In [44]:
```
%pip install stargazer
```

```
Requirement already satisfied: stargazer in c:\users\usuario\anaconda3\lib\site-pa
ckages (0.0.6)
Note: you may need to restart the kernel to use updated packages.
```

In [45]:
```
import pandas as  pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np
import statsmodels.api as sm
from scipy.stats import norm
from stargazer.stargazer import Stargazer
from scipy.stats import pearsonr
```

Jubilaciones en Chile.

Los seres humanos experimentamos en el transcurso de la vida diversos cambios: en la apariencia física, en el modo de pensar, en el modo de aprender y en la forma de relacionarnos con las demás personas.

Estos cambios constantes, en su conjunto, nos permiten identificar distintas etapas en la vida. Según el Ministerio de Salud de Chile considera que son cuatro, según su edad: la infancia, la adolescencia y juventud, la etapa de la adultez y finalmente el adulto mayor. Para este último, el Servicio Nacional del Adulto Mayor establece la edad de 60 años, independientemente del género de la persona.

En cuanto a la vejez, se puede decir que, pese a ser una vivencia ineludible de la experiencia humana a la vez igual se considera como una experiencia lejana: El viejo es otro, no nosotros.

Aunque la vejez y la muerte son acontecimientos continuos en el final de nuestras vidas, debemos considerar que, la muerte nos acecha y circunda en todo momento, no solo en la vejez, sino siempre; es impredecible en nuestras vidas.

La vejez, en cambio, se sucede gradualmente con el lento paso del tiempo; no se ve. Pero, de igual modo, se nos mete en el cuerpo sin darnos cuenta. Por lo mismo no nos reconocemos en los viejos, ya que negamos el carácter invasor de la vejez y la catalogamos en una “subcultura” distinta a la nuestra (idea extraída de la tesis de Gerardo Salinas, quien leyó las obras de la filósofa Simone de Beauvour).

Desde una perspectiva económica y política, la vejez marca el inicio del deterioro en las capacidades de una persona para generar sus propios recursos, como sí lo hacía en su etapa adulta. Es en este momento cuando la persona se incorpora al mecanismo de protección social, específicamente al sistema de pensiones. Dicho sistema que establece la edad de jubilación, siendo los 60 años para las mujeres y los 65 para los hombres.

El sistema de pensiones actual en Chile presenta dos modalidades de administración. El primero conocido como, “sistema de reparto”, que va en retirada y el segundo, “sistema de capitalización individual”, creado en los años ochenta, destinado a cubrir todo el sistema de pensiones en el futuro.

En el sistema de reparto las pensiones se administran de forma estatal y su financiamiento incluía los aportes del empleador, los recursos estatales y un porcentaje del sueldo de los trabajadores. Montos que se iban a un fondo común para posteriormente ser distribuidos entre los mismos afiliados.

En cambio, el sistema de capitalización individual, administrado por privados a través de "Administradora de Fondo de Pensiones" (AFP), basan su sistema de pensiones a través del ahorro individual del trabajador. Dinero que es rentabilizado para posteriormente ser entregado al mismo trabajador como pensión de vejez.

No obstante, en la actualidad, el sistema de las AFP se ha visto duramente cuestionado. La promesa ofrecida en sus comienzos, entregar mejores pensiones que las del sistema de reparto no se cumplió lo cual ha obligado al Estado a hacerse cargo de la situación.

Las razones de este incumplimiento, como destaca Carlos Hunneus, están vinculadas más que nada a las deficiencias estructurales del mercado del trabajo, tales como la inestabilidad laboral y los bajos salarios.

Finalmente, para evidenciar la magnitud del problema, es relevante de mencionar que la Comisión Presidencial de Pensiones de 2015 concluyó que "al llegar a la edad de jubilación, más del 79% de los pensionados recibe mensualmente una pensión menor al sueldo mínimo"

Datos actualizados del Sistema de Pensiones

Las Administradoras de Pensiones, a pesar de ser entidades privadas, están sujetas a la supervisión de la Superintendencia de Pensiones en Chile. Esta institución actúa como el órgano representativo del Estado dentro del sistema de pensiones del país. Se trata de una entidad autónoma cuya máxima autoridad es el Superintendente o Superintendenta. La relación de la Superintendencia de Pensiones con el gobierno se establece a través del Ministerio del Trabajo y Previsión Social, en coordinación con la Subsecretaría de Previsión Social.

En la actualidad la entidad cuenta con un Departamento de Estudios, cuya misión es analizar y preparar información relevante sobre las distintas facetas del Sistema. Dentro de sus funciones se tiene la Ficha Estadística Regional del Sistema de Pensiones, el cual muestra el panorama detallado del sistema previsional respecto de la participación de hombres y de mujeres, a través de las regiones.

En el informe identificado con el número ocho, marzo 2023, se puede destacar lo siguiente: Las regiones que concentran el mayor número de pensionadas y pensionados, en tanto, son la Metropolitana, seguida por la de Valparaíso y la del Biobío.

En cuanto al monto promedio de la pensión por regiones, es Antofagasta la región donde las y los pensionados obtuvieron un mayor monto autofinanciado, con 399.542 pesos. Por otro lado, la región donde el monto de autofinanciamiento fue el menor a nivel país es la región del Maule, con 202.597 pesos.

Finalmente, se puede señalar que bien el monto promedio de la pensión autofinanciada en Chile a marzo pasado alcanzó a 294.226 pesos, la diferencia entre la de hombres y mujeres es significativa. Mientras los hombres recibieron, en promedio, 355.289 pesos, las mujeres obtuvieron un pago promedio de 218.591 pesos, es decir, 136.698 pesos menos.

Tabla de pensiones pagadas en vejez edad y vejez anticipada en las modalidades de Retiro Programado, Renta Temporal y Renta Vitalicia. Cantidad de hombres y mujeres por región.

Tabla N° 6

Pensiones pagadas (D.L. 3.500) de vejez [1] según región[2] y sexo

Marzo 2023

Región	Número			Participación por región	Participación por sexo	
	Hombres	Mujeres	Total		% Hombres	% Mujeres
Arica	7.492	5.712	13.204	1,2%	56,7%	43,3%
Tarapacá	9.690	7.153	16.843	1,5%	57,5%	42,5%
Antofagasta	20.295	12.053	32.348	3,0%	62,7%	37,3%
Atacama	10.755	6.005	16.760	1,5%	64,2%	35,8%
Coquimbo	25.542	18.062	43.604	4,0%	58,6%	41,4%
Valparaíso	67.499	54.720	122.219	11,2%	55,2%	44,8%
Metropolitana	256.338	233.577	489.915	44,8%	52,3%	47,7%
O'Higgins	33.168	21.945	55.113	5,0%	60,2%	39,8%
Maule	34.736	24.451	59.187	5,4%	58,7%	41,3%
Ñuble	13.014	9.998	23.012	2,1%	56,6%	43,4%
Biobío	54.566	38.490	93.056	8,5%	58,6%	41,4%
La Araucanía	25.392	20.440	45.832	4,2%	55,4%	44,6%
Los Ríos	12.574	9.000	21.574	2,0%	58,3%	41,7%
Los Lagos	22.719	18.270	40.989	3,7%	55,4%	44,6%
Aysén	2.977	2.352	5.329	0,5%	55,9%	44,1%
Magallanes	7.096	5.515	12.611	1,2%	56,3%	43,7%
Sin inf.	1.684	1.122	2.806	0,3%	60,0%	40,0%
Total	605.537	488.865	1.094.402	100,0%	55,3%	44,7%

(1) Corresponde a las pensiones pagadas en vejez edad y vejez anticipada en las modalidades de Retiro Programado, Renta Temporal y Renta Vitalicia.

(2) La región corresponde al domicilio que el afiliado pensionado tiene registrado en la AFP.

Tabla de las pensiones pagadas en las modalidades de Retiro Programado, Renta Temporal y Renta Vitalicia a pensionados por vejez.

Tabla N° 10

Monto promedio de pensiones pagadas (D.L. 3.500)[1] autofinanciadas y con Aporte Previsional Solidario (APS) o Pensión Garantizada Universal (PGU)[2] en vejez según región[3] y sexo

(En pesos)

Marzo 2023

Región	Monto promedio $					
	Hombres		Mujeres		Total	
	Autofinanciada	Final[4]	Autofinanciada	Final[4]	Autofinanciada	Final[4]
Arica	$292.208	$432.008	$187.742	$282.086	$247.016	$367.152
Tarapacá	$323.482	$459.061	$184.770	$274.233	$264.572	$380.567
Antofagasta	$507.817	$617.273	$217.226	$302.747	$399.542	$500.079
Atacama	$351.469	$481.539	$186.603	$278.319	$292.399	$408.727
Coquimbo	$330.258	$458.200	$189.723	$281.211	$272.045	$384.887
Valparaíso	$342.229	$470.494	$200.198	$284.515	$278.639	$387.227
Metropolitana	$411.690	$536.698	$243.890	$324.575	$331.688	$435.564
O'Higgins	$324.892	$451.163	$183.218	$270.174	$268.480	$379.096
Maule	$220.344	$354.869	$177.385	$271.463	$202.597	$320.413
Ñuble	$237.797	$370.872	$190.045	$284.331	$217.050	$333.273
Biobío	$329.970	$460.018	$206.668	$296.163	$278.970	$392.244
La Araucanía	$235.327	$367.952	$196.915	$291.315	$218.196	$333.774
Los Ríos	$256.850	$390.816	$201.283	$293.801	$233.669	$350.344
Los Lagos	$246.784	$378.340	$188.118	$275.673	$220.635	$332.579
Aysén	$251.023	$384.551	$176.391	$264.678	$218.083	$331.644
Magallanes	$420.720	$534.912	$221.798	$296.078	$333.728	$430.466
Sin info	$349.649	$449.516	$227.912	$305.751	$300.972	$392.031
Total	$355.289	$482.543	$218.591	$303.554	$294.226	$402.589

(1) Corresponde a las pensiones pagadas en las modalidades de Retiro Programado, Renta Temporal y Renta Vitalicia a pensionados por vejez.
(2) Nuevo beneficio generado a partir de la Ley 21.419 de 29.01.2022. A contar del mes de junio de 2022 el valor de la PGU aumentó a $193.917, el cual comenzó a ser pagado desde el mes de julio. Desde febrero de 2023 el valor de la PGU es de $206.173. Los montos del beneficio pagado reflejados en esta estadística no incluyen el pago retroactivo de la diferencia generada respecto del valor anterior. A partir de agosto de 2022, según la Ley 21.419 se consideran nuevos requisitos de acceso a la PGU, ampliándose la cobertura a quienes no integren un grupo familiar perteneciente al 10% más rico de la población de 65 o más años, según el nuevo instrumento de focalización definido en la misma Ley.
(3) La región corresponde al domicilio que el pensionado tiene registrado en la AFP.
(4) Incluye el componente autofinanciado más APS o PGU en los casos que corresponda.

Encuesta de Caracterización Socioeconómica Nacional: CASEN

Los resultados obtenidos, de la Encuesta de Caracterización Socioeconómica Nacional (Casen), desempeñan un papel fundamental para el diagnóstico y la evaluación de las políticas sociales implementadas por el Estado ya que estos resultados entregan una fotografía numérica de la realidad social del momento. El tener claro tanto la finalidad de la CASEN como las dimensiones involucradas: ingresos, educación, salud, vivienda, trabajo, entre otras permite saber de qué se trata cuando dicen CASEN y de ese modo, tener más claro qué tipo de pregunta se puede hacer al utilizar los datos de esta encuesta.

El monto de la jubilación recibido por los adultos mayores, es parte del módulo ingresos. Dimensión que tiene como objetivo en la Casen determinar la situación de pobreza del país; en términos de ruralidad, urbanidad como a nivel regional. Otro objetivo es conocer las brechas salarias que se dan entre distintos grupos sociales; niños y adolescentes, adultos mayores, mujeres y hombres, personas inmigrantes o pertenecientes a grupos indígenas, entre otros.

La dimensión “ingresos” aborda las diferentes categorías de ingresos que reciben las personas y los hogares, esto es, los ingresos primarios, constituidos por los ingresos provenientes del trabajo (de los asalariados y de los empleadores y trabajadores por cuenta propia) y de la propiedad (retornos por activos financieros y no financieros), así como las transferencias corrientes, compuestas por las jubilaciones, pensiones y montepíos, los subsidios o transferencias monetarias del Estado y las diversas transferencias corrientes entre hogares.

Las preguntas miden todos los tipos pensiones que existen y se suman en el total de ingresos que recibe la persona o el hogar. Sin embargo, para este estudio solo estudiaremos la variable; monto de jubilación entregada por las AFP en modo corregido: ‘y2803c’ y se excluye todas las otras mediciones que agregan en su monto los bonos entregados por el Estado.
Como se señala, la encuesta en la dimensión ingresos se enfoca en los totales percibidos por la persona y el hogar al cual pertenece. El monto de las pensiones pagadas por la AFP solo es una variable más entre tantas, por lo tanto, las otras variables de la base de datos no están enfocadas en aquello.

En este caso, para indagar y medir la relación de las variables que puedan influir en una mejor jubilación de la AFP y considerando la base de datos de la Casen, tiene variables que hablan del presente de la persona y el monto de la pensión es producto de la vida laboral del jubilado o jubilada, tiempo pasado, y eso no es parte de la base de datos, por lo tanto, las variables utilizadas para este análisis son variables que caracterizan a la persona en cuanto lugar de residencia, sexo, etc. No se considera la dimensión vida laboral y previsional previa ni se considera la dimensión del mercado laboral asociado a la vida laboral del cotizante.

Pregunta de investigación

¿Existe una relación significativa entre el nivel educativo de los jubilados y el monto de sus jubilaciones entregadas por la AFP?

Hipótesis Nula (H0): βaños de educación = 0

Hipótesis Alternativa (H1): βaños de educación ≠ 0

La regla de decisión:

Si el valor p es menor a 0.05, podemos rechazar la hipótesis nula y concluir que hay una relación significativa entre los años de educación y el monto pagado de jubilación.

Los tipos de error:

Error del Tipo I

Sería aceptar que hay una relación significativa entre los años de educación y el monto pagado de jubilación cuando no la hay. Es decir, aceptar la hipótesis alternativa cuando la verdadera es la hipótesis nula.

Error Tipo II

Implicaría rechazar una relación significativa entre los años de educación y el monto pagado de jubilación cuando sí hay relación significativa. Es decir, aceptar la hipotesis nula cuando la verdadera es la hipótesis alternativa.

Objetivo

Explorar el impacto que tiene los años de educación sobre el monto de la jubilación entregado por la AFP

Carga, revisión, limpieza y análisis descriptivo de los datos

In [54]:
```
df = pd.read_stata('Base de datos Casen 2022 STATA.dta', convert_categoricals=False
print(df.shape)
df.head()
```

```
(202231, 917)
```

Out[54]:

	id_vivienda	folio	id_persona	region	area	cod_upm	nse	estrato	hogar	expr	...	m
0	1000901	100090101	1	16	2	10009	4	1630324	1	43	...	
1	1000901	100090101	2	16	2	10009	4	1630324	1	43	...	
2	1000901	100090101	3	16	2	10009	4	1630324	1	44	...	
3	1000902	100090201	1	16	2	10009	4	1630324	1	51	...	
4	1000902	100090201	2	16	2	10009	4	1630324	1	51	...	

5 rows × 917 columns

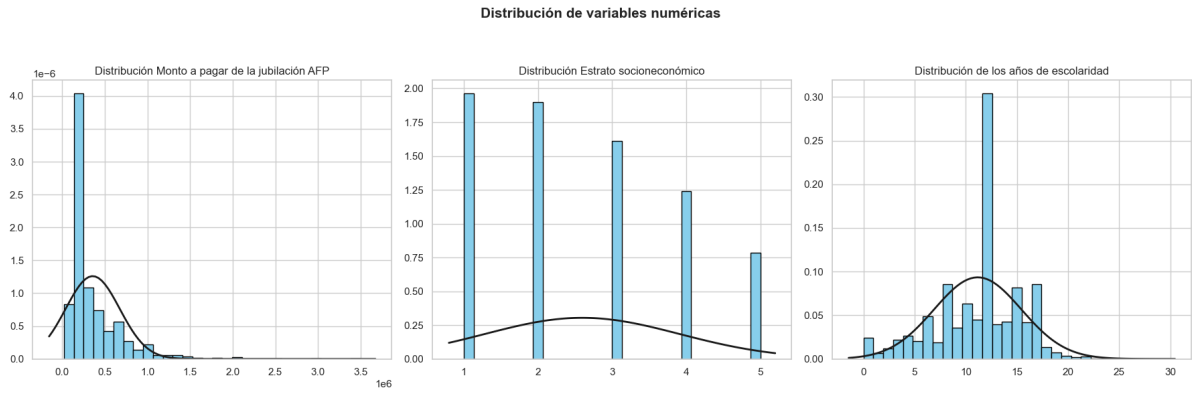
In [55]:
```
# Listar el tipo de dato de cada variable a estudiar

# Generar lista con las variables a revisar|
variables_interesantes = [
'region','area','qaut','edad','sexo','esc',
'y7', 'yah1','yah2', 'yrut', 'yre1','yre2', 'yre3',
'y2803c', 's13', 'v17', 'oficio1_08'
]

# Obtener los tipos de datos de las variables seleccionadas
tipos_de_datos = df[variables_interesantes].dtypes

# Visualizar las filas deseadas
print(tipos_de_datos.head(25))
```

```
region           int8
area             int8
qaut          float64
edad            int16
sexo             int8
esc           float64
y7            float64
yah1          float64
yah2          float64
yrut          float64
yre1          float64
yre2          float64
yre3          float64
y2803c        float64
s13              int8
v17           float64
oficio1_08    float64
dtype: object
```

In [56]:

```
# Crear una figura y subgráficos (subplots)
fig, axs = plt.subplots(1, 3, figsize=(18, 6))


# Subgráfico 1 - Histograma de Años de Escolaridad
axs[0].hist(df['y2803c'], bins=30, color='skyblue', edgecolor='black', density=True

# Ajuste de distribución normal
xmin, xmax = axs[0].get_xlim()
x = np.linspace(xmin, xmax, 100)

# Utilizar la media y la desviación estándar de la columna
mean_y2803c = np.mean(df['y2803c'])
std_y2803c = np.std(df['y2803c'])
p = norm.pdf(x, mean_y2803c, std_y2803c)

axs[0].plot(x, p, 'k', linewidth=2)
axs[0].set_title('Distribución Monto a pagar de la jubilación AFP')


# Subgráfico 2 - Histograma de Años de Escolaridad
axs[1].hist(df['qaut'], bins=30, color='skyblue', edgecolor='black', density=True)

# Ajuste de distribución normal
xmin, xmax = axs[1].get_xlim()
x = np.linspace(xmin, xmax, 100)

# Utilizar la media y la desviación estándar de la columna
mean_qaut = np.mean(df['qaut'])
std_qaut = np.std(df['qaut'])

# Utilizar la media y la desviación estándar de la columna
p = norm.pdf(x, mean_qaut, std_qaut)

axs[1].plot(x, p, 'k', linewidth=2)
axs[1].set_title('Distribución Estrato socioeconómico')


# Subgráfico 2 - Histograma de Años de Escolaridad
axs[2].hist(df['esc'], bins=30, color='skyblue', edgecolor='black', density=True)

# Ajuste de distribución normal
xmin, xmax = axs[2].get_xlim()
x = np.linspace(xmin, xmax, 100)
```

```
# Utilizar la media y la desviación estándar de la columna
mean_esc = np.mean(df['esc'])
std_esc = np.std(df['esc'])

# Utilizar la media y la desviación estándar de la columna
p = norm.pdf(x, mean_esc, std_esc)

axs[2].plot(x, p, 'k', linewidth=2)
axs[2].set_title('Distribución de los años de escolaridad')


fig.tight_layout()
plt.subplots_adjust(top=0.8)  # Ajuste para que el título no se solape con los subg
fig.suptitle('Distribución de variables numéricas', fontsize=15, fontweight="bold")
plt.show()
```



In [57]:
```
# limpieza de datos

df = df[df['y7'] != -88]
df = df[df['s13'] != -88]
df = df[df['v17'] != -88]


# Crear un nuevo DataFrame con las columnas concatenadas
new_df = pd.concat([df, pd.DataFrame({'sexo_dic': df['sexo'].replace({1: 0, 2: 1}),
                                      'area_dic': df['area'].replace({1: 0, 2: 1})}

# Si es necesario, puedes asignar el nuevo DataFrame de vuelta a 'df'
df = new_df.copy()

#Realizar estadistica descriptiva a las variables seleccionadas:
columnas=['region','area_dic','qaut','edad','sexo_dic','esc',
'y7', 'yah1','yah2', 'yrut', 'yre1','yre2', 'yre3',
'y2803c', 's13', 'v17'
]

estadistica = df[columnas].describe().round(0)

# Obtener el índice del DataFrame estadistica
columnas_estadistica = estadistica.index

# Seleccionar solo algunos estadísticos específicos
estadisticos_seleccionados = estadistica.loc[columnas_estadistica.isin(['count', 'm

# Transponer el DataFrame (intercambiar filas y columnas)
estadisticos_seleccionados_transpuesto = estadisticos_seleccionados.transpose()

# Mostrar el DataFrame transpuesto
print("\nEstadística Descriptiva:")
(estadisticos_seleccionados_transpuesto)
```

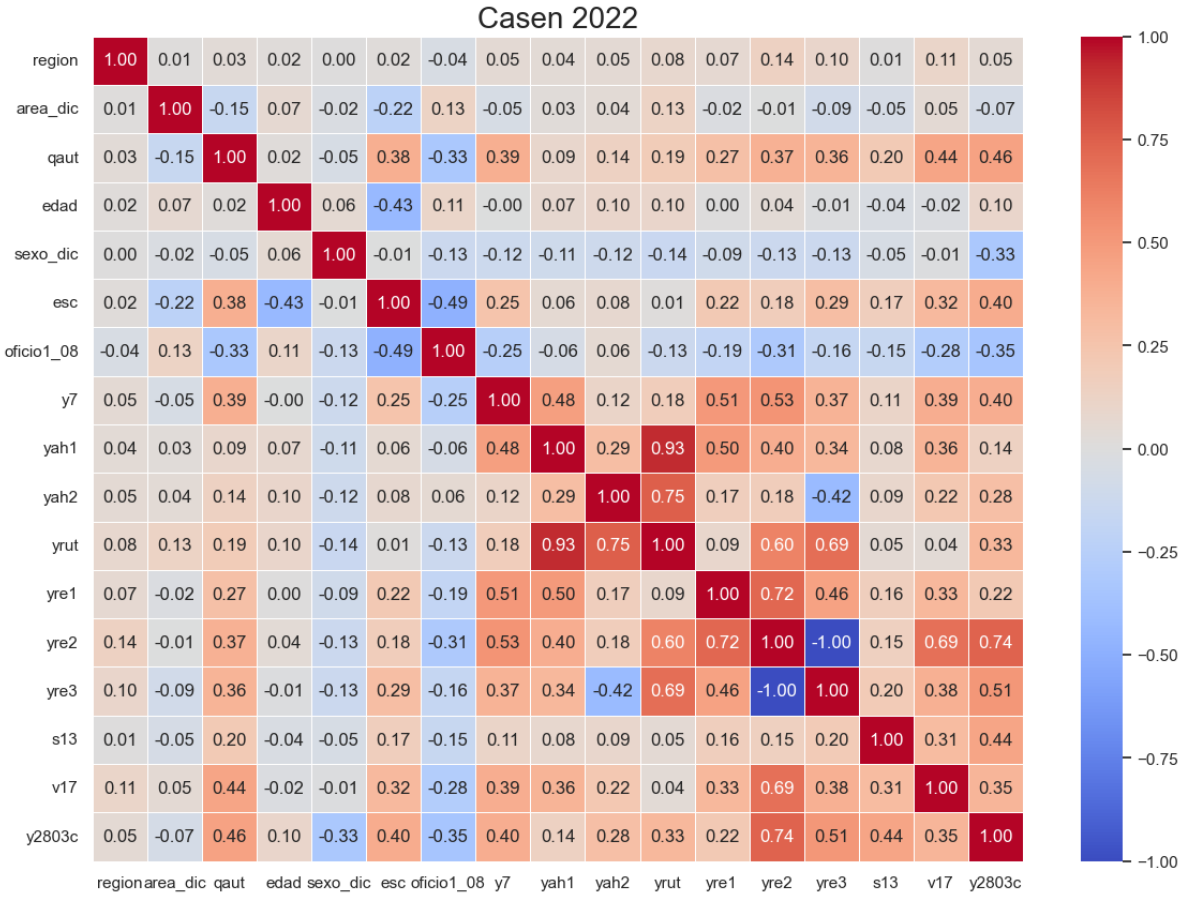
Estadística Descriptiva:

Out[57]:

	count	mean	std	min	max
region	198372.0	9.0	4.0	1.0	16.0
area_dic	198372.0	0.0	0.0	0.0	1.0
qaut	198260.0	3.0	1.0	1.0	5.0
edad	198372.0	39.0	23.0	0.0	120.0
sexo_dic	198372.0	1.0	0.0	0.0	1.0
esc	161939.0	11.0	4.0	0.0	29.0
y7	22596.0	452331.0	810807.0	0.0	40000000.0
yah1	927.0	90836.0	356969.0	83.0	6000000.0
yah2	276.0	203000.0	656081.0	167.0	5416667.0
yrut	292.0	693334.0	1501038.0	417.0	16666667.0
yre1	4039.0	420676.0	579264.0	10000.0	20000000.0
yre2	164.0	200831.0	381990.0	3333.0	3000000.0
yre3	278.0	162311.0	204438.0	3333.0	2000000.0
y2803c	7418.0	357016.0	315286.0	20000.0	3500000.0
s13	198372.0	1.0	1.0	1.0	5.0
v17	18365.0	343379.0	297027.0	22000.0	4000000.0

In [58]:

```
# Selecciona las columnas de interés
columnas_interes = df[['region','area_dic','qaut','edad','sexo_dic','esc', 'oficio1
                       'y7', 'yah1','yah2', 'yrut', 'yre1','yre2', 'yre3',
                       's13','v17', 'y2803c'
                       ]]

# Configura el estilo de Seaborn (opcional)
sns.set(style="whitegrid")

# Crea un mapa de calor a partir de las columnas seleccionadas
plt.figure(figsize=(14, 10))  # Tamaño de la figura
sns.heatmap(columnas_interes.corr(), annot=True, cmap="coolwarm", linewidths=0.5, f


plt.title("Casen 2022", fontsize=20)  # Título del gráfico
plt.show()
```



El monto de la jubilación recibido por los adultos mayores, es parte del módulo ingresos. Dimensión que tiene como objetivo determinar la situación de pobreza del país; en terminos de ruralidad, urbanidad, a nivel regional y de conocer las brechas salarias que se dan entre distintos grupos sociales: niños y adolescentes, adultos mayores, mujeres y hombres, personas inmigrantes o pertenecientes a grupos indígenas, entre otros.

La dimensión “ingresos” aborda las diferentes categorías de ingresos que reciben las personas y los hogares, esto es, los ingresos primarios, constituidos por los ingresos provenientes del trabajo (de los asalariados y de los empleadores y trabajadores por cuenta propia) y de la propiedad (retornos por activos financieros y no financieros), así como las transferencias corrientes, compuestas por las jubilaciones, pensiones y montepíos, los subsidios o transferencias monetarias del Estado y las diversas transferencias corrientes entre hogares.

Las preguntas miden todos los tipos pensiones que existen y se suman en el total de ingresos que recibe la persona o el hogar. Sin embargo, para este estudio solo estudiaremos la variable; monto de jubilación entregada por las AFP y que esta corregida (‘y2803c’), se excluye todas las otras mediciones que agregan en su monto bonos entregados por el Estado.

Tal como se señala, la encuesta se enfoca en los ingresos totales y el monto de las pensiones pagadas por la AFP es una variable más entre tantas. Como la Casen no es un estudio relacionado con las jubilaciones, las otras variables de la base de datos no están en relación directa con el fenomeno de la pensiones entregadas por las AFP.

Sin embargo, se ha querido estudiar el fenómeno de la jubilaciones en relación a las variables que se presentan en el estudio, buscando alguna relación entre esta variable y

otras variables del tipo de caracterización e identificación (sexo, región, estrato social, etc.)

Regresiones

Especificación.

La función base es entre monto jubilacion de la AFP y educación:

$$\text{Monto jubilacion AFP} = f(Educacion))$$

Modelo poblaciónal y estimación

Modelo según la población:

$$\text{Monto de la jubilación AFP}_i = \alpha + \beta \text{Años de educación}_i + \mu_i$$

Estimadores para la muestra:

$$\text{Monto de la } \hat{\text{J}}\text{ubilacion AFP}_i = \hat{\alpha} + \hat{\beta} \text{Años de educación}_i$$

Regresión Lineal Simple

Primer modelo:

En este modelo los datos se agruparon en la variable categórica región, quedando la muestra con solo 16 casos para hacer la regresión.

Modelo según la población:

$$\text{Monto de la jubilación AFP}_i = \alpha + \beta \text{Años de educación}_i + \mu$$

Datos agrupados por región

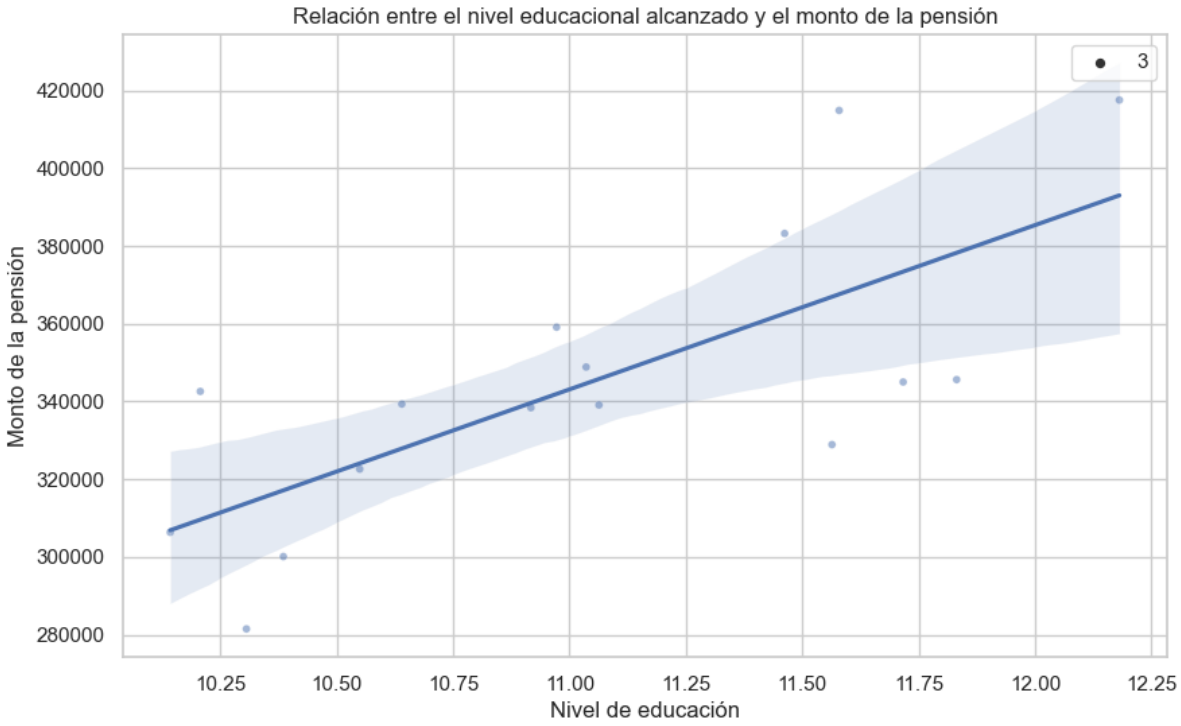
In [59]:
```
# Agrupar por 'region' y aplicar funciones de agregación
#df_prep = df.agg({'y28_2c': 'mean', 'y1': 'mean'}).reset_index()
df_regiones = df.groupby('region').agg({'y2803c': 'mean', 'esc': 'mean'}).reset_inc

# Resultados agregados por región
df_regiones.head(30)
```

Out[59]:

	region	y2803c	esc
0	1	328890.517375	11.563784
1	2	344992.243056	11.716684
2	3	339083.072519	11.063391
3	4	338381.748031	10.917202
4	5	383196.991407	11.462059
5	6	322602.229656	10.549499
6	7	281480.021033	10.306163
7	8	359100.048780	10.972305
8	9	300087.055416	10.385506
9	10	306333.966480	10.142479
10	11	348841.134752	11.035799
11	12	414856.698305	11.579208
12	13	417525.815182	12.181102
13	14	339337.454294	10.639926
14	15	345610.176056	11.831498
15	16	342576.506667	10.207157

In [60]:

```
# Configurar el tamaño de la figura
plt.figure(figsize=(10, 6))

# Crear el gráfico de dispersión con línea de regresión
sns.set(style="whitegrid")
sns.scatterplot(data=df_regiones, x='esc', y='y2803c', alpha=0.5, size=3)
sns.regplot(data=df_regiones, x='esc', y='y2803c', scatter=False)

plt.xlabel("Nivel de educación")
plt.ylabel("Monto de la pensión")
plt.title("Relación entre el nivel educacional alcanzado y el monto de la pensión")

# Mostrar el gráfico
plt.show()
```



In [61]:
```
# Correlación lineal entre las dos variables
# ==============================================================================
corr_test = pearsonr(x = df_regiones['y2803c'], y =  df_regiones['esc'])
print("Coeficiente de correlación de Pearson: ", corr_test[0])
print("P-value: ", corr_test[1])
```

```
Coeficiente de correlación de Pearson:  0.7286640032013907
P-value:  0.0013662686245193446
```

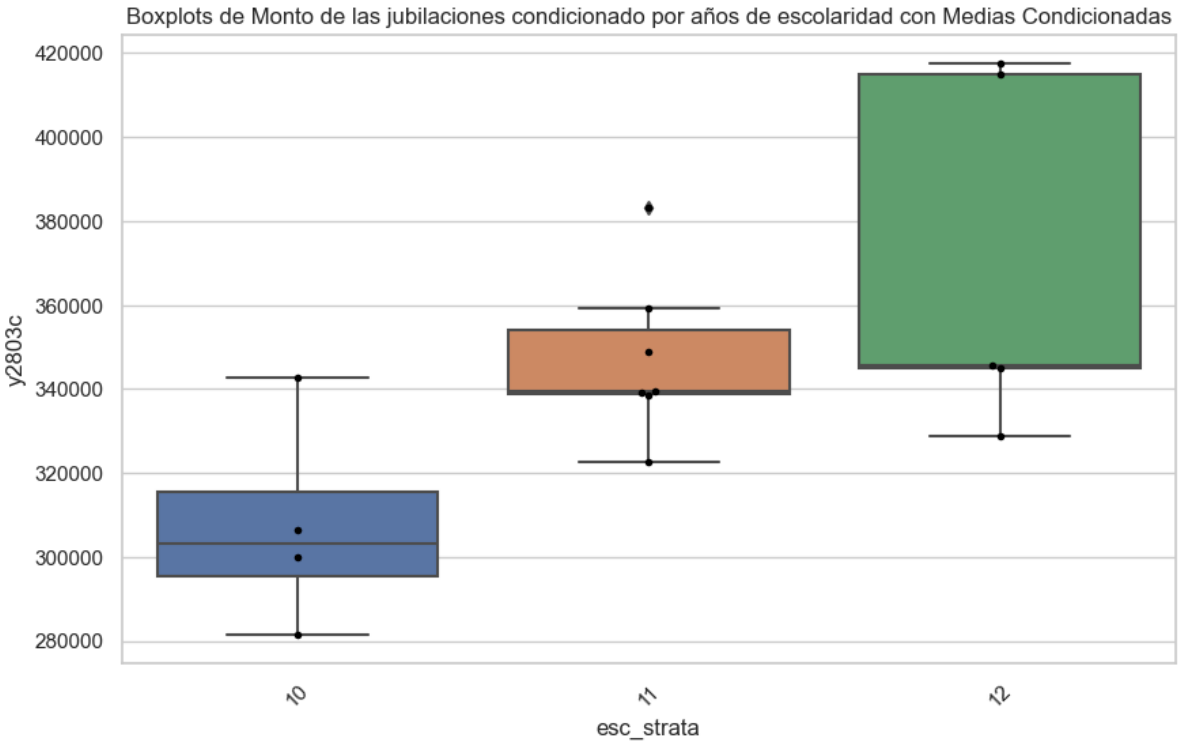
In [ ]:
```
promedio_monto_pensiones = df_regiones['y2803c'].mean()
sd_monto_pensiones = df_regiones['y2803c'].std()
promedio_escolaridad = df_regiones['esc'].mean()
sd_escolaridad = df_regiones['esc'].std()

resumen_estadistico = pd.DataFrame({
    'promedio_Monto jubilaciones': [promedio_monto_pensiones],
    'sd_Monto jubilaciones': [sd_monto_pensiones],
    'promedio_escolaridad': [promedio_escolaridad],
    'sd_escolaridad': [sd_escolaridad]
})

resumen_estadistico.head().round(3)
```

In [ ]:
```
# DataFrame a analizar
sns.regplot(data=df_regiones, y='y2803c', x='esc', ci=95, line_kws={'color': 'sprir

# El argumento ci controla el intervalo de confianza
plt.ylabel('Monto jubilación')
plt.xlabel('Nivel escolaridad')
plt.title('Scatter Plot con Línea de Regresión y Intervalo de Confianza')
plt.show()
```

In [63]:
```
# DataFrame a analizar
sns.regplot(df_regiones, y='y2803c', x='esc', ci=95, line_kws={'color': 'springgree

# Procesar y agregar etiquetas de región a los puntos
for i, label in enumerate(df_regiones['region']):
    last_word = str(label).split()[-1]  # Obtener la última palabra de la etiqueta
    plt.text(df_regiones['esc'][i], df_regiones['y2803c'][i], last_word, fontsize=8
```

```
plt.ylabel('Monto de la Pension')
plt.xlabel('Escolaridad')
plt.title('Scatter Plot con Línea de Regresión e Identificacion de la Región')
plt.show()
```

In [64]:
```
# Crear una nueva columna 'esc_strata' con los valores redondeados de 'esc'
df_regiones['esc_strata'] = df_regiones['esc'].round().astype(int)

# Crear el gráfico de boxplots
plt.figure(figsize=(10, 6))  # Tamaño del gráfico
sns.boxplot(data=df_regiones, x='esc_strata', y='y2803c')

# Agregar puntos para mostrar las medias condicionadas
sns.swarmplot(data=df_regiones, x='esc_strata', y='y2803c', color='black', size=4)

plt.xlabel('esc_strata')
plt.ylabel('y2803c')
plt.title('Boxplots de Monto de las jubilaciones condicionado por años de escolaric
plt.xticks(rotation=45)  # Rotar etiquetas del eje x si es necesario

plt.show()
```



In [65]:
```
# Eliminar filas con valores NaN
df_clean = df_regiones.dropna(subset=['y2803c', 'esc'])

# Agregar una columna de constantes para el término constante en el modelo
df_clean['constante'] = 1

# Definir las variables dependiente e independiente
y = df_clean['y2803c']
X = df_clean[['constante', 'esc']]  # Usar 'constante' como término constante

# Ajustar el modelo de regresión lineal
simple_agrup_region = sm.OLS(y, X).fit()

# Imprimir un resumen del modelo
print(simple_agrup_region.summary())

# Prueba de hipótesis sobre la variable de escolaridad
print("\nPrueba de hipótesis:")
print("\nH0: La escolaridad no tiene efecto significativo en el monto a pagar de la
print("H1: La escolaridad tiene un efecto significativo en el monto a pagar de la j

# Definimos nuestra hipótesis nula
# El valor p está asociado con la variable escolaridad (coeficiente)
valor_p_escolaridad = simple_agrup_region.pvalues[1]

# Nivel de significancia (usualmente 0.05)
nivel_significancia = 0.05

if valor_p_escolaridad < nivel_significancia:
    print(f"\nEl valor p es {valor_p_escolaridad}, por lo tanto, rechazamos la hipó
    print("\nLa escolaridad tiene un efecto significativo en el monto a pagar de la
else:
    print(f"\nEl valor p es {valor_p_escolaridad}, por lo tanto, no podemos rechaza
    print("\nNo hay suficiente evidencia para afirmar que la escolaridad tenga un e
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 y2803c   R-squared:                       0.531
Model:                            OLS   Adj. R-squared:                  0.497
Method:                 Least Squares   F-statistic:                     15.85
Date:                Mon, 08 Jan 2024   Prob (F-statistic):            0.00137
Time:                        13:48:24   Log-Likelihood:                -184.31
No. Observations:                  16   AIC:                             372.6
Df Residuals:                      14   BIC:                             374.2
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
constante  -1.218e+05   1.17e+05     -1.038      0.317   -3.73e+05     1.3e+05
esc         4.226e+04   1.06e+04      3.981      0.001    1.95e+04     6.5e+04
==============================================================================
Omnibus:                        0.492   Durbin-Watson:                   1.602
Prob(Omnibus):                  0.782   Jarque-Bera (JB):                0.549
Skew:                           0.088   Prob(JB):                        0.760
Kurtosis:                       2.109   Cond. No.                         201.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly s
pecified.

Prueba de hipótesis:

H0: La escolaridad no tiene efecto significativo en el monto a pagar de la jubilac
ion de la AFP
H1: La escolaridad tiene un efecto significativo en el monto a pagar de la jubilac
ion de la AFP

El valor p es 0.0013662686245193552, por lo tanto, rechazamos la hipótesis nula.

La escolaridad tiene un efecto significativo en el monto a pagar de la jubilacion
de la AFP.
C:\Users\usuario\anaconda3\Lib\site-packages\scipy\stats\_stats_py.py:1806: UserWa
rning: kurtosistest only valid for n>=20 ... continuing anyway, n=16
  warnings.warn("kurtosistest only valid for n>=20 ... continuing "
```

Interpretación del modelo

La constante ($\beta 0$) cuyo valor es -125059.781 representa el valor esperado del monto de la jubilación cuando todas las variables independientes son cero. Sin embargo, en el contexto de la escolaridad no es significativa.

El coeficiente de escolaridad (βAños de educación cuyo valor es 42548.134) indica que por cada año de estudio el monto de la jubilación pagada por la AFP aumenta en 42.548 pesos.

Dado que el coeficiente de escolaridad es positivo, se interpreta que un aumento en la escolaridad está asociado con un aumento en el monto de la jubilación, el resto se mantiene constante.

Segundo modelo:

A la agrupación regional se suma la agrupación de la variable área, lo cual por el cruce de las variables se obtiene una muestra de 32 casos.

Se mantiene el modelo de regresión lineal simple porque se mantiene la variable dependiente educación y no se agrega otra variable al modelo.

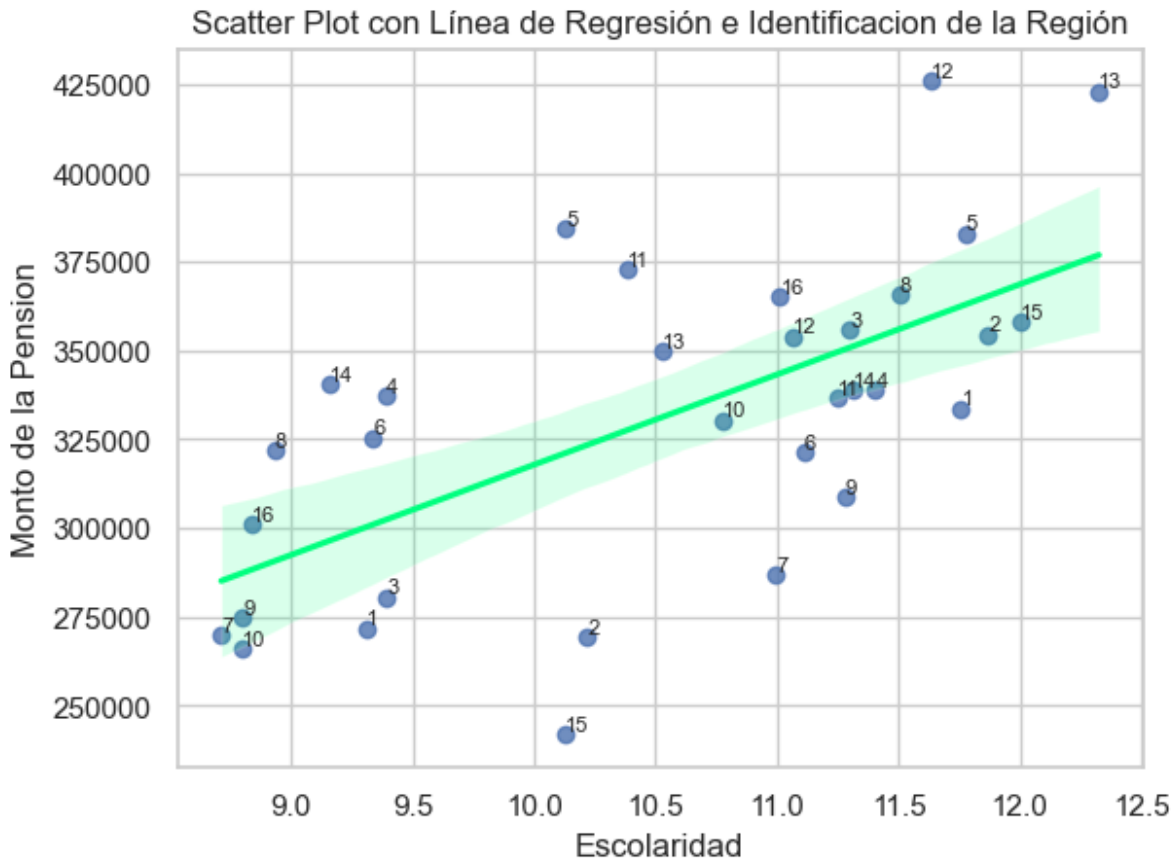
Modelo según la población:

$$\text{Monto de la jubilación AFP}_i = \alpha + \beta \text{Años de educación}_i + \mu$$

Datos agrupados por región y área

In [66]:
```
# Agrupar por 'region' y área
df_agrupado = df.groupby(['region','area']).agg({'y2803c': 'mean', 'esc': 'mean'}).

# Resultados agregados
df_agrupado.head(20)

# Obtener los valores mínimos y máximos

min_esc = df_agrupado['esc'].min()
max_esc = df_agrupado['esc'].max()

min_jub = df_agrupado['y2803c'].min()
max_jub = df_agrupado['y2803c'].max()

# Imprimir los resultados
print(f'Mínimo de escolaridad:{round(min_esc, 1)}')
print(f'Máximo de escolaridad:{round(max_esc, 1)}')
print(f'Mínimo de jubilación: {round(min_jub, 1)}')
print(f'Máximo de jubilación: {round(max_jub), 1}')

# DataFrame a analizar
sns.regplot(df_agrupado, y='y2803c', x='esc', ci=95, line_kws={'color': 'springgree

# Procesar y agregar etiquetas de región a los puntos
for i, label in enumerate(df_agrupado['region']):
    last_word = str(label).split()[-1]  # Obtener la última palabra de la etiqueta
    plt.text(df_agrupado['esc'][i], df_agrupado['y2803c'][i], last_word, fontsize=8

plt.ylabel('Monto de la Pension')
plt.xlabel('Escolaridad')
plt.title('Scatter Plot con Línea de Regresión e Identificacion de la Región')
plt.show()
```

```
Mínimo de escolaridad:8.7
Máximo de escolaridad:12.3
Mínimo de jubilación: 241724.0
Máximo de jubilación: (425811, 1)
```



In [67]:
```
# Eliminar filas con valores NaN
df_agrupado = df_agrupado.dropna(subset=['y2803c', 'esc', 'region','area'])

# Agregar una columna de constantes para el término constante en el modelo
df_agrupado['constante'] = 1


# Definir las variables dependiente e independiente
y = df_agrupado['y2803c']
X = df_agrupado[['constante', 'esc']]

# Ajustar el modelo de regresión lineal
simple_agrup_reg_area = sm.OLS(y, X).fit()

# Imprimir un resumen del modelo
print(simple_agrup_reg_area.summary())

# Prueba de hipótesis sobre la variable de escolaridad
print("\nPrueba de hipótesis:")
print("\nH0: La escolaridad no tiene efecto significativo en el monto a pagar de la
print("H1: La escolaridad tiene un efecto significativo en el monto a pagar de la j

# Definimos nuestra hipótesis nula
# El valor p está asociado con la variable escolaridad (coeficiente)
valor_p_escolaridad = simple_agrup_reg_area.pvalues[1]

# Nivel de significancia (usualmente 0.05)
nivel_significancia = 0.05

if valor_p_escolaridad < nivel_significancia:
    print(f"\nEl valor p es {valor_p_escolaridad}, por lo tanto, rechazamos la hipó
    print("\nLa escolaridad tiene un efecto significativo en el monto a pagar de la
else:
    print(f"\nEl valor p es {valor_p_escolaridad}, por lo tanto, no podemos rechaza
    print("\nNo hay suficiente evidencia para afirmar que la escolaridad tenga un e
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 y2803c   R-squared:                       0.403
Model:                            OLS   Adj. R-squared:                  0.383
Method:                 Least Squares   F-statistic:                     20.26
Date:                Mon, 08 Jan 2024   Prob (F-statistic):           9.48e-05
Time:                        13:49:07   Log-Likelihood:                -379.47
No. Observations:                  32   AIC:                             762.9
Df Residuals:                      30   BIC:                             765.9
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
constante   6.397e+04   5.96e+04      1.073      0.292   -5.78e+04    1.86e+05
esc         2.539e+04   5640.393      4.501      0.000    1.39e+04    3.69e+04
==============================================================================
Omnibus:                        0.030   Durbin-Watson:                   1.820
Prob(Omnibus):                  0.985   Jarque-Bera (JB):                0.203
Skew:                          -0.054   Prob(JB):                        0.903
Kurtosis:                       2.625   Cond. No.                         102.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly s
pecified.

Prueba de hipótesis:

H0: La escolaridad no tiene efecto significativo en el monto a pagar de la jubilac
ion de la AFP
H1: La escolaridad tiene un efecto significativo en el monto a pagar de la jubilac
ion de la AFP

El valor p es 9.48447513058634e-05, por lo tanto, rechazamos la hipótesis nula.

La escolaridad tiene un efecto significativo en el monto a pagar de la jubilacion
de la AFP.
```

Interpretación del modelo

La constante (57420) representa el valor esperado del monto de la jubilación cuando todas las variables independientes son cero. Sin embargo en este contexto no tiene significado práctico.

El coeficiente de escolaridad (26000) significa que el aumento en un año de escolaridad se asocia con un aumento de 26.000 pesos en el monto de la jubilación, manteniendo constantes las otras variables en el modelo.

La significancia del modelo en conjunto (Prob (F-statistic)) y el porcentaje de variabilidad explicada (R-squared) sugieren que este modelo tiene utilidad para explicar el monto de la jubilación en el contexto específico de la agrupación por región.

Tercer modelo:

El modelo se aplica a la muestra total sin agrupamiento de ningún tipo. Lo que se obtiene una muestra de 7349 casos.

Solo se mantiene la variable dependiente educación y no se agrega otra variable al modelo por lo tanto, se mantiene el mismo modelo que describe la relación.

Modelo según la población:

$$\text{Monto de la jubilación AFP}_i = \alpha + \beta \text{Años de educación}_i + \mu$$

Datos totales

In [68]:
```
# Eliminar filas con valores NaN
df_clean = df.dropna(subset=['y2803c', 'esc'])

# Agregar una columna de constantes para el término constante en el modelo
df_clean['constante'] = 1


# Definir las variables dependiente e independiente
y = df_clean['y2803c']
X = df_clean[['constante', 'esc']]

# Ajustar el modelo de regresión lineal
simple_totaldatos = sm.OLS(y, X).fit()

# Imprimir un resumen del modelo
print(simple_totaldatos.summary())

# Prueba de hipótesis sobre la variable de escolaridad
print("\nPrueba de hipótesis:")
print("\nH0: La escolaridad no tiene efecto significativo en el monto a pagar de la
print("H1: La escolaridad tiene un efecto significativo en el monto a pagar de la j

# Definimos nuestra hipótesis nula
# El valor p está asociado con la variable escolaridad (coeficiente)
valor_p_escolaridad = simple_totaldatos.pvalues[1]

# Nivel de significancia (usualmente 0.05)
nivel_significancia = 0.05

if valor_p_escolaridad < nivel_significancia:
    print(f"\nEl valor p es {valor_p_escolaridad}, por lo tanto, rechazamos la hipó
    print("\nLa escolaridad tiene un efecto significativo en el monto a pagar de la
else:
    print(f"\nEl valor p es {valor_p_escolaridad}, por lo tanto, no podemos rechaza
    print("\nNo hay suficiente evidencia para afirmar que la escolaridad tenga un e
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 y2803c   R-squared:                       0.163
Model:                            OLS   Adj. R-squared:                  0.163
Method:                 Least Squares   F-statistic:                     1431.
Date:                Mon, 08 Jan 2024   Prob (F-statistic):          3.04e-286
Time:                        13:49:10   Log-Likelihood:            -1.0284e+05
No. Observations:                7349   AIC:                         2.057e+05
Df Residuals:                    7347   BIC:                         2.057e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
constante   1.036e+05   7509.355     13.796      0.000    8.89e+04    1.18e+05
esc         2.586e+04    683.705     37.826      0.000    2.45e+04    2.72e+04
==============================================================================
Omnibus:                     4147.761   Durbin-Watson:                   1.872
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            50206.149
Skew:                           2.475   Prob(JB):                         0.00
Kurtosis:                      14.810   Cond. No.                         24.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly s
pecified.

Prueba de hipótesis:

H0: La escolaridad no tiene efecto significativo en el monto a pagar de la jubilac
ion de la AFP
H1: La escolaridad tiene un efecto significativo en el monto a pagar de la jubilac
ion de la AFP

El valor p es 3.042544001279201e-286, por lo tanto, rechazamos la hipótesis nula.

La escolaridad tiene un efecto significativo en el monto a pagar de la jubilacion
de la AFP.
C:\Users\usuario\AppData\Local\Temp\ipykernel_8764\2588165157.py:5: SettingWithCop
yWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stabl
e/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_clean['constante'] = 1
```

Interpretación del modelo

En relación a los coeficientes que son parte de la regresión:

La constante (102000) representa el valor esperado del monto de la jubilación cuando todas las variables independientes son cero. En este contexto no tiene significado.

El coeficiente de escolaridad (26050) indica que por cada cada de estudios formales el monto de la jubilación se ve incrementado en 26.050 pesos, manteniendo constantes otras variables en el modelo.

En resumen, según este modelo, la escolaridad tiene un impacto significativo en el monto de la jubilación. Un aumento en la escolaridad se asocia con un aumento de 26050 unidades en el monto de la jubilación, manteniendo constantes otras variables en el modelo.

La significancia del modelo en conjunto (Prob (F-statistic)) y el porcentaje de variabilidad explicada (R-squared) sugieren que este modelo tiene utilidad para explicar el monto de la jubilación en el contexto de los datos totales.

Regresión Lineal Múltiple

A partir de este momento se empieza con el enriquecimiento del modelo. A la variable educación se le agrega la variable dicotómica sexo, se suma además la interacción entre esta variable y la variable educación

Cuarto modelo:

Modelo según la población:

$$\text{Monto de la jubilación AFP}_i = \beta_0 + \beta_1 educ + \beta_2 sexo + (\beta_1 e$$

Datos enriquecidos

In [69]:
```
# Columna a estudiar
frecuencia_sex_dic = df['sexo_dic'].value_counts()

# Mostrar la frecuencia de la variable
print(frecuencia_sex_dic,"\n")

# Crear una variable de interacción entre 'esc' y 'sexo'
df['interaccion_sexo_esc'] = df['esc'] * df['sexo_dic']

# Eliminar filas con valores NaN
df_clear = df.dropna(subset=['y2803c', 'esc', 'sexo_dic', 'interaccion_sexo_esc'])

# Agregar una columna de constantes para el término constante en el modelo
df_clear['constante'] = 1


# Ajustar el modelo de regresión lineal con la variable de interacción
modelo = sm.OLS(df_clear['y2803c'], sm.add_constant(df_clear[['constante','esc', 's

# Obtener los resultados de la regresión
resultados = modelo.summary()
print(resultados)

# Prueba de hipótesis sobre la variable de escolaridad
print("\nPrueba de hipótesis:")
print("\nH0: La escolaridad no tiene efecto significativo en el monto a pagar de la
print("H1: La escolaridad tiene un efecto significativo en el monto a pagar de la j

# Definimos nuestra hipótesis nula
# El valor p está asociado con la variable escolaridad (coeficiente)
valor_p_escolaridad = modelo.pvalues[1]

# Nivel de significancia (usualmente 0.05)
nivel_significancia = 0.05

if valor_p_escolaridad < nivel_significancia:
```

```
    print(f"\nEl valor p es {valor_p_escolaridad}, por lo tanto, rechazamos la hipó
    print("\nLa escolaridad tiene un efecto significativo en el monto a pagar de la
else:
    print(f"\nEl valor p es {valor_p_escolaridad}, por lo tanto, no podemos rechaza
    print("\nNo hay suficiente evidencia para afirmar que la escolaridad tenga un e

# Define esc_range, por ejemplo, de 0 a 20 con cierto paso
esc_range = np.arange(0, 30, 1)

# Graficar la regresión
plt.xlabel('Años de Escolaridad')
plt.ylabel('Monto de la Jubilación')
plt.title('Regresión Lineal con Interacción Sexo-Escolaridad')

# Añadir la línea de regresión para hombres
interaccion_sexo_hombres = 0  # Si el sexo es 0 (hombres), la interacción es 0
y_pred_hombres = modelo.params['constante'] + modelo.params['esc'] * esc_range + ir
plt.plot(esc_range, y_pred_hombres, color='blue', label='Hombres')

# Añadir la línea de regresión para mujeres
interaccion_sexo_mujeres = 1  # Si el sexo es 1 (mujeres), la interacción es 1
y_pred_mujeres = modelo.params['constante'] + modelo.params['esc'] * esc_range + ir
plt.plot(esc_range, y_pred_mujeres, color='red', label='Mujeres')

plt.legend()
plt.show()
```

```
sexo_dic
1    104859
0     93513
Name: count, dtype: int64
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 y2803c   R-squared:                       0.299
Model:                            OLS   Adj. R-squared:                  0.298
Method:                 Least Squares   F-statistic:                     1043.
Date:                Mon, 08 Jan 2024   Prob (F-statistic):               0.00
Time:                        13:49:14   Log-Likelihood:            -1.0219e+05
No. Observations:                7349   AIC:                         2.044e+05
Df Residuals:                    7345   BIC:                         2.044e+05
Df Model:                           3
Covariance Type:            nonrobust
================================================================================
======
                           coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
------
constante             1.071e+05   9663.298     11.088      0.000    8.82e+04    1.
26e+05
esc                    3.64e+04    875.044     41.599      0.000    3.47e+04    3.
81e+04
sexo_dic              4246.0585   1.38e+04      0.309      0.758   -2.27e+04    3.
12e+04
interaccion_sexo_esc -2.155e+04   1252.133    -17.211      0.000    -2.4e+04   -1.
91e+04
==============================================================================
Omnibus:                     4249.656   Durbin-Watson:                   1.857
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            62278.242
Skew:                           2.483   Prob(JB):                         0.00
Kurtosis:                      16.369   Cond. No.                         64.6
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly s
pecified.
```

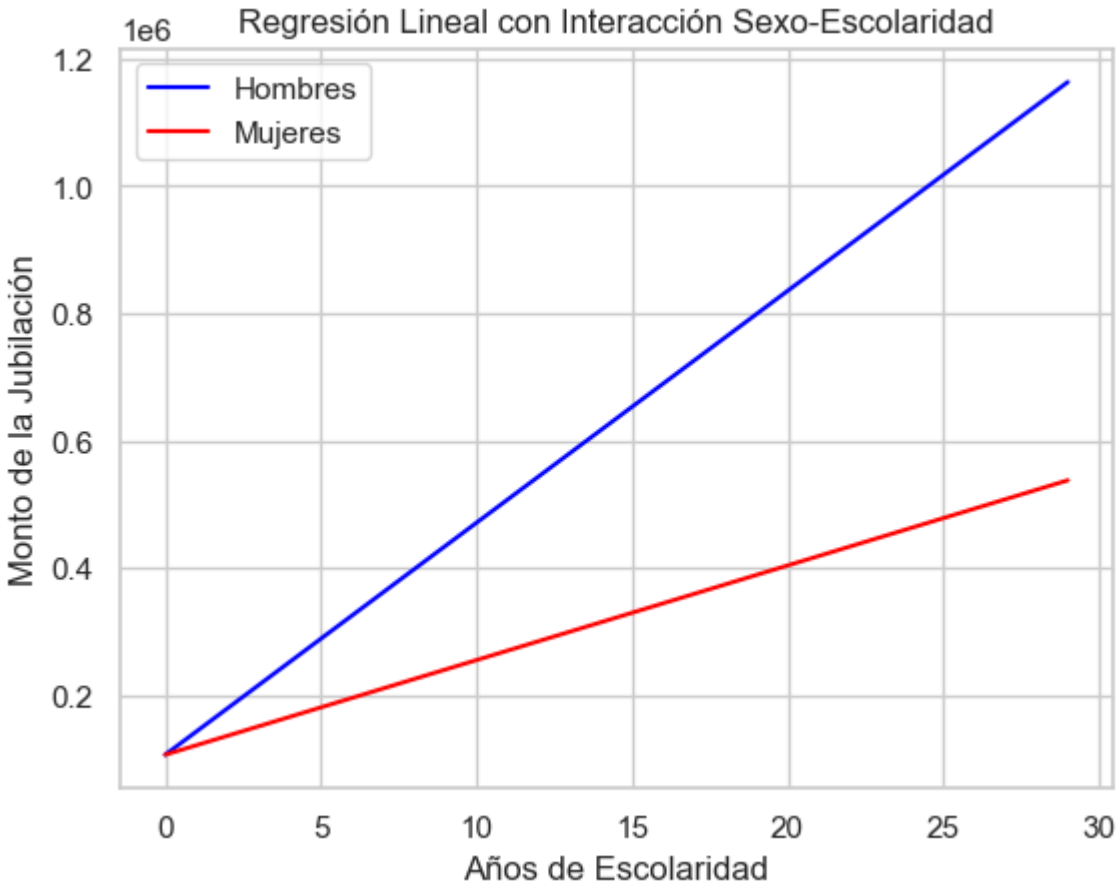
Prueba de hipótesis:

H0: La escolaridad no tiene efecto significativo en el monto a pagar de la jubilac
ión de la AFP
H1: La escolaridad tiene un efecto significativo en el monto a pagar de la jubilac
ión de la AFP

El valor p es 0.0, por lo tanto, rechazamos la hipótesis nula.

La escolaridad tiene un efecto significativo en el monto a pagar de la jubilación
de la AFP.

```
C:\Users\usuario\AppData\Local\Temp\ipykernel_8764\2659305246.py:14: SettingWithCo
pyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stabl
e/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_clear['constante'] = 1
```



Interpretación del modelo

En consideración a los coeficientes:

La constante (107100): Representa el valor esperado del monto de la jubilación cuando todas las variables independientes son cero. En este caso no tiene significado su valor.

Escolaridad (36400): Indica que por cada año más de escolaridad el monto de la pensión aumenta en 36.400 pesos mensuales. manteniendo constante el efecto de otras variables en el modelo.

Sexo (4246.0585): Representa el efecto adicional en el monto de la jubilación cuando la variable sexo es 1 (mujer) en comparación con la variable es 0 (hombre). En este caso las mujeres, en promedio, tendrían un monto de jubilación mayor, en comparación a los hombres a igual nivel de educación. El sueldo para ellas aumenta en 4.246 pesos. Lo demás se mantiene constante

Interacción sexo y escolaridad (-21550): Pero cuando la variable sexo interacciona con la variable educación el resultado es lo opuesto. Dado que es negativo (21.550), se sugiere que la relación entre la escolaridad y el monto de la jubilación es menor para las mujeres en comparación a los hombres a igual nivel de escolaridad. Un año más de educación para las mujures significa 14.492 pesos de jubilación mensual.

En resumen, el nivel de escolaridad está positivamente relacionado con el monto de la jubilación. Además, hay un efecto adicional positivo en el monto de la jubilación para las mujeres en comparación con los hombres.

Sin embargo, el resultado de la la interacción negativa sugiere que el impacto de la escolaridad es menor para las mujeres en comparación con los hombres. Lo que se puede explicar en que en la variable sexo las mujeres contienen algunos valores más altos que distorsionan el resultado del coeficiente.

Quinto modelo:

En este caso se elimina la variable sexo en el modelo de regresión.

Modelo según la población:

$$\text{Monto de la jubilación AFP}_i = \beta_0 + \beta_1 educ + (\beta_1 educ * \beta_2 se$$

Datos enriquecidos

In [70]:
```
# Columna a estudiar
frecuencia_sex_dic = df['sexo_dic'].value_counts()

# Mostrar la frecuencia de la variable
print(frecuencia_sex_dic,"\n")

# Crear una variable de interacción entre 'esc' y 'sexo'
df['interaccion_sexo_esc'] = df['esc'] * df['sexo_dic']

# Eliminar filas con valores NaN
df_clear = df.dropna(subset=['y2803c', 'esc', 'sexo_dic', 'interaccion_sexo_esc'])

# Agregar una columna de constantes para el término constante en el modelo
df_clear['constante'] = 1


# Ajustar el modelo de regresión lineal con la variable de interacción
modelo_1 = sm.OLS(df_clear['y2803c'], sm.add_constant(df_clear[['constante','esc',

# Obtener los resultados de la regresión
resultados = modelo_1.summary()
print(resultados)

# Prueba de hipótesis sobre la variable de escolaridad
print("\nPrueba de hipótesis:")
print("\nH0: La escolaridad no tiene efecto significativo en el monto a pagar de la
print("H1: La escolaridad tiene un efecto significativo en el monto a pagar de la j

# Definimos nuestra hipótesis nula
# El valor p está asociado con la variable escolaridad (coeficiente)
valor_p_escolaridad = modelo_1.pvalues[1]

# Nivel de significancia (usualmente 0.05)
nivel_significancia = 0.05

if valor_p_escolaridad < nivel_significancia:
    print(f"\nEl valor p es {valor_p_escolaridad}, por lo tanto, rechazamos la hipó
    print("\nLa escolaridad tiene un efecto significativo en el monto a pagar de la
else:
    print(f"\nEl valor p es {valor_p_escolaridad}, por lo tanto, no podemos rechaza
    print("\nNo hay suficiente evidencia para afirmar que la escolaridad tenga un e
```

```
# Define esc_range, por ejemplo, de 0 a 20 con cierto paso
esc_range = np.arange(0, 30, 1)

# Graficar la regresión
plt.xlabel('Años de Escolaridad')
plt.ylabel('Monto de la Jubilación')
plt.title('Regresión Lineal con Interacción Sexo-Escolaridad')

# Añadir la línea de regresión para hombres
interaccion_sexo_hombres = 0  # Si el sexo es 0 (hombres), la interacción es 0
y_pred_hombres = modelo_1.params['constante'] + modelo_1.params['esc'] * esc_range
plt.plot(esc_range, y_pred_hombres, color='blue', label='Hombres')

# Añadir la línea de regresión para mujeres
interaccion_sexo_mujeres = 1  # Si el sexo es 1 (mujeres), la interacción es 1
y_pred_mujeres = modelo_1.params['constante'] + modelo_1.params['esc'] * esc_range
plt.plot(esc_range, y_pred_mujeres, color='red', label='Mujeres')

plt.legend()
plt.show()
```

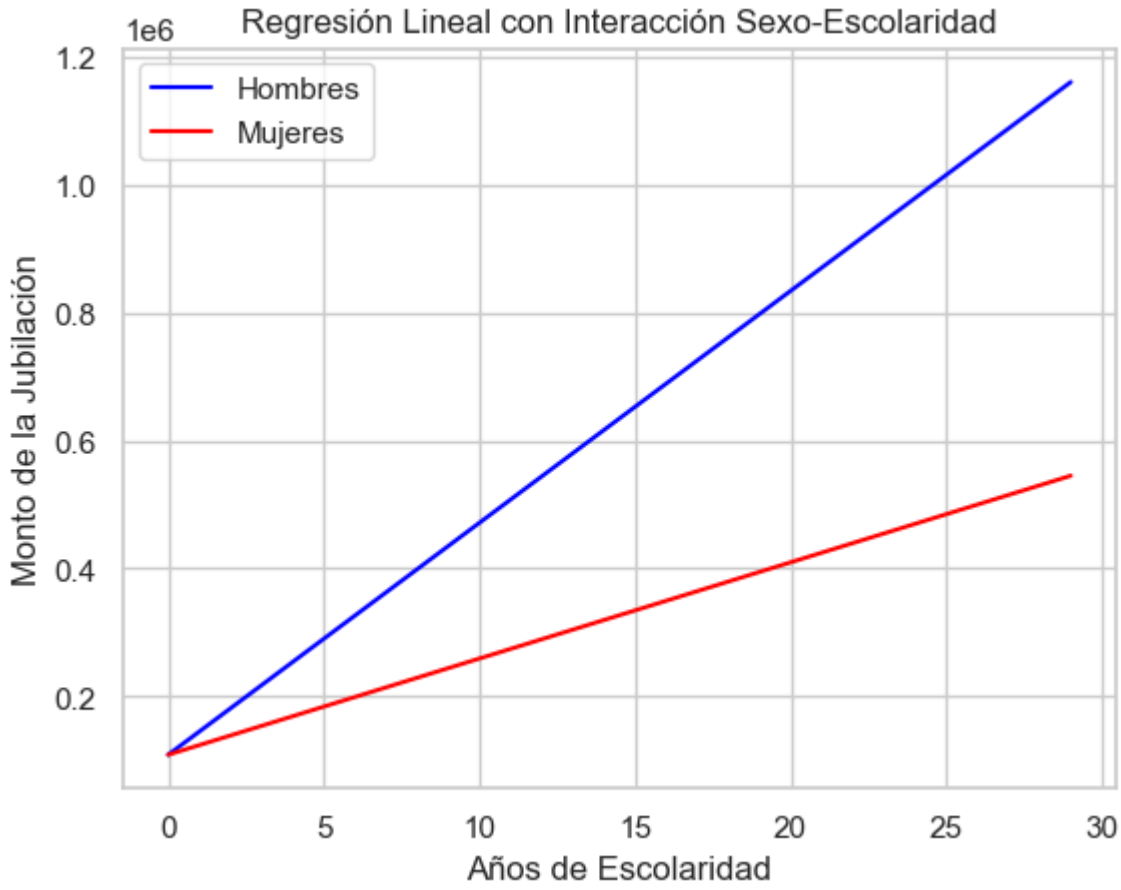
```
sexo_dic
1    104859
0     93513
Name: count, dtype: int64
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  y2803c   R-squared:                       0.299
Model:                             OLS   Adj. R-squared:                  0.299
Method:                  Least Squares   F-statistic:                     1565.
Date:                 Mon, 08 Jan 2024   Prob (F-statistic):               0.00
Time:                         13:49:19   Log-Likelihood:            -1.0219e+05
No. Observations:                7349   AIC:                         2.044e+05
Df Residuals:                    7346   BIC:                         2.044e+05
Df Model:                           2
Covariance Type:            nonrobust
================================================================================
======
                           coef    std err          t      P>|t|      [0.025
0.975]
--------------------------------------------------------------------------------
------
constante             1.092e+05   6875.638     15.888      0.000    9.58e+04    1.
23e+05
esc                   3.623e+04    683.618     53.001      0.000    3.49e+04    3.
76e+04
interaccion_sexo_esc -2.12e+04    562.336    -37.708      0.000   -2.23e+04   -2.
01e+04
==============================================================================
Omnibus:                     4251.029   Durbin-Watson:                   1.857
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            62350.383
Skew:                           2.484   Prob(JB):                         0.00
Kurtosis:                      16.377   Cond. No.                         28.3
==============================================================================
```

```
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly s
pecified.
```

```
Prueba de hipótesis:

H0: La escolaridad no tiene efecto significativo en el monto a pagar de la jubilac
ión de la AFP
H1: La escolaridad tiene un efecto significativo en el monto a pagar de la jubilac
ión de la AFP

El valor p es 0.0, por lo tanto, rechazamos la hipótesis nula.

La escolaridad tiene un efecto significativo en el monto a pagar de la jubilación
de la AFP.
```

```
C:\Users\usuario\AppData\Local\Temp\ipykernel_8764\1181799110.py:14: SettingWithCo
pyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stabl
e/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_clear['constante'] = 1
```



Interpretación del modelo

Constante (1.092e+05): Es la estimación del monto de la jubilación cuando todas las demás variables son cero. En este caso no tiene sentido interpretarla.

Escolaridad (3.623e+04): En este caso, un aumento de un año en la escolaridad se asocia con un aumento de 36.230 pesos en el monto de la jubilación a pagar.Lo demás se mantiene constante

Interacción entre escolaridad y sexo (-2.12e+04): Este coeficiente representa cómo la relación entre la escolaridad y el monto de jubilación difiere para diferentes valores de la variable sexo. En este caso, un aumento en la interacción se asocia con una disminución de 21.200 pesos en el monto de la jubilación pagada por la AFP mensualmente a las mujeres en comparación a los hombres a igual nivel educacional. Lo demás se mantiene constante.

Sexto modelo:

En este modelo a la variable educación se le agrega la variable dicotómica área (rural, urbano). Variable que estará en interacción con la escolaridad dentro del módelo de regresión multiple.

Modelo según la población:

$$\text{Monto de la jubilación AFP}_i = \beta_0 + \beta_1 educ + (\beta_1 educ * \beta_2 \text{á}$$

Datos enriquecidos

In [71]:
```
# Columna a estudiar
frecuencia_area_dic = df['area_dic'].value_counts()

# Mostrar la frecuencia de la variable
print(frecuencia_area_dic,"\n")

# Crear una variable de interacción entre 'esc' y 'sexo'
df['interaccion_area_esc'] = df['esc'] * df['area_dic']

# Eliminar filas con valores NaN
df_clear = df.dropna(subset=['y2803c', 'esc', 'area_dic', 'interaccion_area_esc'])

# Agregar una columna de constantes para el término constante en el modelo
df_clear['constante'] = 1


# Ajustar el modelo de regresión lineal con la variable de interacción
modelo_area = sm.OLS(df_clear['y2803c'], sm.add_constant(df_clear[['constante','esc

# Obtener los resultados de la regresión
resultados = modelo_area.summary()
print(resultados)

# Prueba de hipótesis sobre la variable de escolaridad
print("\nPrueba de hipótesis:")
print("\nH0: La escolaridad no tiene efecto significativo en el monto a pagar de la
print("H1: La escolaridad tiene un efecto significativo en el monto a pagar de la j

# Definimos nuestra hipótesis nula
# El valor p está asociado con la variable escolaridad (coeficiente)
valor_p_escolaridad = modelo_area.pvalues[1]

# Nivel de significancia (usualmente 0.05)
nivel_significancia = 0.05

if valor_p_escolaridad < nivel_significancia:
    print(f"\nEl valor p es {valor_p_escolaridad}, por lo tanto, rechazamos la hipó
    print("\nLa escolaridad tiene un efecto significativo en el monto a pagar de la
else:
    print(f"\nEl valor p es {valor_p_escolaridad}, por lo tanto, no podemos rechaza
    print("\nNo hay suficiente evidencia para afirmar que la escolaridad tenga un e

# Define esc_range, por ejemplo, de 0 a 20 con cierto paso
esc_range = np.arange(0, 31, 1)

# Graficar la regresión
plt.xlabel('Años de Escolaridad')
plt.ylabel('Monto de la Jubilación')
plt.title('Regresión Lineal con Interacción Localidad-Escolaridad')

# Añadir la línea de regresión para urbano
interaccion_area_urbana = 0  # Si el area es 0 (urbana), la interacción es 0
y_pred_hombres = modelo_area.params['constante'] + modelo_area.params['esc'] * esc_
plt.plot(esc_range, y_pred_hombres, color='blue', label='Urbano')

# Añadir la línea de regresión para rural
interaccion_area_rural = 1  # Si el area es 1 (rural), la interacción es 1
y_pred_mujeres = modelo_area.params['constante'] + modelo_area.params['esc'] * esc_
plt.plot(esc_range, y_pred_mujeres, color='red', label='Rural')
```

```
plt.legend()
plt.show()
```

```
area_dic
0    158319
1     40053
Name: count, dtype: int64
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  y2803c   R-squared:                       0.164
Model:                             OLS   Adj. R-squared:                  0.163
Method:                  Least Squares   F-statistic:                     478.8
Date:                 Mon, 08 Jan 2024   Prob (F-statistic):          3.78e-284
Time:                         13:49:25   Log-Likelihood:            -1.0283e+05
No. Observations:                7349   AIC:                         2.057e+05
Df Residuals:                    7345   BIC:                         2.057e+05
Df Model:                           3
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
constante     9.44e+04   8966.435     10.529      0.000    7.68e+04    1.12e+05
esc           2.642e+04    787.059     33.570      0.000    2.49e+04    2.8e+04
area_dic      2.769e+04   1.66e+04      1.667      0.096   -4872.392    6.02e+04
interaccion_area_esc -1126.4279   1728.142     -0.652      0.515   -4514.081    2261.226
==============================================================================
Omnibus:                     4143.351   Durbin-Watson:                   1.872
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            50032.503
Skew:                           2.472   Prob(JB):                         0.00
Kurtosis:                      14.788   Cond. No.                         57.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

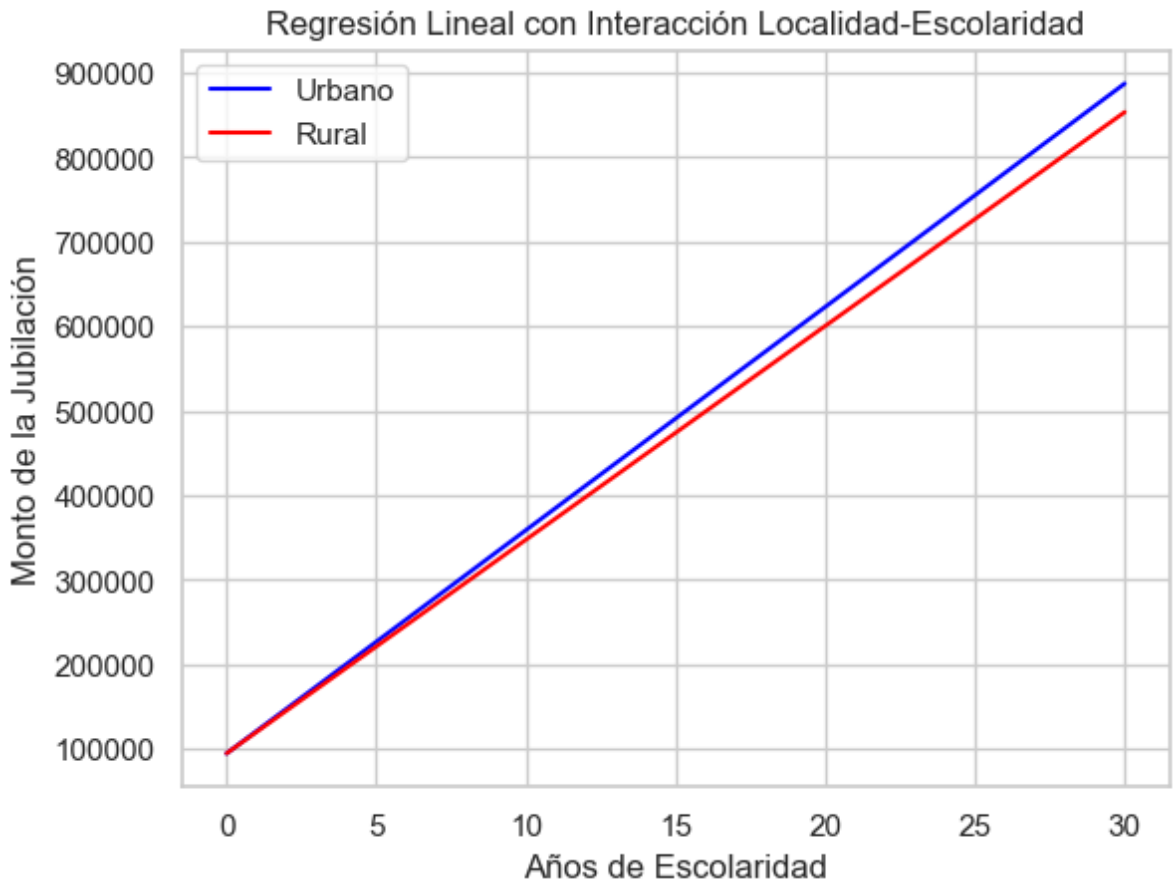
Prueba de hipótesis:

H0: La escolaridad no tiene efecto significativo en el monto a pagar de la jubilación de la AFP
H1: La escolaridad tiene un efecto significativo en el monto a pagar de la jubilación de la AFP

El valor p es 5.596957229219431e-230, por lo tanto, rechazamos la hipótesis nula.

La escolaridad tiene un efecto significativo en el monto a pagar de la jubilación de la AFP.

```
C:\Users\usuario\AppData\Local\Temp\ipykernel_8764\2603975039.py:14: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df_clear['constante'] = 1
```



Interpretación del modelo

Constante (94.400): Es la estimación del monto de la jubilación cuando todas las demás variables son cero (escolaridad, área, interacción). En este caso no es interpretable.

Escolaridad (26.420): Indica que un año extra de educación formal se asocia con un aumento de 26.420 pesos en el monto de la jubilación. Lo demás se mantiene constante.

Area_(27.690): Representa el cambio esperado en el monto de la jubilación al comparar el área rural con el área urbana. Las personas jubiladas de las zonas rurales ganan 27.690 pesos más en el monto de la jubilación en comparación a las personas de las zonas urbanas a igual nivel de estudios. Lo demás se mantiene constante.

Interacción área y escolaridad (-1126): Representa cómo la relación entre la escolaridad y la zona de residencia del jubilado difiere. En este caso, un año más de educación para los que viven en la zona rural significa la disminución de su monto de la jubilacion en 1.126 pesos en comparación a los que viven en zonas urbanas a igual nivel educacional. Lo demás se mantiene constante.

La probabilidad asociada al estadístico F es extremadamente baja, lo que sugiere que al menos una de las variables explicativas en el modelo es significativamente diferente de cero.

Comparación de Modelos

In [72]:
```
stargazer = Stargazer([simple_agrup_region,simple_agrup_reg_area, simple_totaldatos
# Configurar el título con dos líneas
stargazer.title("\nComparación de Modelos: Monto de la jubilación de la AFP.\n(valo
stargazer.custom_columns(["Agrup_Reg", 'Agrup_Reg_y_area', 'No agrup_lineal', 'Inte
```

```
stargazer.add_custom_notes(["Notas personalizadas para la tabla."])  # Notas persor
stargazer
```

Out[72]: Comparación de Modelos: Monto de la jubilación de la AFP. (valores expresados en pesos chilenc

					Depe
	Agrup_Reg	Agrup_Reg_y_area	No agrup_lineal	Interacción:Esc_sexo	Interacción_Esc
	(1)	(2)	(3)	(4)	(5)
area_dic					
constante	-121751.586	63970.854	103599.456***	107142.437***	109238.613***
	(117316.652)	(59624.123)	(7509.355)	(9663.298)	(6875.638)
esc	42258.636***	25389.159***	25862.130***	36401.058***	36232.435***
	(10615.321)	(5640.393)	(683.705)	(875.044)	(683.618)
interaccion_area_esc					
interaccion_sexo_esc				-21549.897***	-21204.507***
				(1252.133)	(562.336)
sexo_dic				4246.059	
				(13753.222)	
Observations	16	32	7349	7349	7349
R^2	0.531	0.403	0.163	0.299	0.299
Adjusted R^2	0.497	0.383	0.163	0.298	0.299
Residual Std. Error	26037.141 (df=14)	35303.418 (df=30)	289070.643 (df=7347)	264629.527 (df=7345)	264613.232 (df=7346)
F Statistic	15.848*** (df=1; 14)	20.262*** (df=1; 30)	1430.839*** (df=1; 7347)	1043.052*** (df=3; 7345)	1564.723*** (df=2; 7346)
Note:					*p<(
					Notas perso

Evaluación y Conclusiones

Tanto la revisión de los resultados propios de cada regresión como la revisión de la literatura y los informes de la Superintendencia de Pensiones permiten señalar que el modelo de la regresión cinco es el mejor, al considerar la diferencia significativa del monto de autofinancimiento de la pensión que se da entre hombres y mujeres es significativa, antecedente entregado por la superintendencia.

La interacción de la variable sexo y educación muestra que la mujer tiene desventaja en comparación al hombre aun cuando tengan el mismo nivel de educación. Por otro lado, revisando y comparando los valores obtenidos de cada regresión: R^2 ajustado**: 0.299, R^2: 0.299, Prob (F-statistic): 0.00, el AIC: 204.2y el BIC: 204.4, en su conjunto son valores acordes con la decisión.

En este modelo, en relació a la escolaridad, se puede decir que un año de educación se asocia con un aumento de 36.230 pesos en el monto de la jubilación a pagar.Lo demás se mantiene constante. En tanto, la relación entre la escolaridad y el monto de jubilación difiere para diferentes valores de la variable sexo. Un aumento en la interacción se asocia con una disminución de 21.200 pesos en el monto de la jubilación pagada por la AFP mensualmente a las mujeres en comparación a los hombres a igual nivel educacional. Lo demás se mantiene constante. Es decir, el monto de 36.230 es para los hombres, pero para las mujeres es de 14.630 pesos.

Sin embargo, se puede considerar la omisión de variables relevantes para el análisis de este fenómeno. Por un lado se debe considerar variables propias de la vida laboral y previsional que tuvo el trabajador, tales como: años de cotizaciones, periodos sin cotización previsional, saber si fue un trabajador dependiente o independiente, saber su monto imponible, saber el pago no imponible de su sueldo, dato no menor porque se sabe que las empresas tratan de bajar costos y el ciertos pagos no son imponibles y la empresa las usa para ahorrar a todo evento, saber si ahorro dinero en la modalidad APV, ahorro previsional voluntario, edad de jubilación, saber si la empresa entregó algún bono para que el trabajador jubilará y no siguiera trabajando.

De otro lado, se puede considerar variables externas asociadas a la dimensión del mercado laboral, tales como: la estabilidad ofrecida en los trabajos, es decir, qué tipo de contrato es más ofrecido por la empresa; indefinido, fijo, por faena u obra, conocer el sueldo imponible promedio del mercado laboral, saber el impacto de la flexibilización laboral en las cotizaciones de los trabajadores, variables asociadas a la tercerización de los trabajadores, es decir, saber el impacto que tiene todo cambio en el mercado laboral.

En la revisión de la base de datos y sus variables, trabajo no menor, se constata y se ejercita que se debe tener clara la direccionalidad de la variable, es decir, en algún momento, se hizo la regresión lineal usando el estrato socioeconómico corregido, pero se determinó que no está claro si esta variable describe a la pensión o al revés si el monto de la pensión describe el estrato de la persona.

Lo otro, ver mapa de calor, la relación que se da entre otros ingresos asociados al patrimonio y el monto de la pensión, es decir, refuerza la idea de que a mayor patrimonio mejor la pensión, pero de igual modo es un dato muy aislado, pero interesante de abordar, porque como se piensa, para una persona sería mejor imponer por el mínimo y destinar ese dinero en adquirir patrimonio para tener una mejor vejez.

Expuesto lo anterior, se puede señalar que el resultado de la regresión solo aportaría al debate, pero por si sola no permite tomar alguna decisión en relación a aplicar alguna política pública.
Solo señalar que la educación es el pilar de todo país que quiere mejorar y salir de la pobreza, sin educación difícil se hace superar las dificultades presentes en la vida de la

persona. Lo anterior describe tanto las fortalezas y debilidades de este trabajo, y en esa misma senda deja abierto el camino a otros trabajos futuros que mezclen las dimensiones anteriores: Mercado laboral, trayectoria laboral y previsional, caracterización del trabajador para determinar el peso de cada uno y así determinar donde se debe hacer hincapié a un fenómeno problemático que los gobiernos de turno deben sortear de la mejor manera.

Bibliografía

Stuardo, N., Zavala, M., Merino, J., . (2016). La decisión de jubilar. Dificultades y factores asociados en personal universitario de Concepción. Revista Perspectivas, 28, pp. 79-107.

Salina Galvez, G. (2010). La paradoja de la vejez en Nuestro sistema económico (Seminario de Titulo, Universidad de Chile, Chile). Recuperado de https://repositorio.uchile.cl/handle/2250/108012

Asociacion de AFP Chile. (2023). Recuperado de https://www.aafp.cl/

Superintendencia de Pensiones. SP (2023) Recuperado de https://www.spensiones.cl/portal/institucional/594/w3-channel.html

Espinoza, C. (21 de Noviembre de 2014). Investigadores buscan resolver misterio del ejército de terracota. La Tercera. Recuperado de http://www.latercera.com/noticia/tendencias/2014/11/659-605461-9-investigadores-buscan-resolver-misterio-del-ejercito-de-terracota.shtml

Cainupán, K. (5 de marzo de 2013). Los montos que se deben tener en la AFP para una pensión sobre $ 500.000. La Tercera. Recuperado de https://www.latercera.com/noticia/los-montos-que-se-deben-tener-en-la-afp-para-una-pension-sobre-500-000/

Huneuus, C. (30 de octubre de 2018). La crisis del sistema privado de pensiones: un punto de inflexión de nuestra democracia. Ciper. Recuperado de https://www.ciperchile.cl/2018/10/30/la-crisis-del-sistema-privado-de-pensiones-un-punto-de-inflexion-de-nuestra-democracia/