

# Proyecto Final Econometría I

David Silva Ábrego 552999  
Aline Rivera Mata 552977

# Base de datos “SMOKE”

## Variable dependiente

- *cigs*

## Variables explicativas

- *educ*
- *cigpric*
- *white*
- *age*
- *income*
- *restaurn*
- *lincome*
- *agesq*
- *lcigpric*

# ¿Por qué escogimos este tema?

A pesar de saber que es dañino para el cuerpo las personas siguen fumando.

Escogimos este tema principalmente porque se nos hizo interesante ver qué factores tienen un mayor impacto para que una persona fume más o fume menos.

## ¿Por qué es importante?

Desafortunadamente, la supervivencia de varios fumadores hoy en día está sujeta a que abandonen este hábito o que por lo menos lo reduzcan. Con este trabajo se pueden sentar las bases para identificar por medio de qué factores pueden lograrlo así como para identificar qué factores pueden ser un impedimento para ello.





# Literatura



- **educ:** Conforme aumentan los años de escolaridad de una persona se espera que disminuya la cantidad de cigarros consumidos por día. Frecuentemente el nivel de educación depende del nivel socioeconómico de las personas por lo que *educ*, *income* y *lincome* podrían estar correlacionadas entre sí.
- **cigpric y lcigpric:** La elasticidad precio de los cigarros oscila alrededor de 0.85.
- **white:** Se ha comprobado que las personas caucásicas tienden a fumar más que las personas afroamericanas, hispanas y asiático americanas.
- **age y agesq:** La proporción de mayor de fumadores se encuentra en los rangos de edad adulta avanzada. Sin embargo, el consumo de tabaco suele ser más intensivo entre aquellos fumadores de la edad joven adulta.
- **income y lincome:** Debido a que la nicotina es una sustancia adictiva y se crea una dependencia hacia ella, los fumadores son propensos a intentar satisfacer esta adicción independientemente de su nivel de ingresos.
- **restaurn:** Se ha comprobado que políticas antitabaquismo frecuentemente son efectivas para reducir el consumo de cigarros.



# Descripción de Datos



# Frecuencias y porcentajes

Educ	Freq.	Percent
6	43	5.33
8	54	6.69
10	126	15.61
12	259	32.09
13.5	128	15.86
15	32	3.97
16	86	10.66
18	79	9.79
<b>Total</b>	<b>807</b>	<b>100</b>

Age	Freq.	Percent
17-29	253	31.33
30-59	407	50.41
60-88	147	18.19
<b>Total</b>	<b>807</b>	<b>100</b>

Cigpric	Freq.	Percent
44-49	39	4.75
50-59	276	33.88
60-70	492	60.51
<b>Total</b>	<b>807</b>	<b>100</b>

White	Freq.	Percent
0	98	12.14
1	709	87.86
<b>Total</b>	<b>807</b>	<b>100</b>

Agesq	Freq.	Percent
289-2500	563	69.73
2601-4900	199	24.64
5041-7744	45	5.56
<b>Total</b>	<b>807</b>	<b>100</b>

Income	Freq.	Percent
500	5	0.62
1500	10	1.24
2500	10	1.24
3500	19	2.35
4500	13	1.61
5500	15	1.86
6500	32	3.97
8500	61	7.56
12500	125	15.49
20000	247	30.61
30000	270	33.46
<b>Total</b>	<b>807</b>	<b>100</b>

Restaurn	Freq.	Percent
0	608	75.34
1	199	24.66
<b>Total</b>	<b>807</b>	<b>100.</b>

Agesq	Freq.	Percent
289-2500	563	69.73
2601-4900	199	24.64
5041-7744	45	5.56
<b>Total</b>	<b>807</b>	<b>100</b>

Lcigpric	Freq.	Percent
3.70-3.99	87	10.61
4.00-4.25	720	88.53
<b>Total</b>	<b>807</b>	<b>100</b>

Lincome	Freq.	Percent
6.214.608	5	0.62
7.313.221	10	1.24
7.824.046	10	1.24
8.160.519	19	2.35
8.411.833	13	1.61
8.612.503	15	1.86
8.779.557	32	3.97
9.047.821	61	7.56
9.433.484	125	15.49
9.903.487	247	30.61
1.030.895	270	33.46
<b>Total</b>	<b>807</b>	<b>100</b>



# Estadísticos descriptivos

La **media** es el promedio de los valores de una variable.

La **desviación estándar** muestra la dispersión de los datos de acuerdo a la media.

Variable	Media	Desv. Estándar
Educ	12.47088	3.057161
Cigpric	60.30041	4.738469
White	.8785626	.3268375
Age	41.23792	17.02729
Income	19304.83	9142.958
Cigs	8.686493	13.72152
Restaurn	.2465923	.4312946
Lincome	9.687315	.7126952
Agesq	1990.135	1577.166
Lcigpric	4.096032	.0829194

# Resultados



Buscamos un modelo que explique el comportamiento del consumo de cigarros por día de la siguiente forma:

$$\text{cigs} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{cigpric} + \beta_3 \text{white} + \beta_4 \text{age} + \beta_5 \text{income} + \beta_6 \text{restaurn} + \beta_7 \text{lincome} + \beta_8 \text{agesq} + \beta_9 \text{lcigpric}$$



# Regresión Múltiple con Todas las Variables

Source	SS	df	MS
Model	8369.80824	9	929.978693
Residual	143383.875	797	179.904485
Total	151753.683	806	188.280003

Number of obs = 807  
 F(3, 803) = 5.17  
 Prob > F = 0  
 R-squared = 0.0552  
 Adj R-squared = 0.0445  
 Root MSE = 13.413

cigs	Coef.	Std. Err.	t	P>t	[95% Conf. Interval]	
educ	-0.4947807	0.1681802	-2.94	0.003	-0.8249092	-0.1646522
cigpric	2.002239	1.492834	1.34	0.18	-0.9281119	4.932589
white	-0.5310485	1.460722	-0.36	0.716	-3.398365	2.336268
age	0.7783598	0.1605556	4.85	0	0.4631979	1.093522
income	-0.0000462	0.0001335	-0.35	0.729	-0.0003082	0.0002158
restaurn	-2.644241	1.129999	-2.34	0.02	-4.862367	-0.4261154
lincome	1.404058	1.708164	0.82	0.411	-1.948974	4.75709
agesq	-0.0091504	0.0017493	-5.23	0	-0.0183007	-0.0057166
lcigpric	-115.2718	85.42447	-1.35	0.178	-282.9553	52.41172
_cons	340.7993	260.016	1.31	0.19	-169.5978	851.1965

cigs = 340.7993-0.4947807educ+2.002239cigpric-0.5310485white+0.7783598age-0.0000462income-2.644241restaurn+1.404058lincome-0.0091504agesq-115.2718lcigpric

# Modelo Restringido y Prueba F

Source	SS	df	MS
Model	7739.48459	4	1934.87115
Residual	144014.198	802	179.568826
Total	151753.683	806	188.280003

Number of obs = 807  
 F(4, 802) = 10.78  
 Prob > F = 0  
 R-squared = 0.051  
 Adj R-squared = 0.0463  
 Root MSE = 13.4

cigs	Coef.	Std. Err.	t	P>t	[95% Conf. Interval]	
educ	-0.4504	0.1614857	-2.79	0.005	-0.7673845	-0.1334156
age	0.822327	0.1541866	5.33	0.000	0.51967	1.124984
restaurn	-2.746372	1.09685	-2.5	0.012	-4.899408	-0.5933367
agesq	-0.0095886	0.0016779	-5.71	0.000	-0.0128822	-0.006295
_cons	0.1521404	3.503322	0.04	0.965	-6.724623	7.028904

$H_0 = \hat{\beta}_{cigpric} = \hat{\beta}_{white} = \hat{\beta}_{income} = \hat{\beta}_{lincome} = \hat{\beta}_{lcigpric} = 0$      $H_a =$  Algún estimador es diferente de cero.

$gl = 807 - 9 - 1 = 797$   
 $q = 5$   
 $c = 2.21$

$$R_{nr}^2 = 0.0552$$

$$R_r^2 = 0.0510$$

$$F = \frac{(R_{nr}^2 - R_r^2)/q}{(1 - R_{nr}^2)/(n - k - 1)} = \frac{(0.0552 - 0.0510)/5}{(1 - 0.0552)/(807 - 9 - 1)} = 0.708594$$

- $0.708594 < 2.21$  por lo tanto no se rechaza  $H_0$
- Las variables *cigpric*, *white*, *income*, *lincome* y *lcigpric* no son conjuntamente significativas.

# Modelo Seleccionado

Source	SS	df	MS
Model	7739.48459	4	1934.87115
Residual	144014.198	802	179.568826
Total	151753.683	806	188.280003

Number of obs = 807  
F(4, 802) = 10.78  
Prob > F = 0  
R-squared = 0.051  
Adj R-squared = 0.0463  
Root MSE = 13.4

cigs	Coef.	Std. Err.	t	P>t	[95% Conf. Interval]	
educ	-0.4504	0.1614857	-2.79	0.005	-0.7673845	-0.1334156
age	0.822327	0.1541866	5.33	0.000	0.51967	1.124984
restaurn	-2.746372	1.09685	-2.5	0.012	-4.899408	-0.5933367
agesq	-0.0095886	0.0016779	-5.71	0.000	-0.0128822	-0.006295
_cons	0.1521404	3.503322	0.04	0.965	-6.724623	7.028904

$$\text{cigs} = 0.1521404 - 0.4504\text{educ} + 0.822327\text{age} - 2.746327\text{restaurn} - 0.0095886\text{agesq}$$





# Multicolinealidad

# Factor de Inflación de la Varianza

Variable	VIF	1/VIF
agesq	31.43	0.031813
age	30.94	0.032323
educ	1.09	0.914095
restaurn	1	0.995528
Mean VIF	16.12	

Debido que el promedio de los factores de inflación de las varianzas de las 4 variables independientes es mayor a 10, se tiene presencia de multicolinealidad

# Correlaciones de las Variables Independientes

	educ	age	restaurn	agesq
educ	1			
age	-0.1806	1		
restaurn	0.0605	-0.0389	1	
agesq	-0.2188	0.9831	-0.0408	1

- A excepción de las variables *age* y *agesq*, las correlaciones entre las variables independientes caen dentro del marco de lo normal.
- Era esperado que *age* y *agesq* tuvieran una alta correlación muy cercana a 1 puesto que *agesq* es simplemente  $age^2$ .
- Una opción para controlar la multicolinealidad sería eliminar una de estas dos variables.



## Regresión si se elimina age :

				Number of obs = 807	
				F(3, 803) = 4.72	
				Prob > F = 0.0028	
				R-squared = 0.0173	
				Adj R-squared = 0.0137	
				Root MSE = 13.627	
Source	SS	df	MS		
Model	2631.76906	3	877.256353		
Residual	149121.914	803	185.705995		
Total	151753.683	806	188.280003		

  

cigs	Coef.	Std. Err.	t	P>t	[95% Conf. Interval]	
educ	-0.2841778	0.1611345	-1.76	0.078	-0.6004723	0.0321167
restaurn	-2.7682	1.115428	-2.48	0.013	-4.9577	-0.5787002
agesq	-0.0007907	0.000312	-2.53	0.011	-0.00140403	-0.0001782
_cons	14.48664	2.285165	6.34	0.000	10.00104	18.97225

- $R^2$  disminuye de 0.0510 a 0.0173.
- Ninguna variable se convierte en no estadísticamente significativa.

## Regresión si se elimina agesq :

				Number of obs = 807	
				F(3, 803) = 3.35	
				Prob > F = 0.0186	
				R-squared = 0.0124	
				Adj R-squared = 0.0087	
				Root MSE = 13.662	
Source	SS	df	MS		
Model	1875.3028	3	625.100932		
Residual	149878.38	803	186.648045		
Total	151753.683	806	188.280003		

  

cigs	Coef.	Std. Err.	t	P>t	[95% Conf. Interval]	
educ	-0.2393804	0.160276	-1.49	0.136	-0.5539898	0.0752289
age	-0.0439352	0.0287458	-1.53	0.127	-0.100361	0.0124905
restaurn	-2.736921	1.118261	-2.45	0.015	-4.93198	-0.5418624
_cons	14.15848	2.552075	5.55	0.000	9.148954	19.16801

- $R^2$  disminuye de 0.0510 a 0.0124.
- Dos de las variables restantes se vuelven no estadísticamente significativas (incluyendo age).

# Prueba F de *age* y *agesq*

$$H_0 = \hat{\beta}_{age} = \hat{\beta}_{agesq} = 0$$

$H_a$  = Algún estimador es diferente de cero.

$$gl = 807 - 4 - 1 = 802$$

$$R_{nr}^2 = 0.0510$$

$$q = 2$$

$$c = 3.00$$

$$R_r^2 = 0.0173$$

$$F = \frac{(R_{nr}^2 - R_r^2)/q}{(1 - R_{nr}^2)/(n - k - 1)} = \frac{(0.0510 - 0.0173)/2}{(1 - 0.0510)/(807 - 4 - 1)} = 14.2399$$

- $14.2399 > 3.00$  por lo tanto se rechaza  $H_0$
- Las variables *age* y *agesq* son conjuntamente significativas.
- **Conclusión de Multicolinealidad:** Lo más conveniente es no actuar ante la presencia de multicolinealidad en nuestro modelo.



# Heterocedasticidad



# Resultados de Prueba de Heterocedasticidad de Breusch-Pagan

## Breusch-Pagan / Cook-Weisberg test for heteroskedasticity

Ho: Constant variance  
Variables: fitted values of cigs

chi2(1) = 53.50  
Prob > chi2 = 0.0000

Debido a que Prob > chi2 es menor a un valor de significancia de 0.05, se rechaza Ho por lo que existe presencia de heterocedasticidad en el modelo.

# Corrigiendo con “robust” (Método de White)

Linear regression

Number of obs = 807  
F(4,802)= 15.78  
Prob > F= 0  
R-squared= 0.051  
Root MSE= 13.4

cigs	Coef.	Robust Std. Err.	t	P>t	[95% Conf. Interval]	
educ	-0.4504	0.1559712	-2.89	0.004	0.7565601	-0.14424
age	0.822327	0.1359125	6.05	0	0.5555408	1.089113
restaurn	-2.746372	0.9915171	-2.77	0.006	4.692647	-0.8000975
agesq	-0.0095886	0.0014409	-6.65	0	0.0124169	-0.0067603
_cons	0.1521404	3.18116	0.05	0.962	6.092242	6.396523

Ahora contamos con la estimación menos eficiente pero insesgada de Mínimos Comunes Ordinarios. Asimismo, la matriz de las varianzas y covarianzas de los estimadores está estimada correctamente.

# Comparación de los Modelos

Variable	SW	CW
educ	-0.450***	-0.450***
age	0.822***	0.822***
restaurn	-2.746**	-2.746***
agesq	-0.010***	-0.010***
_cons	0.152	0.152

legend: \*  $p < .1$ ; \*\*  $p < .05$ ; \*\*\*  $p < .01$

Al aplicar el Método de White con STATA, *restaurn* aumentó su nivel de significancia pues pasó del 5% al 1%.





# Conclusiones

# De acuerdo a nuestro modelo:

- Se espera que el incremento de un año de escolaridad disminuya en aproximadamente 0.45 unidades el número de cigarros consumidos en un día por una persona.
- Se espera que el aumento de un año de edad aumente en aproximadamente 0.34 unidades el número de cigarros consumidos en un día por una persona.
- Se espera que si hay restricciones a fumar en restaurantes, el número de cigarros consumidos en un día por una persona disminuya en promedio 2.75 unidades aproximadamente.

$$\widehat{\Delta cigs} = -0.4504 \Delta educ$$

$$\widehat{\Delta cigs} = -0.4504(1)$$

$$\widehat{\Delta cigs} = -0.4504$$

$$\widehat{\Delta cigs} = (\widehat{\beta}_{age} + 2\widehat{\beta}_{agesq} age) \Delta age$$

$$\widehat{\Delta cigs} = (0.822327 + (2)(-0.0095886)(25))(1)$$

$$\widehat{\Delta cigs} = 0.342897$$

$$\widehat{\Delta cigs} = -2.746372 \Delta restaurn$$

$$\widehat{\Delta cigs} = -2.746372(1)$$

$$\widehat{\Delta cigs} = -2.746372$$

# Conclusiones generales

- A pesar que se cuenta con 9 variables explicativas solamente 4 son significativas para el modelo.
- A pesar de eliminar 5 variables del modelo original la R-cuadrada en el modelo restringido varió muy poco.
- El modelo tiene multicolinealidad debido a la alta correlación entre *age* y *agesq*, sin embargo el modelo es más certero con dichas variables.
- Al arreglar el problema de la heterocedasticidad la variable *restaurn* aumentó su significancia de al menos 5% a al menos 1%.
- No se sospecha que haya endogeneidad en el modelo.



# Referencias



- Azevedo, G., Gonçalves, J., de Almeida, L., de Moura, E. & Carvalho, D. (2009). Tobacco smoking and level of education in Brazil, 2006. *Rev Saúde Pública*. 43 (Supl. 2), 1-9.
- Baker, T., Brandon, T. & Chassin, L. (2004). Motivational influences on cigarette smoking. *Rev psychol*. 55, 463-491.
- Blanco, M., Cifuentes, T., Rodríguez, C. & Suárez, C. (2009). *Factores que Influyen en el Consumo de Tabaco*. Recuperado de: <https://intellectum.unisabana.edu.co/bitstream/handle/10818/2600/121974.pdf?sequence=1>
- Chávez, R. (2016). Elasticidad precio de la demanda de cigarrillos y alcohol en Ecuador con datos de hogares. *Revista Panamericana de Salud Pública*. 40, n. 4, 222-228.
- Martí, P. (2016). *Cuando el tabaco era bueno para la salud*. Recuperado de: <https://www.lavanguardia.com/vida/20160210/302057569916/tabaco-bueno-salud.html>
- Pampel, F., Mollborn, S., y Lawrence, E. (2015). *Life Course Transitions in Early Adulthood and SES Disparities in Tobacco Use*. Recuperado de: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3840392/#>

# Código de Honor

Damos nuestra palabra que hemos  
realizado esta actividad con  
Integridad Académica.

