Unidad 1. Regresión lineal simple y correlación

Medidas de dispersión

Suma de los cuadrados de las desviaciones de los valores de X con respecto a su media:

$$S_{xx}=\sum_{i=1}^{n}\left(X_i-\acute{X}\right)^2$$

Suma de los productos de las desviaciones de los valores de X y Y con respecto a sus medias:

$$S_{xy}=\sum_{i=1}^{n}\left(X_i-\acute{X}\right)\left(Y_i-\acute{Y}\right)$$

Suma de los cuadrados de las desviaciones de los valores de Y con respecto a su media:

$$S_{xx}=\sum_{i=1}^{n}\left(Y_i-\acute{Y}\right)^2$$

Coeficiente de correlación y de determinación

Coeficiente de correlación de Pearson:

$$r=\frac{S_{xy}}{\sqrt{S_{xx}\cdot S_{yy}}}$$

Coeficiente de determinación:

$$r^2$$

Recta de regresión ajustada

La regresión lineal ajustada se representa mediante estadísticos:

$$\hat{Y}=b_0+b_1X$$

donde $\hat{Y}$ representa el valor de Y obtenido mediante la recta de regresión ajustada (no la verdadera Y). Los estadísticos b_0 y b_1 se calculan de la siguiente manera:

$$b_1=\frac{S_{xy}}{S_{xx}}$$

$$b_0=\acute{Y}-b_1\acute{X}$$

Cálculo de residuales

Residuales:

$$e_i = Y_i - \hat{Y}_i$$

Sumas de cuadrados SS (Sum of Squares)

Suma de los Cuadrados de los Errores:

$$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2 = \sum_{i=1}^{n} e_i^2$$

Suma total de cuadrados:

$$SST = \sum_{i=1}^{n} \left(Y_i - \acute{Y} \right)^2$$

Suma de cuadrados de regresión:

$$SSR = \sum_{i=1}^{n} \left(\hat{Y}_i - \acute{Y} \right)^2$$

- SST: Mide la variabilidad total de los datos observados.
- SSR: Mide la variabilidad de los datos que el modelo de regresión explica.
- SSE: Mide la variabilidad no explicada por el modelo (es decir, los residuos).

Intervalo de confianza

Estadístico de prueba t:

$$t = \frac{b_1}{SE(b_1)}$$

Error estándar de b_1:

$$SE(b_1) = \frac{\sqrt{SSE/(n-2)}}{\sqrt{S_{xx}}}$$

Intervalo de confianza para b_1:

$$b_1 - t_{\alpha/2} \cdot SE(b_1) < \beta_1 < b_1 + t_{\alpha/2} \cdot SE(b_1)$$

donde n representa la cantidad de pares de datos.

Comprobación de supuestos

Comprobar suposiciones:

- Test de shapiro a los residuales e_i: Para comprobar si la distribución es normal sobre la recta.

- Grafico X vs Y: Para observar si los datos soportan la suposición de linealidad.
- Gráfico de residuales: Para observar si los datos soportan la suposición de linealidad, complementario al coeficiente de correlación
- Test de Breusch-Pagan: Para detectar heteroscedasticidad en regresion lineal

Test de Shapiro: `from scipy.stats import shapiro` Después, se obtiene el valor-p: `_, valor_p_sh = shapiro(data)`

- H_0: Los datos siguen una distribución normal
- H_1: Los datos no siguen una distribución normal

Test de Breusch-Pagan: `from statsmodels.stats.api import het_breuschpagan` Después, se obtiene el valor-p: `_, valor_p_bp, _, _ = het_breuschpagan(residuales, X)`

- H_0: Hay homoscedasticidad
- H_1: Hay heteroscedasticidad

ANOVA en regresión lineal

Fuente de variación	Suma de cuadrados (SS)	Grados de libertad (df)	Promedio de los cuadrados (MS)	Estadístico F
Regresión	$ SSR$	p	$MSR=\frac{SSR}{p}$	$F=\frac{MSR}{MSE}$
Error	SSE	$n - p - 1 $	$MSE = \frac{SSE}{n - p - 1} $	
Total	$SST $	$n-1$		

donde p es el número de parámetros para la recta de regresión ajustada (en la regresión simple $p=1$). Las hipótesis son:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

Problemario de la Unidad 1

Problema 1

Un profesor intenta mostrar a sus estudiantes la importancia de los exámenes cortos, aun cuando el 90% de la calificación final esté determinada por los exámenes parciales. Él cree que cuanto más altas sean las calificaciones de los exámenes cortos, más alta será la calificación final. Seleccionó una muestra aleatoria de 15 estudiantes de su clase con los siguientes datos:

Promedio de exámenes cortos	Promedio final
5	
9	64
92	

Promedio de exámenes cortos	Promedio final
84	
72	77
90	80
95	77
87	81
89	80
77	84
76	80
65	69
97	83
42	40
94	78
62	65
91	90

1. Establesca una variable dependiente (Y) y una variable independiente (X). 2 . Realice un diagrama de dispersión para estos datos.
2. ¿Los datos soportan la suposición de linealidad?
3. Calcule el coeficiente de correlación e interprete el resultado.
4. Calcule el coeficiente de determinación e interprete el resultado.
5. Obtenga la recta de regresión ajustada y grafíquelo sobre el gráfico de dispersión.
6. Obtenga un intervalo de confianza del 95% para la pendiente de la recta de regresión ajustada (b_1)
7. Calcule los residuales y trace un nuevo gráfico de dispersión. Comente, ¿Parece que se verifican los supuestos?
8. Realice la prueba de Shapiro para los residuales y comente el resultado.
9. Realice la prueba de Brausch-Pagan para los residuales y comente el resultado.
10. Tres estudiantes sacaron 70, 75 y 84 de calificación. Según la recta de regresión ajustada, ¿cuáles son los resultados esperados para estos tres alumnos?
11. Realice una tabla ANOVA e interprete el resultado.

```
# 1. Establezca una variable dependiente ( Y ) y un variable
#    independiente ( X ).
# Varible independiente: examenes cortos
# Variable dependiente: examen final
import numpy as np
X = np.array([59, 92, 72, 90, 95, 87, 89, 77, 76, 65, 97, 42, 94, 62,
91])
Y = np.array([64, 84, 77, 80, 77, 81, 80, 84, 80, 69, 83, 40, 78, 65,
90])
# 2. Realice un diagrama de dispersión para estos datos.
import matplotlib.pyplot as plt
plt.scatter(X, Y, color = 'pink')
```

```python
plt.xlabel('Examenes cortos')
plt.ylabel('Examen final')
plt.title('Grafico de dispersión')
plt.grid()


# 3. ¿ Loa dato soportan la suposicion de linealidad?
# 4. Calcule el coeficiente de correlacion e interprete el resultado.
SXX = np.sum((X - np.mean(X))**2)
SXY = np.sum((X - np.mean(X))*(Y - np.mean(Y)))
SYY = np.sum((Y - np.mean(Y))**2)
r = SXY / np.sqrt(SXX * np.sum((Y - np.mean(Y))**2))
print("Coeficiente de correlacion: ", r)

# 5. Calcule el coeficiente de determinación e interprete el
resultado.
print('Coeficiente de determinación:', r ** 2)
# 6. Obtenga la recta de regreción ajustada y grafíquenlo sobre el
grafico de disperción.
b1 = SXY / SXX
b0 = np.mean(Y) - b1 * np.mean(X)
print('Pendiente:', b1) #Corrected print statement
Yc = b0 + b1 * X

plt.plot(X, Yc, '--', color = 'pink')

# 7. Intervalo de confianza para b1
nivel_de_significancia = 0.05
from scipy.stats import t
t_value = t.ppf(1- nivel_de_significancia / 2, len(Y) - 2)
se_b1 = np.sqrt(np.sum((Y - Yc)**2) / (2 - 2)) / np.sqrt(Sxx)
confianza_b1 = (b1 - t_value * se_b1, b1 + t_value * se_b1)
print(f'Intervalo de confianza para b1: ', confianza_b1)
# 8. Calcule los residuales y trace un nuevo gráfico de dispersión.
Comente, ¿Parece que se verifican los supuestos?
residuales = Y - Yc
plt.figure()
plt.scatter(X, Residuales, color = 'pink')
plt.xlabel('examenes cortos')
plt.ylabel('residuales')

plt.title('Grafico de dispersión de los residuales')
plt.axhline(y=0,color = "pink", linestyle = "--")
# 9. Realice la prueba de Shapiro para los residuales y comente el
resultado.
from scipy.stats import shapiro
_, valor_p_sh = shapiro(residuales)
print('Valor-p de shapiro: ', valor_p_sh)
# 10. Realice la prueba de Brausch-Pagan para los residuales y comente
el resultado.
```

```python
# 11. Tres estudiantes sacaron 70, 75 y 84 de calificación. Según la recta de regresión ajustada, ¿cuáles son los resultados esperados para estos tres alumnos?
print(F"Yc para 70: {b1 * 70 + b0}")
print(F"Yc oara 75:{b1 * 75 + b0} ")
print(F"Yc para 84:{b1 * 84 + b0}")
# 12.Realice una tabla ANOVA e interprete el resultado.
print(F"SXX:{SXX}")
print(F"SXY:{SXY}")
print(F"SYY:{SYY}")
print(F"SSE:{np.sum(residuales**2)}")
print(F"SST:{np.sum((Y-np.mean(Y))**2)}")
print(F"SSR:{np.sum((Yc-np.mean(Y))**2)/(len(X)-2)}")
print(F"MSR:{np.sum((Yc-np.mean(Y))**2)/(len(X)-2)}")
print(F"MSE:{np.sum(residuales**2)/(len(X)-2)}")
print(F"F:{np.sum((Yc - 
np.mean(Y))**2)/1/(np.sum(residuales**2)/(len(Y)-2))}")
```

```
Coeficiente de correlacion:  0.8646014213752985
Coeficiente de determinación: 0.7475356178441864
Pendiente: 0.6431798623063684
Intervalo de confianza para b1:  (-inf, inf)
Valor-p de shapiro:  0.901827735700704
Yc para 70: 69.54941193344808
Yc oara 75:72.76531124497993
Yc para 84:78.55393000573724
SXX:3718.3999999999996
SXY:2391.6
SYY:2057.733333333333
SSE:519.5043746414227
SST:2057.733333333333
SSR:118.32530451476238
MSR:118.32530451476238
MSE:39.961874972417135
F:38.49241207409918
```

```
<ipython-input-29-285ad97e26f1>:39: RuntimeWarning: divide by zero encountered in scalar divide
  se_b1 = np.sqrt(np.sum((Y - Yc)**2) / (2 - 2)) / np.sqrt(Sxx)
```

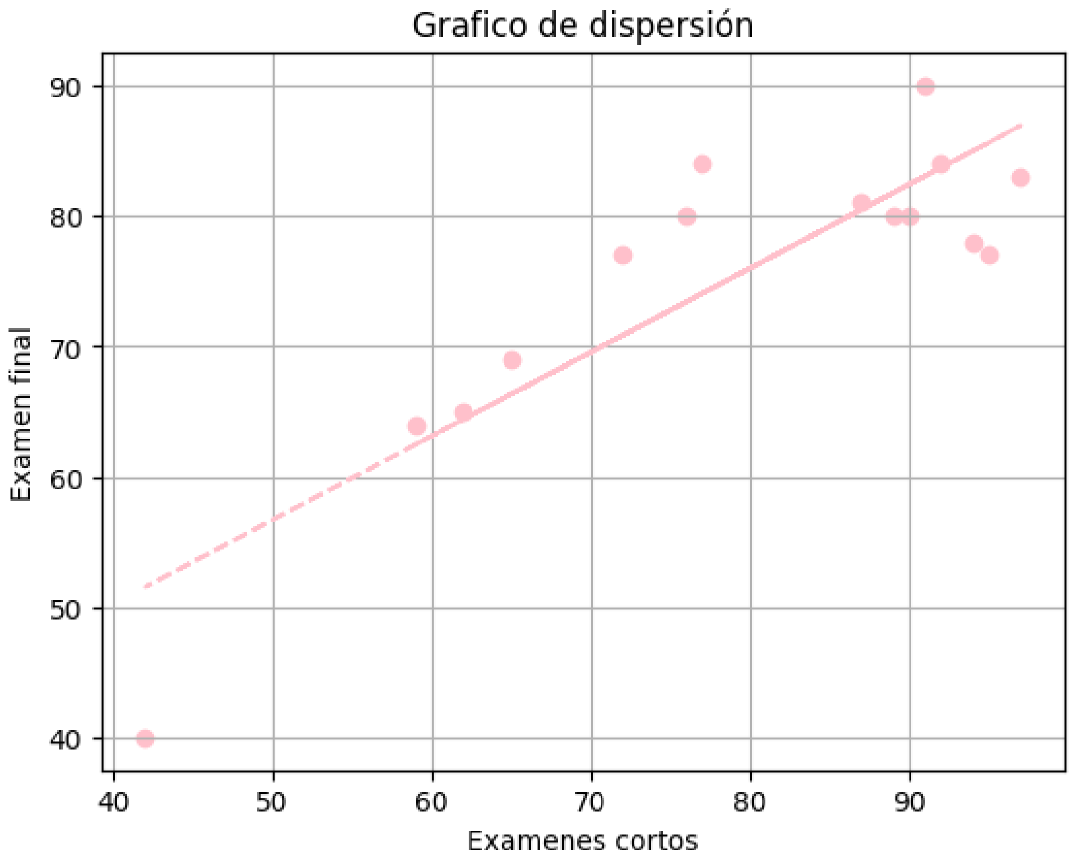
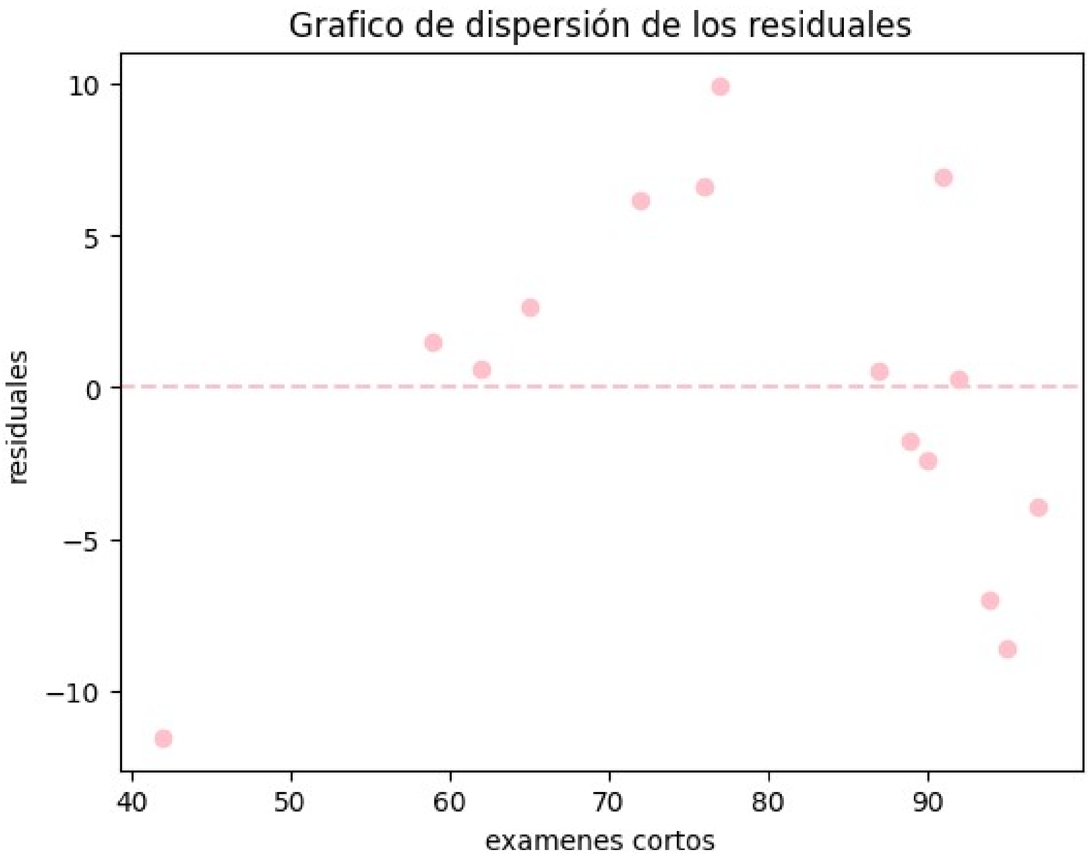


Grafico de dispersión
Examen final
Examenes cortos
90
80
70
60
50
40
40
50
60
70
80
90



Problema 2

William Hawkins, vicepresidente de personal de la International Motors, trabaja en la relación entre el salario de un trabajador y el porcentaje de ausentismo. Hawkins dividió el intervalo de salarios de International en 12 grados o niveles (1 es el menor grado, 12 el más alto) y después muestreó aleatoriamente a un grupo de trabajadores. Determinó el grado de salario de cada trabajador y el número de días que ese empleado había faltado en los últimos 3 años.

Categoría de salario	11	10	8	5	9	7	3
Ausencias	18	17	29	36	11	28	35

Categoría de salario	11	8	7	2	9	8	3
Ausencias	14	20	32	39	16	31	40

1. Establesca una variable dependiente (Y) y una variable independiente (X).
2. Realice un diagrama de dispersión para estos datos.

3. ¿Los datos soportan la suposición de linealidad?
4. Calcule el coeficiente de correlación e interprete el resultado.
5. Calcule el coeficiente de determinación e interprete el resultado.
6. Obtenga la recta de regresión ajustada y grafíquelo sobre el gráfico de dispersión.
7. Obtenga un intervalo de confianza del 95% para la pendiente de la recta de regresión ajustada (b_1)
8. Calcule los residuales y trace un nuevo gráfico de dispersión. Comente, ¿Parece que se verifican los supuestos?
9. Realice la prueba de Shapiro para los residuales y comente el resultado.
10. Realice la prueba de Brausch-Pagan para los residuales y comente el resultado.
11. Utiliza la recta de regresión para interpolar dos valores y extrapolar uno. Comenta estos resultados.
12. Realice una tabla ANOVA e interprete el resultado.

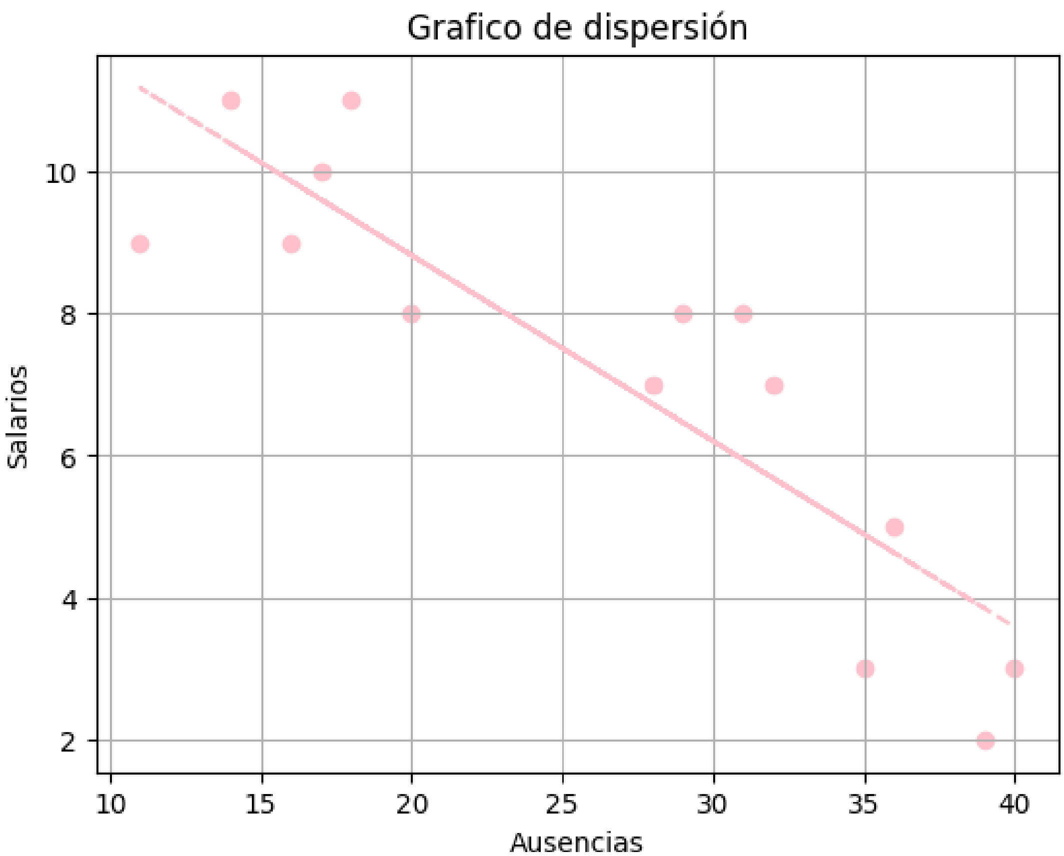
```
# 1. Establesca una variable dependiente ( Y ) y una variable
independiente ( X ).
# 1. Establezca una variable dependiente ( Y ) y un variable
#    independiente ( X ).
# Varible independiente: ausensias
# Variable dependiente: salarios
import numpy as np
Y = np.array([11, 10, 8, 5, 9, 7, 3, 11, 8, 7, 2, 9, 8, 3])
X = np.array([18, 17, 29, 36, 11, 28, 35, 14, 20, 32, 39, 16, 31, 40])
#. 2 Realice un diagrama de dispersión para estos datos.
import matplotlib.pyplot as plt
plt.scatter(X, Y, color = 'pink')
plt.xlabel('Ausencias')
plt.ylabel('Salarios')
plt.title('Grafico de dispersión')
plt.grid()
#. 3 ¿Los datos soportan la suposición de linealidad?
#. 4 Calcule el coeficiente de correlación e interprete el resultado.
SXX = np.sum((X - np.mean(X))**2)
SXY = np.sum((X - np.mean(X))*(Y - np.mean(Y)))
SYY = np.sum((Y - np.mean(Y))**2)
r = SXY / np.sqrt(SXX * np.sum((Y - np.mean(Y))**2))
print("Coeficiente de correlacion: ", r)
#. 5 Calcule el coeficiente de determinación e interprete el
resultado.
print('Coeficiente de determinación:', r ** 2)
#. 6 Obtenga la recta de regresión ajustada y grafíquelo sobre el
gráfico de dispersión.
b1 = SXY / SXX
b0 = np.mean(Y) - b1 * np.mean(X)
print('Pendiente:', b1) #Corrected print statement
Yc = b0 + b1 * X
plt.plot(X, Yc, '--', color = 'pink')
#. 7 Obtenga un intervalo de confianza del 95% para la pendiente de la
```

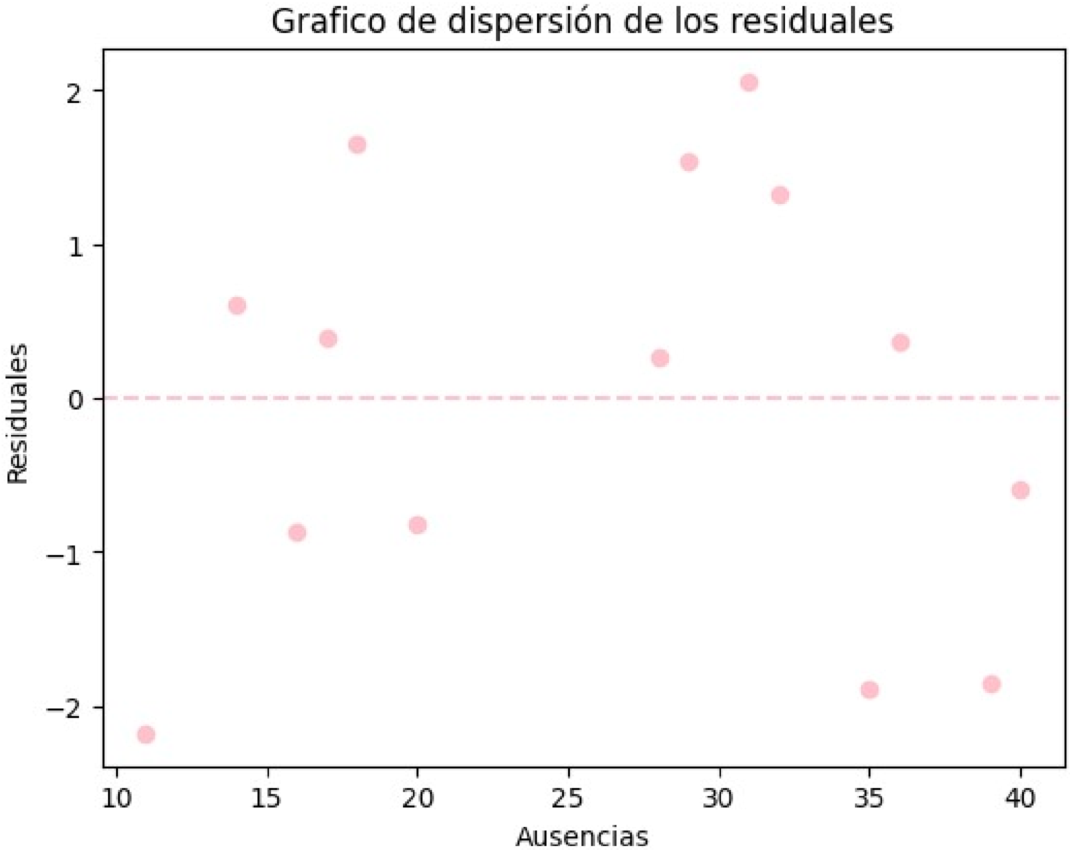
```python
recta de regresión ajustada (b1).
nivel_de_significancia = 0.05
from scipy.stats import t
t_value = t.ppf(1- nivel_de_significancia / 2, len(Y) - 2)
se_b1 = np.sqrt(np.sum((Y - Yc)**2) / (2 - 2)) / np.sqrt(SXX)
confianza_b1 = (b1 - t_value * se_b1, b1 + t_value * se_b1)
print(f'Intervalo de confianza para b1: ', confianza_b1)
#. 8 Calcule los residuales y trace un nuevo gráfico de dispersión.
Comente, ¿Parece que se verifican los supuestos?
residuales = Y - Yc
plt.figure()
plt.scatter(X, residuales, color = 'pink')
plt.xlabel('Ausencias')
plt.ylabel('Residuales')
plt.title('Grafico de dispersión de los residuales')
plt.axhline(y=0,color = "pink", linestyle = "--")
#. 9 Realice la prueba de Shapiro para los residuales y comente el
resultado.
from scipy.stats import shapiro
_, valor_p_sh = shapiro(residuales)
print('Valor-p de shapiro: ', valor_p_sh)
#. 10 Realice la prueba de Brausch-Pagan para los residuales y comente
el resultado.
#. 11 Utiliza la recta de regresión para interpolar dos valores y
extrapolar uno. Comenta estos resultados.
print(F"Yc para 18:{b1 * 18 + b0}")
print(F"Yc para 35:{b1 * 35 + b0}")
print(F"Yc para 40:{b1 * 40 + b0}")
#. 12 Realice una tabla ANOVA e interprete el resultado.
print(F"SXX:{SXX}")
print(F"SXY:{SXY}")
print(F"SYY:{SYY}")
print(F"SSE:{np.sum(residuales**2)}")
print(F"SST:{np.sum((Y-np.mean(Y))**2)}")
print(F"SSR:{np.sum((Yc-np.mean(Y))**2)/(len(X)-2)}")
print(F"MSR:{np.sum((Yc-np.mean(Y))**2)/(len(X)-2)}")
print(F"MSE:{np.sum(residuales**2)/(len(X)-2)}")
print(F"F:{np.sum((Yc -
np.mean(Y))**2)/1/(np.sum(residuales**2)/(len(Y)-2))}")
```

```
Coeficiente de correlacion:  -0.8801262960169057
Coeficiente de determinación: 0.7746222969404379
Pendiente: -0.2618136813681368
Intervalo de confianza para b1:  (-inf, inf)
Valor-p de shapiro:  0.4172971767713699
Yc para 18:9.346197119711972
Yc para 35:4.895364536453645
```

```
Yc para 40:3.586296129612961
SXX:1269.7142857142858
SXY:-332.42857142857144
SYY:112.35714285714285
SSE:25.32279477947794
SST:112.35714285714285
SSR:7.252862339805411
MSR:7.252862339805411
MSE:2.1102328982898286
F:41.24395375894251

<ipython-input-2-e45e7687d600>:35: RuntimeWarning: divide by zero
encountered in scalar divide
  se_b1 = np.sqrt(np.sum((Y - Yc)**2) / (2 - 2)) / np.sqrt(SXX)
```





Problema 3

A menudo, quienes hacen la contabilidad de costos estiman los gastos generales con base en el nivel de producción. En Standard Knitting Co. han reunido información acerca de los gastos generales y las unidades producidas en diferentes plantas.

Gastos generales	191	170	272	155	280	173	234	116	153	178
Unidades	40	42	53	35	56	39	48	30	37	40

1. Establesca una variable dependiente (Y) y una variable independiente (X).
2. Realice un diagrama de dispersión para estos datos.
3. ¿Los datos soportan la suposición de linealidad?
4. Calcule el coeficiente de correlación e interprete el resultado.
5. Calcule el coeficiente de determinación e interprete el resultado.
6. Obtenga la recta de regresión ajustada y grafíquelo sobre el gráfico de dispersión.
7. Obtenga un intervalo de confianza del 95% para la pendiente de la recta de regresión ajustada (b_1)
8. Calcule los residuales y trace un nuevo gráfico de dispersión. Comente, ¿Parece que se verifican los supuestos?
9. Realice la prueba de Shapiro para los residuales y comente el resultado.
10. Realice la prueba de Brausch-Pagan para los residuales y comente el resultado.

11. Utiliza la recta de regresión para interpolar dos valores y extrapolar uno. Comenta estos resultados.
12. Realice una tabla ANOVA e interprete el resultado.

```
#. 1 Establesca una variable dependiente ( Y ) y una variable
independiente ( X ).


#variable independiente: gastos generales
#variable dependiente: unidades
import numpy as np
Y = np.array([40, 42, 53, 35, 56, 39, 48, 30, 37, 40])
X = np.array([191, 170, 272, 155, 280, 173, 234, 116, 153, 178])
#. 2 Realice un diagrama de dispersión para estos datos.
import matplotlib.pyplot as plt
plt.scatter(X, Y, color = 'pink')
plt.xlabel('Gastos generales')
plt.ylabel('Unidades')
plt.title('Grafico de dispersión')
plt.grid()
#. 3 ¿Los datos soportan la suposición de linealidad?
#. 4 Calcule el coeficiente de correlación e interprete el resultado.
SXX = np.sum((X - np.mean(X))**2)
SXY = np.sum((X - np.mean(X))*(Y - np.mean(Y)))
SYY = np.sum((Y - np.mean(Y))**2)
r = SXY / np.sqrt(SXX * np.sum((Y - np.mean(Y))**2))
print("Coeficiente de correlacion: ", r)
#. 5 Calcule el coeficiente de determinación e interprete el
resultado.
print('Coeficiente de determinación:', r ** 2)
#. 6 Obtenga la recta de regresión ajustada y grafíquelo sobre el
gráfico de dispersión.
b1 = SXY / SXX
b0 = np.mean(Y) - b1 * np.mean(X)
print('Pendiente:', b1) #Corrected print statement
Yc = b0 + b1 * X
plt.plot(X, Yc, '--', color = 'pink')
#. 7 Obtenga un intervalo de confianza del 95% para la pendiente de la
recta de regresión ajustada (b1).
nivel_de_significancia = 0.05
from scipy.stats import t
t_value = t.ppf(1- nivel_de_significancia / 2, len(Y) - 2)
se_b1 = np.sqrt(np.sum((Y - Yc)**2) / (2 - 2)) / np.sqrt(SXX)
confianza_b1 = (b1 - t_value * se_b1, b1 + t_value * se_b1)
print(f'Intervalo de confianza para b1: ', confianza_b1)
#. 8 Calcule los residuales y trace un nuevo gráfico de dispersión.
Comente, ¿Parece que se verifican los supuestos?
residuales = Y - Yc
plt.figure()
plt.scatter(X, residuales, color = 'pink')
```

```
plt.xlabel('Gastos generales')
plt.ylabel('Residuales')
plt.title('Grafico de dispersión de los residuales')
plt.axhline(y=0,color = "pink", linestyle = "--")
#. 9 Realice la prueba de Shapiro para los residuales y comente el
resultado.
from scipy.stats import shapiro
_, valor_p_sh = shapiro(residuales)
print('Valor-p de shapiro: ', valor_p_sh)
#. 10 Realice la prueba de Brausch-Pagan para los residuales y comente
el resultado.
#. 11 Utiliza la recta de regresión para interpolar dos valores y
extrapolar uno. Comenta estos resultados.
print(F"Yc para 191:{b1 * 191 + b0}")
print(F"Yc para 272:{b1 * 272 + b0}")
print(F"Yc para 178:{b1 * 178 + b0}")
#. 12 Realice una tabla ANOVA e interprete el resultado.
print(F"SXX:{SXX}")
print(F"SXY:{SXY}")
print(F"SYY:{SYY}")
print(F"SSE:{np.sum(residuales**2)}")
print(F"SST:{np.sum((Y-np.mean(Y))**2)}")
print(F"SSR:{np.sum((Yc-np.mean(Y))**2)/(len(X)-2)}")
print(F"MSR:{np.sum((Yc-np.mean(Y))**2)/(len(X)-2)}")
print(F"MSE:{np.sum(residuales**2)/(len(X)-2)}")
print(F"F:{np.sum((Yc -
np.mean(Y))**2)/1/(np.sum(residuales**2)/(len(Y)-2))}")
```

```
Coeficiente de correlacion:  0.9835155492696092
Coeficiente de determinación: 0.967302835655101
Pendiente: 0.14901075906869252
Intervalo de confianza para b1:  (-inf, inf)
Valor-p de shapiro:  0.30963893537420123
Yc para 191:41.82118708911757
Yc para 272:53.89105857368166
Yc para 178:39.88404722122456
SXX:25615.6
SXY:3817.0
SYY:588.0
SSE:19.22593263480066
SST:588.0
SSR:71.09675842064993
MSR:71.09675842064993
MSE:2.4032415793500825
F:236.66953511973404
```

```
<ipython-input-3-2387b7130ad1>:35: RuntimeWarning: divide by zero
encountered in scalar divide
  se_b1 = np.sqrt(np.sum((Y - Yc)**2) / (2 - 2)) / np.sqrt(SXX)
```

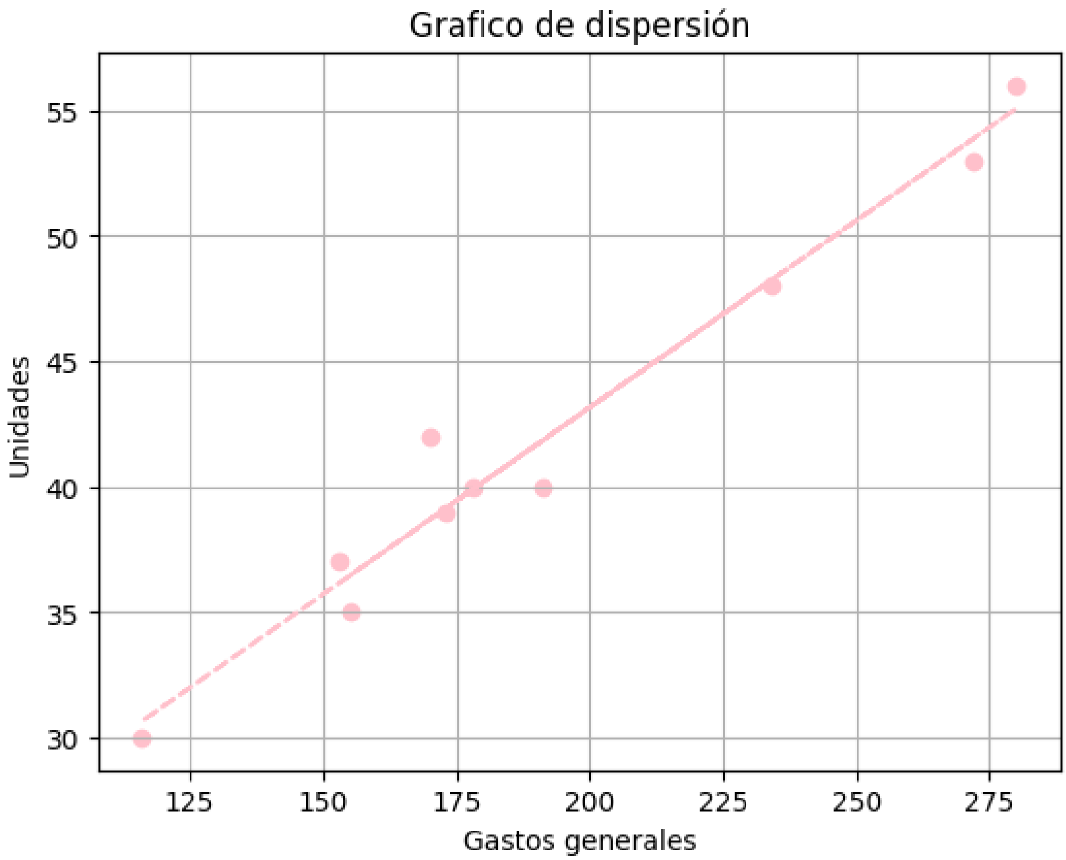
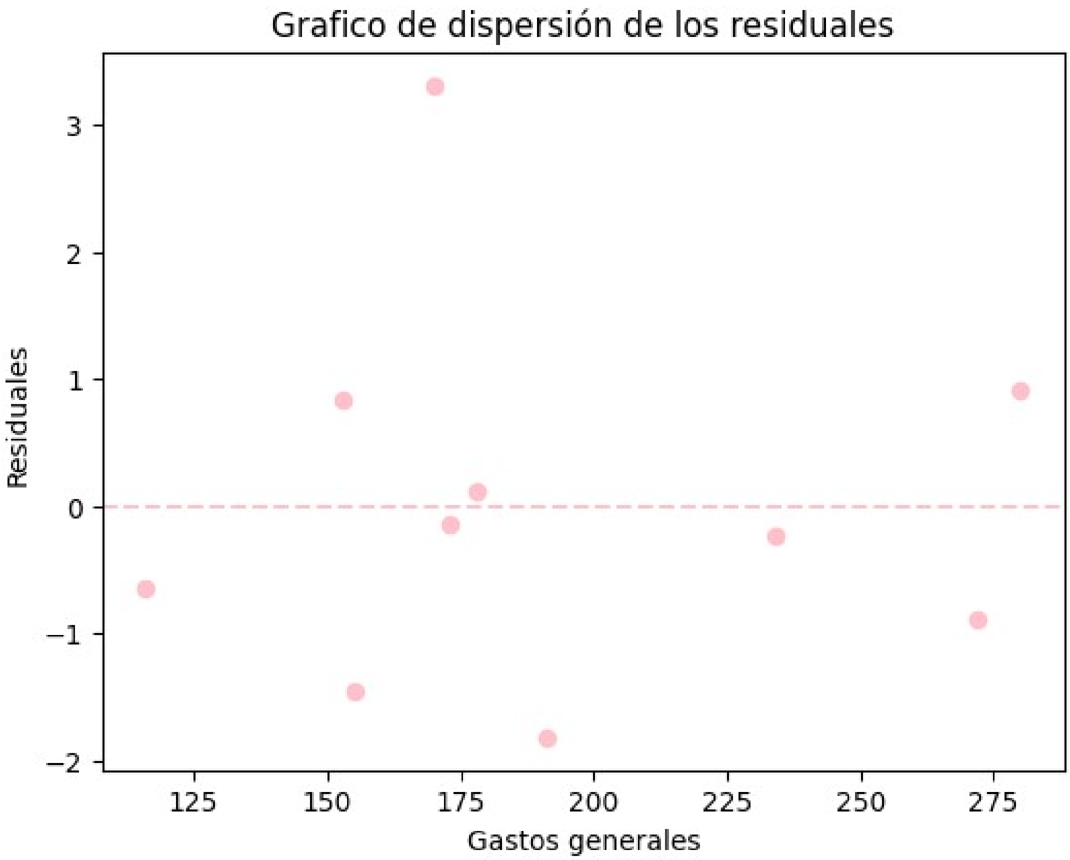


Grafico de dispersión
Unidades
Gastos generales
55
50
45
40
35
30
125
150
175
200
225
250
275



Problema 4

Las ventas de línea blanca varían según el estado del mercado de casas nuevas: cuando las ventas de casas nuevas son buenas, también lo son las de lavaplatos, lavadoras de ropa, secadoras y refrigeradores. Una asociación de comercio compiló los siguientes datos históricos (en miles de unidades) de las ventas de línea blanca y la construcción de casas.

Construcción de casas (miles)	Ventas de línea blanca (miles)
2.0	5.0
2.5	5.5
3.2	6.0
3.6	7.0
3.7	7.2
4.0	7.7
4.2	8.4
4.6	9.0
4.8	9.7
5.0	10.0

1. Establesca una variable dependiente (Y) y una variable independiente (X).
2. Realice un diagrama de dispersión para estos datos.

3. ¿Los datos soportan la suposición de linealidad?
4. Calcule el coeficiente de correlación e interprete el resultado.
5. Calcule el coeficiente de determinación e interprete el resultado.
6. Obtenga la recta de regresión ajustada y grafíquelo sobre el gráfico de dispersión.
7. Obtenga un intervalo de confianza del 95% para la pendiente de la recta de regresión ajustada (b_1)
8. Calcule los residuales y trace un nuevo gráfico de dispersión. Comente, ¿Parece que se verifican los supuestos?
9. Realice la prueba de Shapiro para los residuales y comente el resultado.
10. Realice la prueba de Brausch-Pagan para los residuales y comente el resultado.
11. Utiliza la recta de regresión para interpolar dos valores y extrapolar uno. Comenta estos resultados.
12. Realice una tabla ANOVA e interprete el resultado.

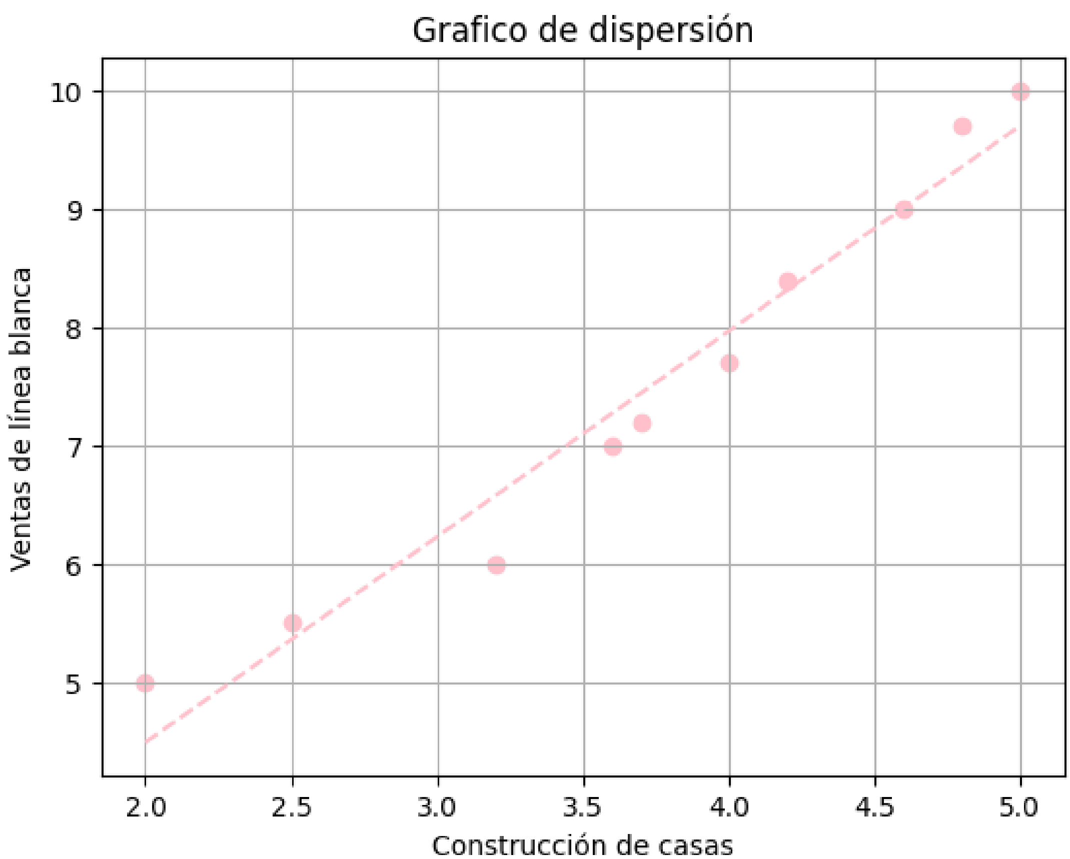
```
#. 1 Establesca una variable dependiente ( Y ) y una variable
independiente ( X ).
import numpy as np
Y = np.array([5, 5.5, 6, 7, 7.2, 7.7, 8.4, 9, 9.7, 10])
X = np.array([2, 2.5, 3.2, 3.6, 3.7, 4, 4.2, 4.6, 4.8, 5])
#. 2 Realice un diagrama de dispersión para estos datos.
import matplotlib.pyplot as plt
plt.scatter(X, Y, color = 'pink')
plt.xlabel('Construcción de casas')
plt.ylabel('Ventas de línea blanca')
plt.title('Grafico de dispersión')
plt.grid()
#. 3 ¿Los datos soportan la suposición de linealidad?
#. 4 Calcule el coeficiente de correlación e interprete el resultado.
SXX = np.sum((X - np.mean(X))**2)
SXY = np.sum((X - np.mean(X))*(Y - np.mean(Y)))
SYY = np.sum((Y - np.mean(Y))**2)
r = SXY / np.sqrt(SXX * np.sum((Y - np.mean(Y))**2))
print("Coeficiente de correlacion: ", r)
#. 5 Calcule el coeficiente de determinación e interprete el
resultado.
print('Coeficiente de determinación:', r ** 2)
#. 6 Obtenga la recta de regresión ajustada y grafíquelo sobre el
gráfico de dispersión.
b1 = SXY / SXX
b0 = np.mean(Y) - b1 * np.mean(X)
print('Pendiente:', b1) #Corrected print statement
Yc = b0 + b1 * X
plt.plot(X, Yc, '--', color = 'pink')
#. 7 Obtenga un intervalo de confianza del 95% para la pendiente de la
recta de regresión ajustada (b1).
nivel_de_significancia = 0.05
from scipy.stats import t
t_value = t.ppf(1- nivel_de_significancia / 2, len(Y) - 2)
```

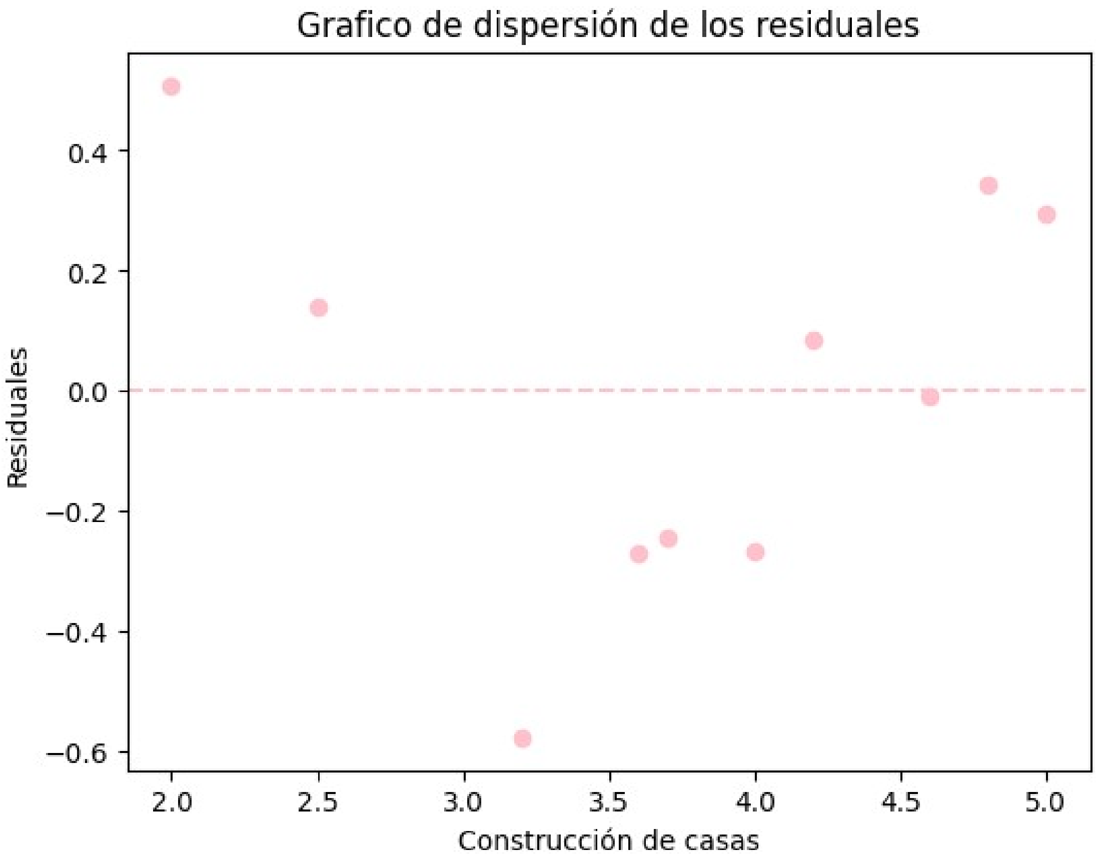
```
se_b1 = np.sqrt(np.sum((Y - Yc)**2) / (2 - 2)) / np.sqrt(SXX)
confianza_b1 = (b1 - t_value * se_b1, b1 + t_value * se_b1)
print(f'Intervalo de confianza para b1: ', (b1 - t_value * se_b1, b1 +
t_value * se_b1))
#. 8 Calcule los residuales y trace un nuevo gráfico de dispersión.
Comente, ¿Parece que se verifican los supuestos?
residuales = Y - Yc
plt.figure()
plt.scatter(X, residuales, color = 'pink')
plt.xlabel('Construcción de casas')
plt.ylabel('Residuales')
plt.title('Grafico de dispersión de los residuales')
plt.axhline(y=0,color = "pink", linestyle = "--")
#. 9 Realice la prueba de Shapiro para los residuales y comente el
resultado.
from scipy.stats import shapiro
_, valor_p_sh = shapiro(residuales)
print('Valor-p de shapiro: ', valor_p_sh)
#. 10 Realice la prueba de Brausch-Pagan para los residuales y comente
el resultado.
#. 11 Utiliza la recta de regresión para interpolar dos valores y
extrapolar uno. Comenta estos resultados.
print(F"Yc para 2:{b1 * 2 + b0}")
print(F"Yc para 3.7:{b1 * 3.7 + b0}")
print(F"Yc para 5:{b1 * 5 + b0}")
#. 12 Realice una tabla ANOVA e interprete el resultado.
print(F"SXX:{SXX}")
print(F"SXY:{SXY}")
print(F"SYY:{SYY}")
print(F"SSE:{np.sum(residuales**2)}")
print(F"SST:{np.sum((Y-np.mean(Y))**2)}")
print(F"SSR:{np.sum((Yc-np.mean(Y))**2)/(len(X)-2)}")
print(F"MSR:{np.sum((Yc-np.mean(Y))**2)/(len(X)-2)}")
print(F"MSE:{np.sum(residuales**2)/(len(X)-2)}")
print(F"F:{np.sum((Yc -
np.mean(Y))**2)/1/(np.sum(residuales**2)/(len(Y)-2))}")
```

```
Coeficiente de correlacion:  0.980773902153433
Coeficiente de determinación: 0.9619174471452717
Pendiente: 1.7375639237563925
Intervalo de confianza para b1:  (-inf, inf)
Valor-p de shapiro:  0.8463507249054649
Yc para 2:4.491887494188749
Yc para 3.7:7.445746164574616
Yc para 5:9.704579265457927
SXX:8.604
SXY:14.95
SYY:27.005
```

```
SSE:1.0284193398419348
SST:27.005
SSR:3.247072582519758
MSR:3.247072582519758
MSE:0.12855241748024185
F:202.06995068101767

<ipython-input-10-ca2685207cf9>:31: RuntimeWarning: divide by zero
encountered in scalar divide
  se_b1 = np.sqrt(np.sum((Y - Yc)**2) / (2 - 2)) / np.sqrt(SXX)
```





Problema 5

William C. Andrews, consultor de comportamiento organizacional de Victory Motorcycles, ha diseñado una prueba para mostrar a los supervisores de la compañía los peligros de sobrevigilar a sus trabajadores. Un trabajador de la línea de ensamble tiene a su cargo una serie de tareas complicadas. Durante el desempeño del trabajador, un inspector lo interrumpe constantemente para ayudarlo a terminar las tareas. El trabajador, después de terminar su trabajo, recibe una prueba psicológica diseñada para medir la hostilidad del trabajador hacia la autoridad (una alta puntuación implica una hostilidad baja). A ocho distintos trabajadores se les asignaron las tareas y luego se les interrumpió para darles instrucciones útiles un número variable de veces (línea X). Sus calificaciones en la prueba de hostilidad se dan en el renglón Y.

número interrupciones al trabajador	5	10	10	15	15	20	20	25
calificación del trabajador en la prueba de hostilidad	58	41	45	27	26	12	16	3

1. Establesca una variable dependiente (Y) y una variable independiente (X).
2. Realice un diagrama de dispersión para estos datos.
3. ¿Los datos soportan la suposición de linealidad?
4. Calcule el coeficiente de correlación e interprete el resultado.
5. Calcule el coeficiente de determinación e interprete el resultado.
6. Obtenga la recta de regresión ajustada y grafíquelo sobre el gráfico de dispersión.

7. Obtenga un intervalo de confianza del 95% para la pendiente de la recta de regresión ajustada (b_1)
8. Calcule los residuales y trace un nuevo gráfico de dispersión. Comente, ¿Parece que se verifican los supuestos?
9. Realice la prueba de Shapiro para los residuales y comente el resultado.
10. Realice la prueba de Brausch-Pagan para los residuales y comente el resultado.
11. Utiliza la recta de regresión para interpolar dos valores y extrapolar uno. Comenta estos resultados.
12. Realice una tabla ANOVA e interprete el resultado.

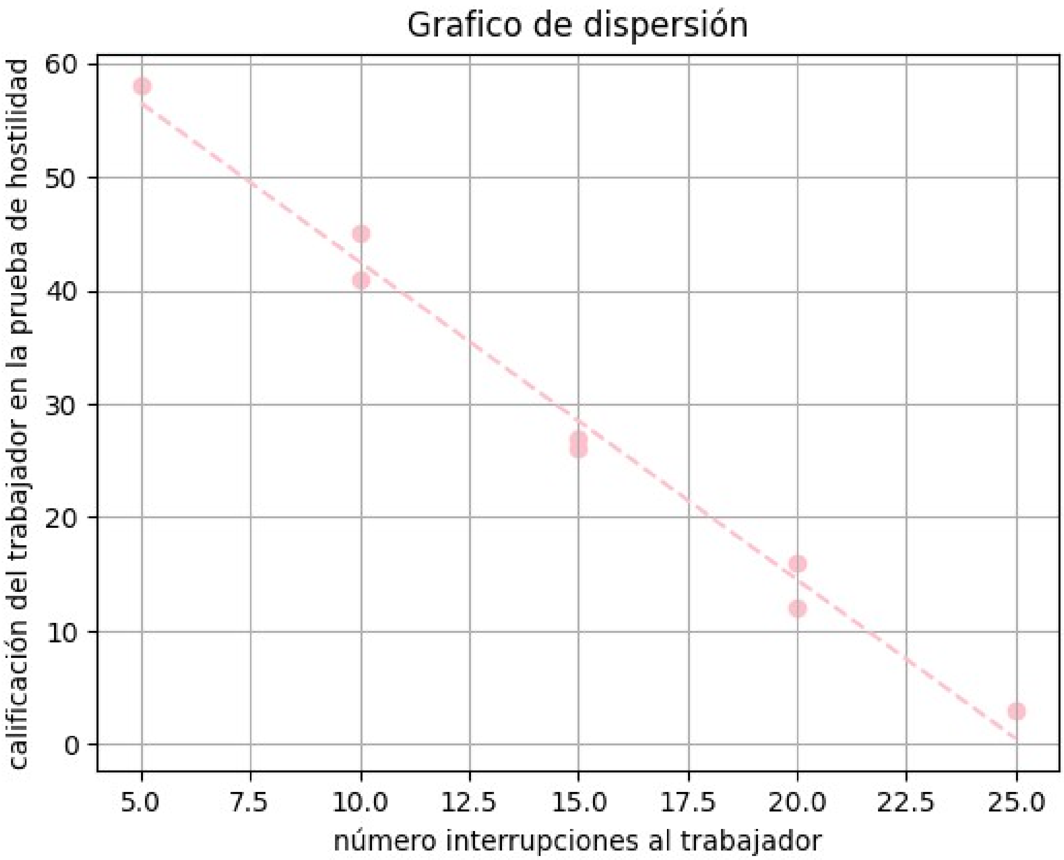
```python
#. 1 Establesca una variable dependiente ( Y ) y una variable
independiente ( X )
import numpy as np
Y = np.array([58, 41, 45, 27, 26, 12, 16, 3])
X = np.array([5, 10, 10, 15, 15, 20, 20, 25])
#. 2 Realice un diagrama de dispersión para estos datos.
import matplotlib.pyplot as plt
plt.scatter(X, Y, color = 'pink')
plt.xlabel('número interrupciones al trabajador')
plt.ylabel('calificación del trabajador en la prueba de hostilidad')
plt.title('Grafico de dispersión')
plt.grid()
#. 3 ¿Los datos soportan la suposición de linealidad?
#. 4 Calcule el coeficiente de correlación e interprete el resultado.
SXX = np.sum((X - np.mean(X))**2)
SXY = np.sum((X - np.mean(X))*(Y - np.mean(Y)))
SYY = np.sum((Y - np.mean(Y))**2)
r = SXY / np.sqrt(SXX * np.sum((Y - np.mean(Y))**2))
print("Coeficiente de correlacion: ", r)
#. 5 Calcule el coeficiente de determinación e interprete el
resultado.
print('Coeficiente de determinación:', r ** 2)
#. 6 Obtenga la recta de regresión ajustada y grafíquelo sobre el
gráfico de dispersión.
b1 = SXY / SXX
b0 = np.mean(Y) - b1 * np.mean(X)
print('Pendiente:', b1) #Corrected print statement
Yc = b0 + b1 * X
plt.plot(X, Yc, '--', color = 'pink')
#. 7 Obtenga un intervalo de confianza del 95% para la pendiente de la
recta de regresión ajustada (b1).
nivel_de_significancia = 0.05
from scipy.stats import t
t_value = t.ppf(1- nivel_de_significancia / 2, len(Y) - 2)
se_b1 = np.sqrt(np.sum((Y - Yc)**2) / (2 - 2)) / np.sqrt(SXX)
print(f'Intervalo de confianza para b1: ', (b1 - t_value * se_b1, b1 +
t_value * se_b1))
#. 8 Calcule los residuales y trace un nuevo gráfico de dispersión.
Comente, ¿Parece que se verifican los supuestos?
```

```python
residuales = Y - Yc
plt.figure()
plt.scatter(X, residuales, color = 'pink')
plt.xlabel('número interrupciones al trabajador')
plt.ylabel('residuales')
plt.title('Grafico de dispersión de los residuales')
plt.axhline(y=0,color = "pink", linestyle = "--")
#. 9 Realice la prueba de Shapiro para los residuales y comente el
resultado.
from scipy.stats import shapiro
_, valor_p_sh = shapiro(residuales)
print('Valor-p de shapiro: ', valor_p_sh)
#. 10 Realice la prueba de Brausch-Pagan para los residuales y comente
el resultado.
#. 11 Utiliza la recta de regresión para interpolar dos valores y
extrapolar uno. Comenta estos resultados.
print(F"Yc para 5:{b1 * 5 + b0}")
print(F"Yc para 15:{b1 * 15 + b0}")
print(F"Yc para 20:{b1 * 20 + b0}")
#. 12 Realice una tabla ANOVA e interprete el resultado.
print(F"SXX:{SXX}")
print(F"SXY:{SXY}")
print(F"SYY:{SYY}")
print(F"SSE:{np.sum(residuales**2)}")
print(F"SST:{np.sum((Y-np.mean(Y))**2)}")
print(F"SSR:{np.sum((Yc-np.mean(Y))**2)/(len(X)-2)}")
print(F"MSR:{np.sum((Yc-np.mean(Y))**2)/(len(X)-2)}")
print(F"MSE:{np.sum(residuales**2)/(len(X)-2)}")
print(F"F:{np.sum((Yc -
np.mean(Y))**2)/1/(np.sum(residuales**2)/(len(Y)-2))}")
```

```
Coeficiente de correlacion:  -0.9928495402404848
Coeficiente de determinación: 0.985750209555742
Pendiente: -2.8
Intervalo de confianza para b1:  (-inf, inf)
Valor-p de shapiro:  0.05481649112766485
Yc para 5:56.5
Yc para 15:28.5
Yc para 20:14.5
SXX:300.0
SXY:-840.0
SYY:2386.0
SSE:34.0
SST:2386.0
```

```
SSR:392.0
MSR:392.0
MSE:5.666666666666667
F:415.05882352941177

<ipython-input-11-66e1ced179ac>:31: RuntimeWarning: divide by zero
encountered in scalar divide
  se_b1 = np.sqrt(np.sum((Y - Yc)**2) / (2 - 2)) / np.sqrt(SXX)
```



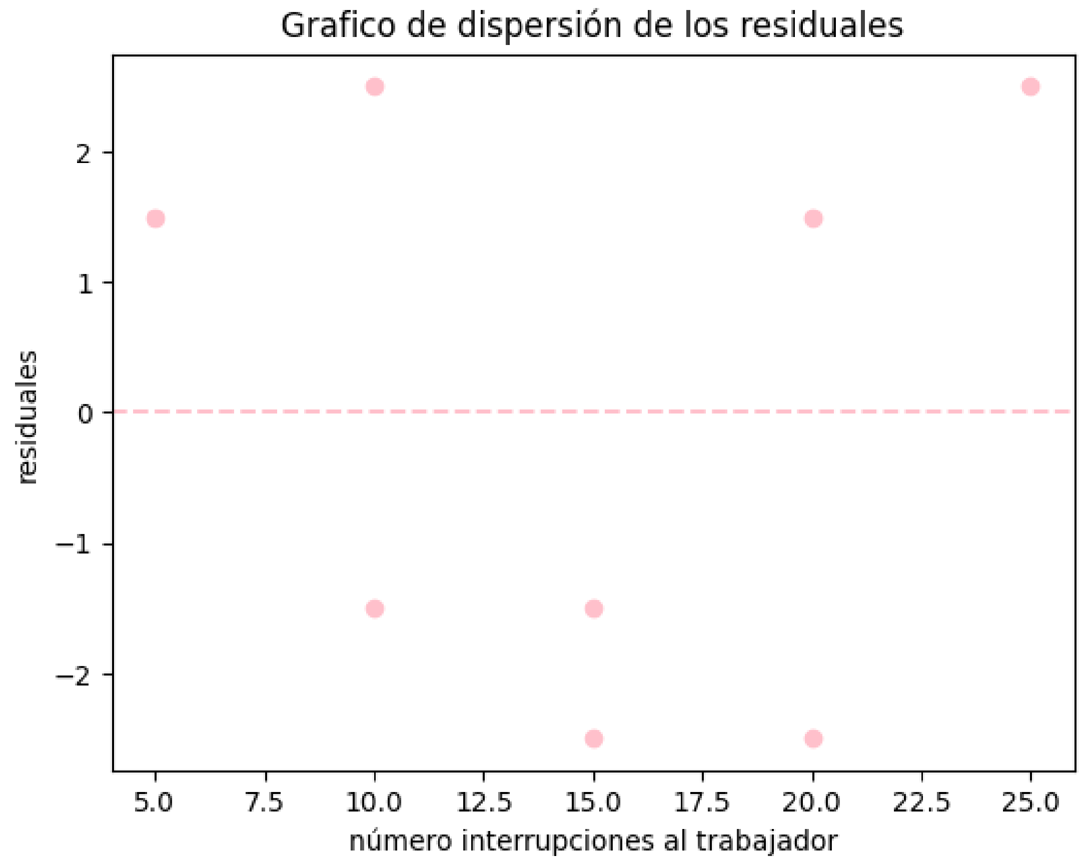


Grafico de dispersión de los residuales
residuales
2
1
0
−1
−2
5.0
7.5
10.0
12.5
15.0
17.5
20.0
22.5
25.0
número interrupciones al trabajador