	$\Pr(T > t) = 0{,}00$

Fuente: Elaboración propia.

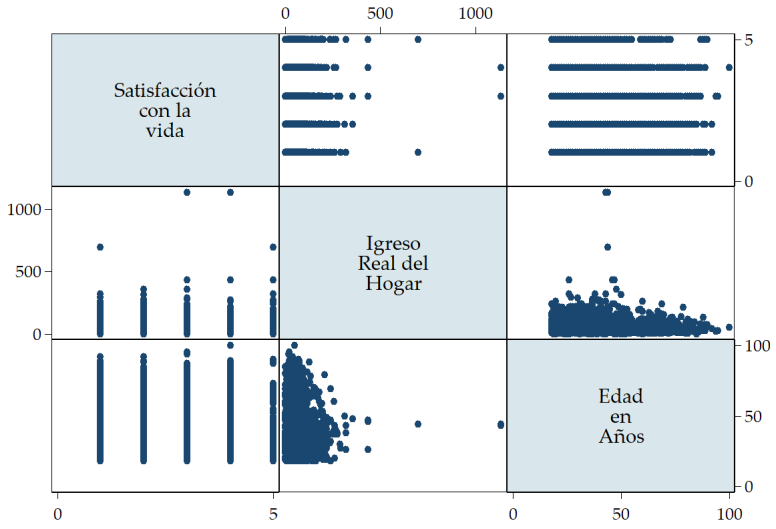
Realizamos los test de diferencia en media se rechazan las hipótesis nulas de que los hombres tienen caderas mas grandes, como la hipótesis nula de que no existen diferencias. Y No se puede rechazar la hipótesis nula

de que las caderas de las mujeres son mas chicas.[2] De este modo rechazamos la hipótesis de Shakira (2005), de que los hombres tienen caderas mas grandes que lass mujeres.

2.7. ¿Cual que se relación detrás la felicidad?

Queremos ver que relación hay detrás de la felicidad, para esto traficamos la dispersión de las observaciones entre variables para darnos una primera idea de cuales pueden ser las variables que influyan en esta.

Figura 2: Matriz de Dispersiones



Fuente: Elaboración propia.

En el presente gráfico podemos ver la correlación entre variables. En la primera fila y/o columna podemos ver el impacto de ingreso y edad contra la satisfacción con la vida, en orden respectivo. A simple vista cuesta ver una tendencia marcada. Al ser la satisfacción una variable categórica, puede haber observaciones donde haya ingreso o edades altas en cada categoría.

Nos disponemos a estimar dos modelos. El primer modelo explica la satisfacción con la vida mediante el ingreso (*tincm_100*) y el gasto (*totexp_100*) cómo variables de bienestar económico, estas fueron rescaladas a 100 unidades monetarias para facilidad de su interpretación. Además incluye el estado civil y emocional de las personas con las variables binarias *marsta*, según este casado, convivan, sean divorciados o viudos, respectivamente. Otras variables que pueden influir son la cobertura médica (*cmedin*) y su evaluación de salud (*evalhl*) como factores que puedan afectar su sensación de certidumbre. Adicionalmente controlamos por la zona en donde vive el individuo, dado que la zona geográfica puede contribuir. Por último incluimos la edad (*yearagel*) en años para detectar si existe un efecto de la vejez en la percepción de la satisfacción.

$$satlif = \beta_0 + \beta_1 tincm\_100 + \beta_2 totexp\_100 + \beta_3 satecc + \beta_4 yearage + \beta_5 cmedin + \beta_6 evalhl + \tag{1}$$
$$\beta_7 marsta1 + \beta_8 marsta_2 + \beta_9 marsta_3 + \beta_{10}\beta_{11} marsta_4 + \beta_{12} Zona_2 + \beta_{13} Zona3 + u$$

Este modelo no nos parece lo suficientemente convincente, por lo que proponemos un modelo alternativo:

$$satlif = \beta_0 + \beta_1 tincm\_100 + \beta_2 totexp\_100 + \beta_3 satecc + \beta_4 yearage + \beta_5 sq\_yearag +$$

[2] Realizamos el test tanto bajo el supuesto de idénticas varianzas, como permitiendo que sean diferentes, en ambos casos se rechazan mismas hipótesis nulas.

$$\beta_6 cmedin + \beta_7 evalhl + \beta_8 evalhl + \beta_9 smokes \beta_{10} alclmo + \beta_{11} marsta1 + \beta_{12} marsta_2 + \beta_{13} marsta_3 \quad (2)$$
$$\beta_{13} marsta_3 + \beta_{14} marsta_4 + \beta_{16} Zona_3 2 + \beta_{17} Zona3 + u$$

En este segundo modelo, incluimos variables asociadas a vicios como si fuma (*smokes*) y si toma alcohol (*alclmo*). Además incorporamos un termino cuadrático para la edad.

Las estimaciones por MCO del modelo descripto en las ecuaciones (1) y (2) se presenta en la Tabla 3.

Tabla 3. Modelos de Regresión por MCO

	Modelo 1	**Modelo 2**
Variables	*Satisfacción con la Vida*	
Ingreso Real del hogar (En cientos)	0.00094	0.0001
	(.0004)***	(.0004)***
Gasto Total del hogar(En cientos)	0.0010	0.0010
	(.0003)***	(.0003)***
Satisfacción Económica	0.5062	0.5002
	(.0231)***	(.0232)***
Edad (en años)	-0.0009	-0.01974
	(.0015)	(.0068)***
$Edad^2$ (en años)	-	0.0002
	-	(0.0001)***
Cobertura Médica	0.2218	0.2366
	(.0610)***	(.0609)***
Autoeval. de Salud	0.22904	0.2276
	(.0299)***	(.0301)***
Fumador	-	0.7523
	-	(.0442)*
Toma Alcohol	-	0.04461
	-	(.0412)
Casado	-0.0825	-0.0297
	(.0618)	(.0661)
Conviviendo	-0.0921	-0.0473
	(.0876)	(.0906)
Divorciado	-0.1456	-0.0821
	(.0861)*	(.0891)
Viudo	-0.1868	-0.1688
	(.0869)**	(.0870)*
Zona 2	-0.0111	-0.0085
	(.0689)	(.0693)
Zona 3	0.0344	0.0496
	(.0560)	(.0564)
Estadísticos		
R^2	0.306	0.3095
N	2,350	2,404

Nota: Los errores estándar fueron corregidos por heterocedasticidad, en paréntesis se presentan errores estándar. Significatividad * 0.10 ** 0.05 *** 0.01

Fuente: Elaboración propia.

En la presente tabla podemos ver los coeficientes resultante de ambos modelos. En los dos casos el ingreso del

hogar es significativo. Cada cien unidades monetarias más de ingreso tiene un efecto positivo sobre el nivel de satisfacción. La interpretación es similar para el gasto. Con respecto a la satisfacción económica, un punto más influye en medio punto más de satisfacción con la vida. El efecto de un año más de edad es bajo en el primer modelo. Sin embargo, en el segundo modelo, al incluir el efecto cuadrático en la edad, los coeficientes se tornan significativos, encontrando así un efecto negativo pero decreciente de la edad sobre la felicidad. También encontramos que los coeficientes de cobertura médica y autoevaluación de la salud son positivos y significativos, son buenos *proxy* de como valora aspectos no económicos, como la salud. Aquellas personas que viven en las Zonas 1 y 3 parecieran ser mas felices que las de las Zona 2, aunque estas diferencias no se encuentran como significativas estadísticamente. En cuanto a la situación marital, se encuentra que las personas viudas son menos felices, lo cual es significativo al 10 %.