

## Computación en Estadística y Optimización

MESIO curso 2018-19

Práctica SAS

Enero de 2019

### Presentación

El presente documento es un breve análisis econométrico de la estructura salarial en España. Los datos utilizados son de elaboración propia, y tienen como único insumo la base de datos entregados por la "Encuesta de estructura salarial año 2006" (EES2006) elaborada por el INE.

El fichero Excel utilizado contiene 21880 observaciones y las variables utilizadas son Salario, logaritmo del salario, antigüedad, antigüedad al cuadrado, sexo, edad, número de horas trabajadas en el mes de octubre de 2006, nacionalidad, control de la empresa, tipo de jornada laboral, tipo de duración del contrato y responsabilidad.

Las observaciones tomadas corresponden a las personas cuyo nivel de educación es "licenciados, ingenieros superiores y doctores". Estos datos se utilizan en los cursos de econometría para analizar cuáles son los principales factores que determinan las diferencias salariales de los trabajadores en España.

En la siguiente web esta la información de esta encuesta [https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica\\_C&cid=1254736177025&menu=resultados&secc=1254736195110&idp=1254735976596](https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177025&menu=resultados&secc=1254736195110&idp=1254735976596)

Logaritmo del salario, y antigüedad al cuadrado son variables creadas a partir de salario y de antigüedad respectivamente, y que se utilizan siguiendo la teoría de las ecuaciones de Mincer para analizar la estructura de los salarios.

## Descripción de los datos

Lista alfabética de variables y atributos					
#	Variable	Tipo	Longitud	Formato	Etiqueta
4	anti	Num	8	BEST.	antigüedad en la empresa
5	anti2	Num	8	BEST.	antigüedad al cuadrado
10	control	Num	8	CONTROL.	empresa estatal o privada
7	edad	Num	8	EDAD.	edad
1	id	Num	8	BEST.	id
3	lsal	Num	8	BEST.	logaritmo del salario
9	nac	Num	8	NAC.	nacionalidad
8	nht	Num	8	BEST.	horas trabajadas
13	respon	Num	8	RESPON.	responsabilidad
2	sal	Num	8	BEST.	salario
6	sexo	Num	8	SEXO.	sexo
12	tipoc	Num	8	TIPOC.	tipo contrato
11	tipoj	Num	8	TIPOJ.	tipo jornada

La variable sal es el salario bruto mensual, definido como la suma de salario base, pagos extras, pagos por horas extras y complementos salariales (salario base, pagos extras y pagos por hora extras y complementos salariales son variables de la EES2006).

La variable lsal es el logaritmo natural de sal. anti es la antigüedad en la empresa medido en años, y es de la EES2006. anti2 es el cuadrado de anti.

La variable sexo es de EES2006 y sus valores son 1=hombre 6=mujer. También de EES2006 es edad cuyos valores son 1=menores de 19, 2=de 20 a 29, 3=de 30 a 39, 4=de 40 a 49, 5=de 50 A 59, 6=más de 59.

nht es el número de horas trabajadas en el mes de octubre de 2006. Esta variable se crea desde las variables de EES2006 horas semanales pactadas y horas extraordinarias en el mes. nac es nacionalidad es de EES2006 y sus valores son 1=españa, 2= resto del mundo.

control es una variable que nos indica la dependencia de la empresa donde se trabaja, es de la EES2006 y sus valores son 1= público, 2= privado. También de EES2006 es tipoj que es el tipo de jornada laboral y sus valores son 1=tiempo completo 6=tiempo parcial

La variable tipoc es la duración del contrato, sus valores son 1=indefinida 2=determinada y respon nos dice si el trabajador tiene o no responsabilidad en la organización y/o supervisión, sus valores son 1= si 6= no, ambas variables de EES2006.

Finalmente la variable id es solo un identificador y asigna el número i a la observación i.

Variable	Etiqueta	Mínimo	Media	Cuantil superior	Máximo
id	id	1.0000000	10940.50	16410.50	21880.00
sal	salario	0	1272.07	1631.91	20180.65
lsal	logaritmo del salario	-4.6051702	6.8643022	7.3981741	9.9124795
anti	antigüedad en la empresa	0.0833333	8.7970902	13.5000000	55.0000000
anti2	antigüedad al cuadrado	0.0069444	171.6925458	182.2500000	3025.00
sexo	sexo	1.0000000	3.7170932	6.0000000	6.0000000
edad	edad	1.0000000	3.3175503	4.0000000	6.0000000
nht	horas trabajadas	5.3142857	157.7212826	177.1428571	265.1428571
nac	nacionalidad	1.0000000	1.0251828	1.0000000	2.0000000
control	empresa estatal o privada	1.0000000	1.8059415	2.0000000	2.0000000
tipoj	tipo jornada	1.0000000	1.7127514	1.0000000	6.0000000
tipoc	tipo contrato	1.0000000	1.2554845	2.0000000	2.0000000
respon	responsabilidad	1.0000000	4.4702925	6.0000000	6.0000000

En el cuadro anterior vemos que no existen errores en los datos. El salario presenta valores extremos superiores, lo cual es normal considerando que las observaciones provienen de los trabajadores con mayor nivel de educación. Por su parte los datos en cuanto a antigüedad y horas trabajadas están en márgenes aceptables.

### Análisis de datos

A continuación presentamos un completo análisis descriptivo de las variables numéricas correspondientes al salario, la antigüedad y el número de horas trabajadas.

#### Variable Sal (salario)

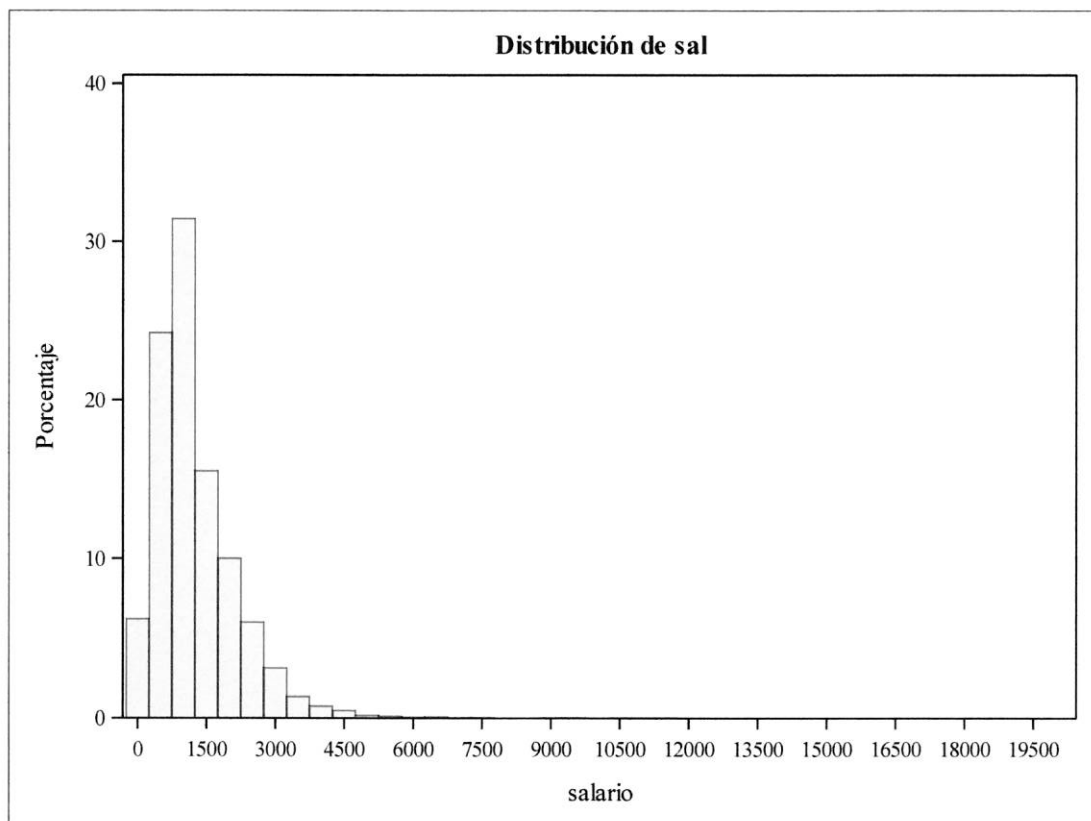
Momentos			
<b>N</b>	21880	<b>Sumar pesos</b>	21880
<b>Media</b>	1272.0703	<b>Observ suma</b>	27832898.2
<b>Desviación std</b>	1060.43785	<b>Varianza</b>	1124528.44
<b>Asimetría</b>	4.82436873	<b>Curtosis</b>	54.1087842
<b>SC no corregida</b>	6.0009E10	<b>SC corregida</b>	2.46036E10
<b>Coef. variación</b>	83.3631484	<b>Media error std</b>	7.16904924

Medidas estadísticas básicas			
Ubicación		Variabilidad	
Media	1272.070	Desviación std	1060
Mediana	1029.600	Varianza	1124528
Moda	453.330	Rango	20181
		Rango intercuartil	975.43000

Tests para posición: $\mu_0=0$				
Test	Estadístico		p valor	
T de Student	t	177.4392	Pr >  t	<.0001
Signo	M	10929.5	Pr >=  M	<.0001
Rango con signo	S	1.1946E8	Pr >=  S	<.0001

Cuantiles (Definición 5)	
Nivel	Cuantil
100% Máx	20180.650
99%	4500.000
95%	2939.120
90%	2427.280
75% Q3	1631.910
50% Mediana	1029.600
25% Q1	656.480
10%	345.020
5%	212.115
1%	66.300
0% Mín	0.000

Observaciones extremas			
Inferior		Superior	
Valor	Observación	Valor	Observación
0	21548	19194.1	21020
0	18151	19461.6	21016
0	17163	19836.3	21880
0	14893	20093.0	21879
0	13886	20180.7	21876



El histograma nos presenta de forma gráfica lo que antes vimos en los estadísticos, que el salario presenta una cola pesada a la derecha (cola larga a la derecha).

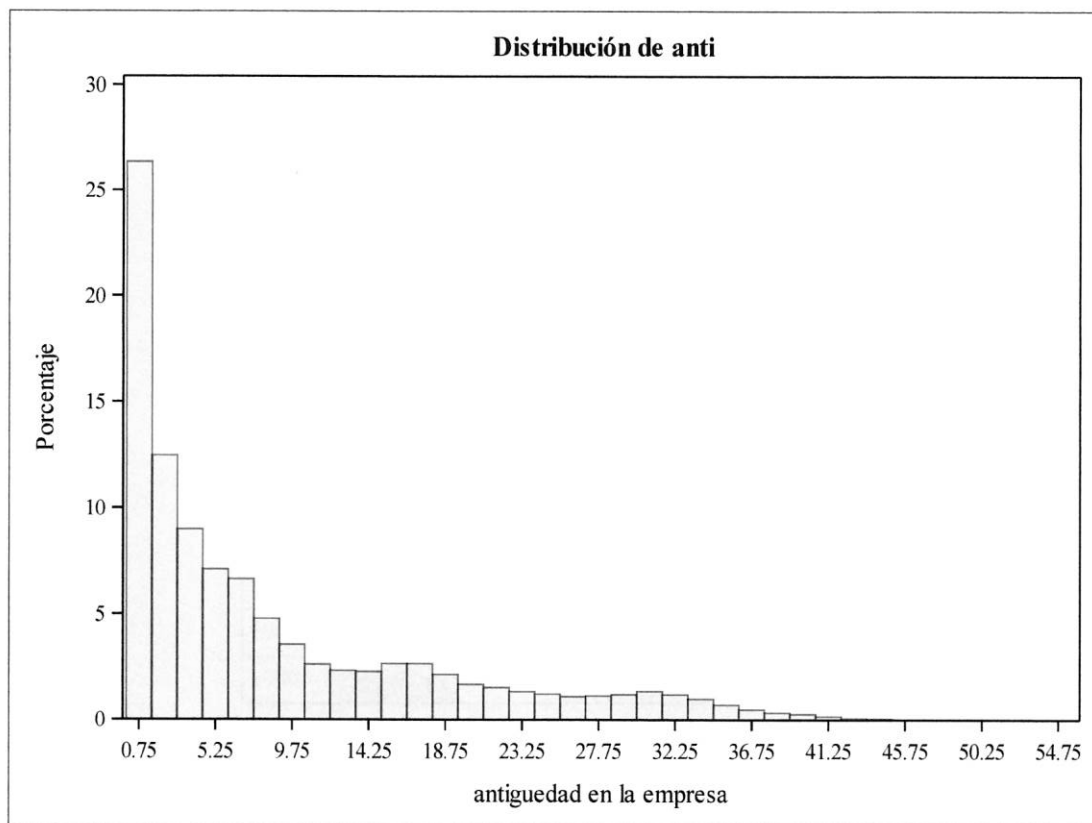
Variable anti (antigüedad en la empresa)

Momentos			
<b>N</b>	21880	<b>Sumar pesos</b>	21880
<b>Media</b>	8.79709019	<b>Observ suma</b>	192480.333
<b>Desviación std</b>	9.71123372	<b>Varianza</b>	94.3080603
<b>Asimetría</b>	1.35607175	<b>Curtosis</b>	0.97940651
<b>SC no corregida</b>	3756632.9	<b>SC corregida</b>	2063366.05
<b>Coef. variación</b>	110.391431	<b>Media error std</b>	0.06565242

Medidas estadísticas básicas			
Ubicación		Variabilidad	
<b>Media</b>	8.797090	<b>Desviación std</b>	9.71123
<b>Mediana</b>	5.000000	<b>Varianza</b>	94.30806
<b>Moda</b>	0.833333	<b>Rango</b>	54.91667
		<b>Rango intercuartil</b>	12.16667

Cuantiles (Definición 5)	
Nivel	Cuantil
<b>100% Máx</b>	55.0000000
<b>99%</b>	37.3333333
<b>95%</b>	30.8333333
<b>90%</b>	24.6666667
<b>75% Q3</b>	13.5000000
<b>50% Mediana</b>	5.0000000
<b>25% Q1</b>	1.3333333
<b>10%</b>	0.5000000
<b>5%</b>	0.3333333
<b>1%</b>	0.0833333
<b>0% Mín</b>	0.0833333

Observaciones extremas			
Inferior		Superior	
Valor	Observación	Valor	Observación
0.0833333	21207	47.0000	13975
0.0833333	21166	49.1667	6504
0.0833333	21137	50.0000	246
0.0833333	21111	50.1667	2650
0.0833333	21067	55.0000	17620



Variable nht (número de horas trabajadas)

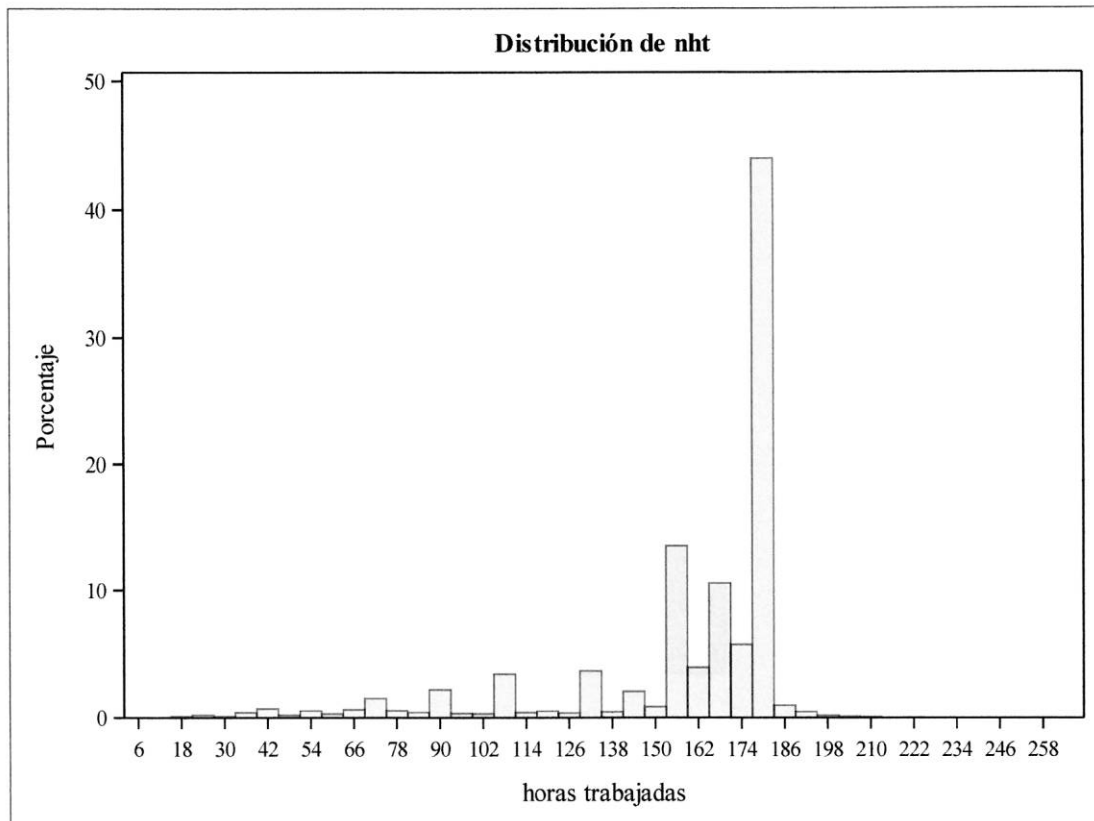
Momentos			
N	21880	Sumar pesos	21880
Media	157.721283	Observ suma	3450941.66
Desviación std	32.4079015	Varianza	1050.27208
Asimetría	-2.0115375	Curtosis	3.73692002
SC no corregida	567265848	SC corregida	22978902.8
Coef. variación	20.5475767	Media error std	0.21909237

Medidas estadísticas básicas			
Ubicación		Variabilidad	
Media	157.7213	Desviación std	32.40790
Mediana	172.7143	Varianza	1050
Moda	177.1429	Rango	259.82857
		Rango intercuartil	22.14286

Tests para posición: $\mu_0=0$				
Test	Estadístico		p valor	
T de Student	t	719.8849	Pr >  t	<.0001
Signo	M	10940	Pr >=  M	<.0001
Rango con signo	S	1.1969E8	Pr >=  S	<.0001

Cuantiles (Definición 5)	
Nivel	Cuantil
100% Máx	265.14286
99%	189.14286
95%	177.14286
90%	177.14286
75% Q3	177.14286
50% Mediana	172.71429
25% Q1	155.00000
10%	110.71429
5%	77.50000
1%	42.07143
0% Mín	5.31429

Observaciones extremas			
Inferior		Superior	
Valor	Observación	Valor	Observación
5.31429	12989	251.238	5921
5.90476	9932	255.143	17341
8.48810	10619	257.143	5213
15.50000	3109	259.198	18906
15.50000	3079	265.143	20390



Como señalamos, en este estudio se pretende abordar qué variables influyen y en qué proporción en el salario de un trabajador que cuenta con los niveles más altos de estudios. La siguiente tabla muestra el promedio del salario y el promedio de horas trabajadas por sexo y por tramos de edad de muestra.

Por ahora, es apresurado sacar conclusiones, más adelante haremos una regresión para ver los coeficientes de cada variable y el impacto en el salario. Podemos si señalar, y como era esperable, que a medida que aumenta la edad, aumenta el promedio de los salarios. La tabla nos indica que el promedio de horas trabaja por las mujeres es inferior al promedio de horas trabajadas por los hombres, y que el promedio en el salario de las mujeres es superior al promedio de los salarios de las mujeres, pero no podemos señalar aún que esa diferencia sea significativa o que se deba a que los hombres en promedio trabajan más horas.

			salario	horas trabajadas
		N	Mean	Mean
edad	sexo			
menor de 19 años	hombre	12	654.59	123.36
	mujer	13	542.93	105.50
entre 20 a 29 años	hombre	1899	975.80	163.87
	mujer	3410	915.30	151.34
entre 30 y 39 años	hombre	3776	1275.91	167.69
	mujer	4555	1102.83	154.89
entre 40 y 49 años	hombre	2350	1626.79	165.67
	mujer	2399	1410.54	150.50
entre 50 y 59 años	hombre	1552	1817.42	164.54
	mujer	1288	1543.77	147.64
más de 59 años	hombre	401	1728.42	144.04
	mujer	225	1304.91	126.51

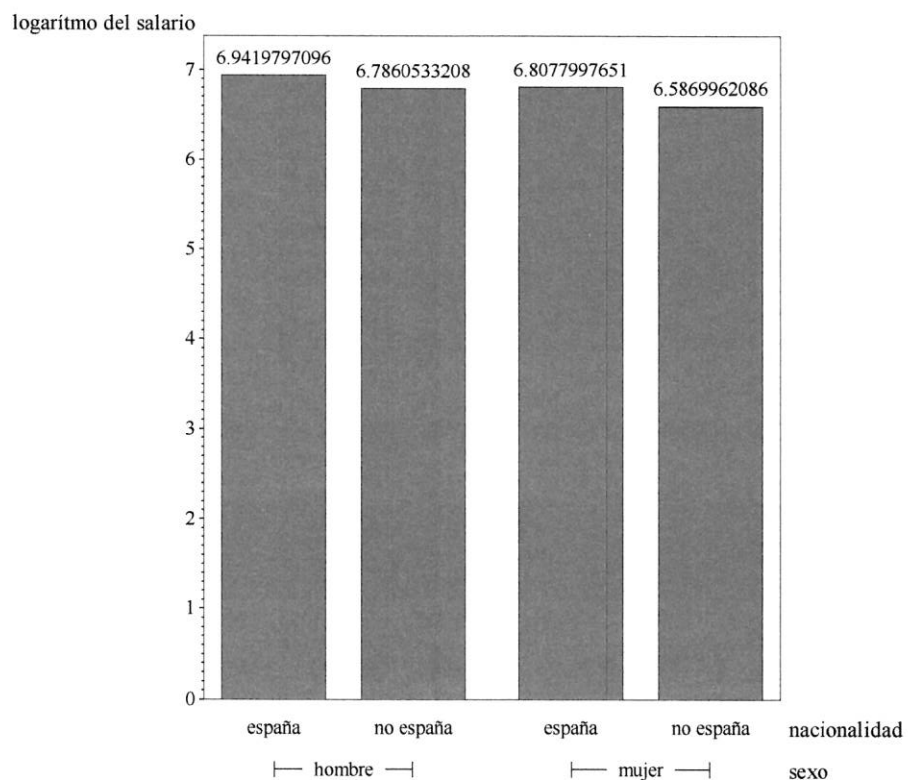
Coeficientes de la correlación de Pearson entre las variables numéricas sal, anti y nht.

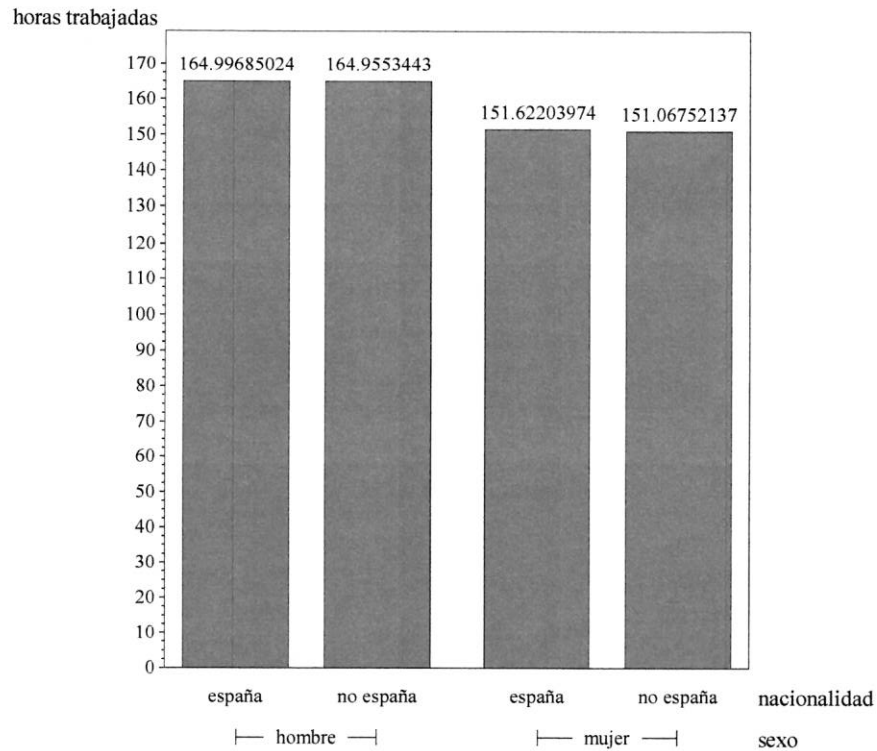
Estadísticos simples							
Variable	N	Media	Dev std	Suma	Mínimo	Máximo	Etiqueta
sal	21880	1272	1060	27832898	0	20181	salario
anti	21880	8.79709	9.71123	192480	0.08333	55.00000	antigüedad en la empresa
nht	21880	157.72128	32.40790	3450942	5.31429	265.14286	horas trabajadas

Coeficientes de correlación Pearson, N = 21880 Prob >  r  suponiendo H0: Rho=0				
		sal	anti	nht
sal salario		1.00000	0.25703 <.0001	0.16634 <.0001
anti antigüedad en la empresa		0.25703 <.0001	1.00000	0.01362 0.0440
nht horas trabajadas		0.16634 <.0001	0.01362 0.0440	1.00000

El siguiente gráfico nos muestra algo interesante, y es que en este tramo de la población trabajadora de España (los con más alto nivel de estudio) la variable nacionalidad no tiene o tiene muy poca incidencia en el salario. No se parecían diferencias significativas entre los trabajadores con nacionalidad española y aquellos de otras nacionalidades. Hemos utilizado el logaritmo del salario pues así se plantean estos problemas en las ecuaciones de Mincer, y es porque de esta forma controlamos la volatilidad de los salarios.

Presentamos también el grafico en relación al promedio de las horas trabajadas.





Ahora planteamos el siguiente modelo de regresión lineal:

$$Lsal = b_0 + b_1 \cdot \text{sexo} + b_2 \cdot \text{edad} + b_3 \cdot \text{anti} + b_4 \cdot \text{anti}^2 + b_5 \cdot \text{nht} + b_6 \cdot \text{nac} + b_7 \cdot \text{control} + b_8 \cdot \text{tipoj} + b_9 \cdot \text{tipoc} + b_{10} \cdot \text{respon} + \text{error}.$$

Y los resultados son:

Número de observaciones leídas	21880
Número de observaciones usadas	21859
Número de observaciones con valores ausentes	21

Análisis de la varianza					
Fuente	DF	Suma de cuadrados	Cuadrado de la media	F-Valor	Pr > F
Modelo	10	2358.76012	235.87601	397.02	<.0001
Error	21848	12980	0.59412		
Total corregido	21858	15339			

Raíz MSE	0.77079	R-cuadrado	0.1538
Media dependiente	6.86430	R-Sq Ajust	0.1534
Coef Var	11.22898		

Estimadores de parámetros						
Variable	Etiqueta	DF	Estimador del parámetro	Error estándar	Valor t	Pr >  t
Intercept	Intercept	1	7.14325	0.07677	93.05	<.0001
sexo	sexo	1	-0.01531	0.00224	-6.84	<.0001
edad	edad	1	0.06091	0.00700	8.69	<.0001
anti	antigüedad en la empresa	1	0.02751	0.00202	13.64	<.0001
anti2	antigüedad al cuadrado	1	-0.00054174	0.00005584	-9.70	<.0001
nht	horas trabajadas	1	0.00310	0.00025948	11.95	<.0001
nac	nacionalidad	1	0.00377	0.03363	0.11	0.9107
control	empresa estatal o privada	1	-0.46614	0.01454	-32.06	<.0001
tipoj	tipo jornada	1	-0.02367	0.00479	-4.94	<.0001
tipoc	tipo contrato	1	-0.06478	0.01424	-4.55	<.0001
respon	responsabilidad	1	-0.02289	0.00242	-9.45	<.0001

## Conclusiones

Lo primero que notamos es que el ajuste del modelo es bajo, del orden del 15%. Esto nos señala que las variables explicativas y la forma funcional del modelo solo explican un 15% del comportamiento de la variable dependiente.

Debemos señalar que este es un modelo muy sencillo, que no ha considerado variables relacionadas a la habilidad o expertiz de los trabajadores, que no ha contemplado la posibilidad de heterocedasticidad en la varianza del error, ni ha utilizado otras herramientas de la econometría, pues escapan a los objetivos del presente trabajo.

Junto con las deficiencias del modelo, es probable que las personas con mayores niveles de educación tengan más oportunidades de empleo, y una variedad amplia de intereses, lo que genera una mayor variabilidad de los salarios, y por lo mismo, sea difícil contar con rangos de ajuste aceptables.

Dejando en claro las debilidades del modelo, atendiendo a los resultados de la regresión podemos señalar entre los aspectos más relevantes que:

- Un hombre en iguales condiciones que una mujer gana un 1.5% más (las cifras oficiales presentan una brecha bastante mayor, que si bien es menor en el sector con mayores niveles de estudio es mucho mayor a lo arrojado en este modelo).
- La edad explica en un 6% el salario, y a medida que aumenta la edad el salario también lo hace.
- La antigüedad influye positivamente y explica en un 2,7% el salario.
- La variable nacionalidad no es significativa para explicar el salario de un trabajador, resultado que habíamos adelantado de gráficos anteriores.

#### **Anexo: código SAS**

```
libname pm"F:\tarea sas\pm";  
proc import out=pm.encuestaine  
datafile ="F:\tarea sas\varpm.xlsx"  
dbms=xlsx replace;  
getnames=yes;  
run;  
proc print data=pm.encuestaine (obs=10);  
run;  
proc format library = pm;  
value sexo 1="hombre" 6="mujer";  
value edad 1="menor de 19 años" 2="entre 20 a 29 años" 3="entre 30 y 39 años" 4="entre 40 y 49 años" 5="entre 50 y 59 años" 6="más de 59 años";  
value nac 1="españa" 2="no españa";  
value control 1="público" 2="privado";  
value tipoj 1="completa" 6="parcial";  
value tipoc 1="indefinido" 2="determinado";
```

```

value respon 1="si" 6="no";

run;

options fmtsearch = (pm);

data pm.encuestaine2;

set pm.encuestaine;

label nht="horas trabajadas" sal = "salario" lsal = "logaritmo del salario" anti= "antigüedad en la
empresa" anti2="antigüedad al cuadrado" nac= "nacionalidad" control="empresa estatal o
privada" tipoj="tipo jornada" tipoc="tipo contrato" respon="responsabilidad";

format sexo sexo. ;

format edad edad.;

format nac nac. ;

format control control.;

format tipoj tipoj. ;

format tipoc tipoc.;

format respon respon. ;

run;

ods rtf file='F:\tarea sas\tareacom.rtf';

proc contents data=pm.encuestaine2;

run;

proc means data=pm.encuestaine2 min mean q3 max;

run;

proc univariate data=pm.encuestaine2;

var sal anti nht;

histogram;

run;

proc tabulate data=pm.encuestaine2 ;

class edad sexo;

var sal nht;

tables edad*sexo,(n sal*mean nht*mean);

```

```

run;

proc iml;

reset print;

proc corr data=pm.encuestaine2;

var sal anti nht;

run;

quit;

PROC GCHART DATA = pm.encuestaine2;

VBAR nac /DISCRETE TYPE = mean space = 4 mean

sumvar = lsal GROUP = sexo;

RUN;

PROC GCHART DATA = pm.encuestaine2;

VBAR nac /DISCRETE TYPE = mean space = 4 mean

sumvar = nht GROUP = sexo;

RUN;

proc reg data=pm.encuestaine2;

model lsal=sexo edad anti anti2 nht nac control tipoj tipoc respon;

run;

%macro uni(bbdd,vnum1,vnum2,vnum3);

proc univariate data=&bbdd ;

var &vnum1 &vnum2 &vnum3;

histogram;

run;

%mend;

%uni(pm.encuestaine2,sal,anti,nht);

ods rtf close;

```