EJERCICIO DE REGRESIÓN LINEAL. GRUPO 5CDM1

Materia: Minería de datos **Periodo escolar: 2022-2**

Nombre del alumno:

De Luna Ocampo Yanina

En una compañía se aplicó un examen para medir el nivel de hostilidad hacia la autoridad, una puntuación alta implica una hostilidad baja. A diez trabajadores se les asignaron 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.
Se desea crear un modelo matemático que represente la relación de los datos. Utilice la guía proporcionada.
Ejercicio adaptado de Levín, Rubín, Balderas, Del Valle y Gómez. (2004). Estadística para administración y economía. Séptima Edición. Prentice-Hall.

X (número de interrupciones al trabajador)	**Y (calificación del trabajador en la prueba de hostilidad)**
5	58
10	41
10	45
15	27
15	26
20	12
20	16
25	3
25	5
30	2

Responder cada uno de los siguientes incisos. Agregar la generación de tablas de cálculos y la presentación de las fórmulas que utilice en cada sección.

1) Generar la gráfica de variables

2) Realice los cálculos *pasos a paso* para generar la ecuación de regresión.

3) Realice la verificación de la ecuación de regresión de una recta generada con el método de mínimos cuadrados.

4) Realice los siguientes cálculos (muestre el proceso)
   a) Suma de cuadrados debida al error
   b) Suma total de cuadrados
   c) Suma de cuadrados debida a la regresión
   d) El coeficiente de determinación

e) Exprese el significado del coeficiente de determinación encontrado
f) El coeficiente de correlación y su significado

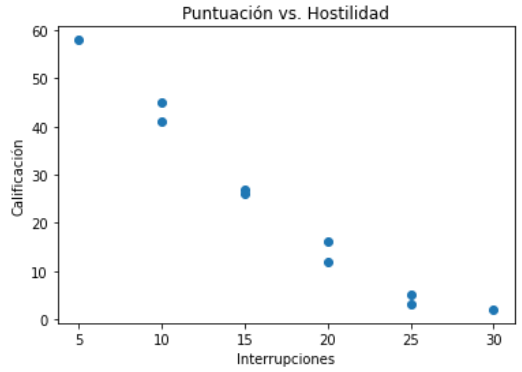
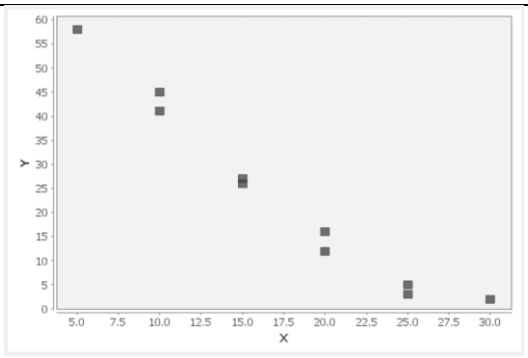
5) Calcule los errores estándar de la estimación

6) Los intervalos de confianza

7) Aplique la prueba t para determinar si el modelo es estadísticamente significativo

8) Genere la ecuación de recta en el Knime incorporando prueba de normalidad y gráfico de residuales.

SOLUCIÓN:

Python	KNIME
	

Pasos realizados en Excel:

1. Colocamos nuestras columnas con el respectivo X y Y.

X	Y
5	58
10	41
10	45
15	27
15	26
20	12
20	16
25	3
25	5
30	2

2. Comenzamos a sacar los valores de $\bar{x}$ y de $\bar{y}$

Xbarra	17.5
Ybarra	23.5

3. Una vez que obtenemos esos valores, sacamos las columnas de x-$\bar{x}$ y de y-$\bar{y}$, recordemos que x y y son las columnas de nuestro csv. Añadiendo también en otras dos columnas el cuadrado de lo obtenido previamente con las restas.

X-Xbarra	Y-Ybarra	(Y-Ybarra)^2	(X-Xbarra)^2
-12.5	34.5	1190.25	156.25
-7.5	17.5	306.25	56.25
-7.5	21.5	462.25	56.25
-2.5	3.5	12.25	6.25
-2.5	2.5	6.25	6.25
2.5	-11.5	132.25	6.25
2.5	-7.5	56.25	6.25
7.5	-20.5	420.25	56.25
7.5	-18.5	342.25	56.25
12.5	-21.5	462.25	156.25

4. Multiplicamos las primeras dos columnas porque las necesitaremos para los cálculos siguientes.

(X-Xbarra)*(Y-Ybarra)
-431.25
-131.25
-161.25
-8.75
-6.25
-28.75
-18.75
-153.75
-138.75
-268.75

5. Obtenemos la suma de cada columna sacada

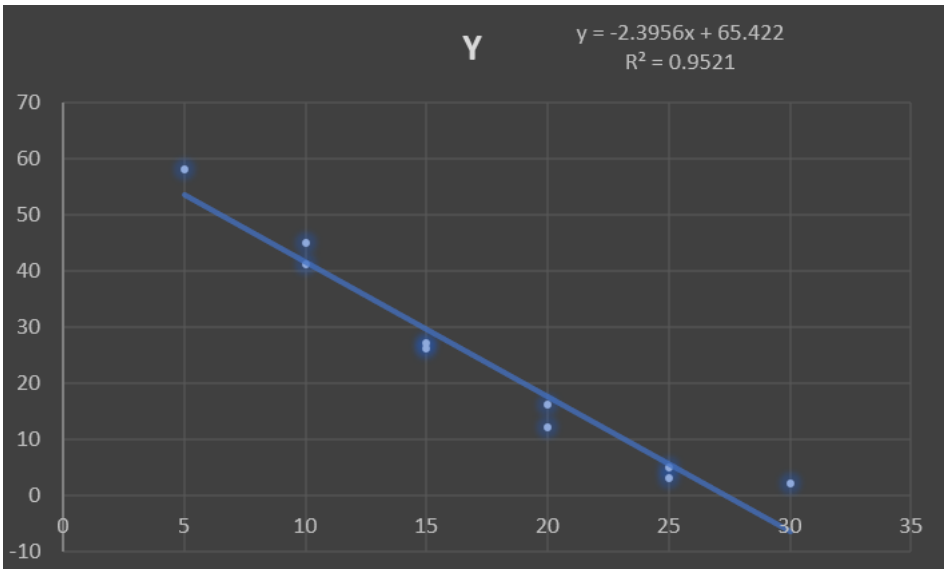
X-Xbarra	Y-Ybarra	(X-Xbarra)*(Y-Ybarra)	(Y-Ybarra)^2	(X-Xbarra)^2
-12.5	34.5	-431.25	1190.25	156.25
-7.5	17.5	-131.25	306.25	56.25
-7.5	21.5	-161.25	462.25	56.25
-2.5	3.5	-8.75	12.25	6.25
-2.5	2.5	-6.25	6.25	6.25
2.5	-11.5	-28.75	132.25	6.25
2.5	-7.5	-18.75	56.25	6.25
7.5	-20.5	-153.75	420.25	56.25
7.5	-18.5	-138.75	342.25	56.25
12.5	-21.5	-268.75	462.25	156.25
0	0	-1347.5	3390.5	562.5

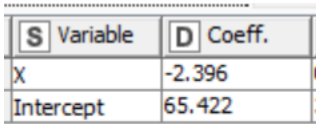
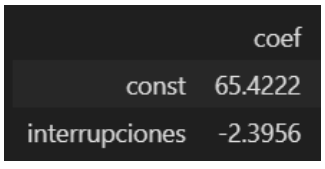
6. Obtenemos Beta0 y Beta1 con ayuda de los cálculos de arriba, entendiendo que:
   B1 = (X-Xbarra * Y-Ybarra) / (X-Xbarra)^2
   B0 = Ybarra – Beta1 * Xbarra

Beta0	65.4222222
Beta1	-2.3955556

7. Con esos datos ya podemos obtener la gráfica que nos ayudará a ver si nuestros Betas fueron correctamente calculamos. Una vez haciendo esto, vemos que se calculó de forma correcta, entonces la ecuación presentada es:

y = -2.3956x + 65.422



KNIME	
Python	

Excel	

Beta0	65.42222222
Beta1	-2.395555556

Obtenemos las estadísticas:

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.975743431
Coeficiente de determinación R^2	0.952075243
R^2 ajustado	0.946084648
Error típico	4.506785008
Observaciones	10

El coeficiente de determinación es la proporción de la varianza en la variable de respuesta que se pueda explicar por la variable explicativa. Aquí entendemos que el 95.2% de la variación en los puntuajes de la hostilidad se debe a las veces de interrupción a cada trabajador.

El error típico o error estándar es la distancia promedio que los valores observados caen desde la línea de regresión. En este caso, los valores observados caen un promedio de 4.5 unidades de la línea de regresión.

ANÁLISIS DE VARIANZA					
	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	1	3228.011111	3228.011111	158.928337	1.47095E-06
Residuos	8	162.4888889	20.31111111		
Total	9	3390.5			

El F es el estadístico general para el modelo de regresión, calculando como MS de regresión / MS residual. Este es de 158.92

El valor crítico de F en este caso es de 1.47E-06, este es el valor asociado con el estadístico F general. Nos dice si es modelo de regresión es estadísticamente significativo o no. De otra forma, nos dice si la variable explicativa tiene una asociación estadísticamente significativa con la variable de respuesta. En este caso, el valor p es menor que 0.05, lo que indica que existe una asociación estadísticamente significativa entre la hostilidad y las veces de interrupción a cada trabajador.

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*	*Inferior 95.0%*	*Superior 95.0%*
Intercepción	65.42222222	3.617925195	18.08280125	8.98036E-08	57.07927176	73.76517268	57.07927176	73.76517268
Variable X 1	-2.395555556	0.190022741	-12.60667827	1.47095E-06	-2.833748781	-1.95736233	-2.833748781	-1.95736233

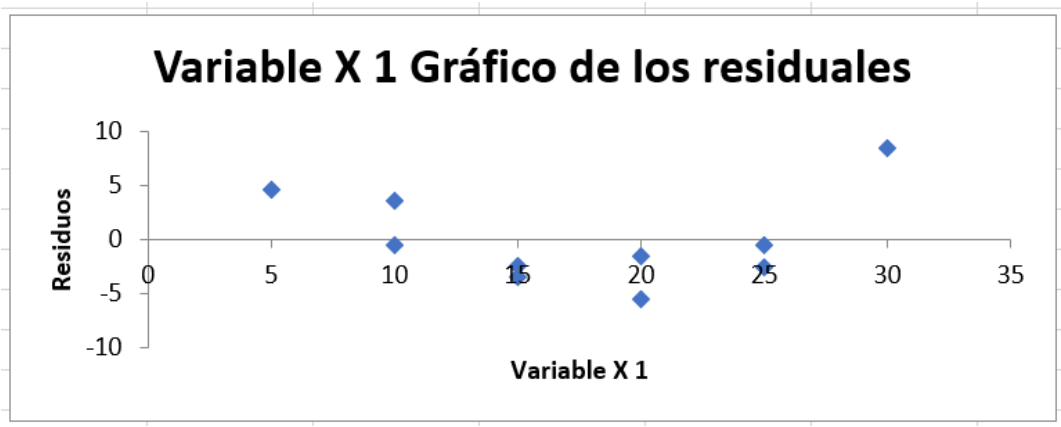
Análisis de los residuales

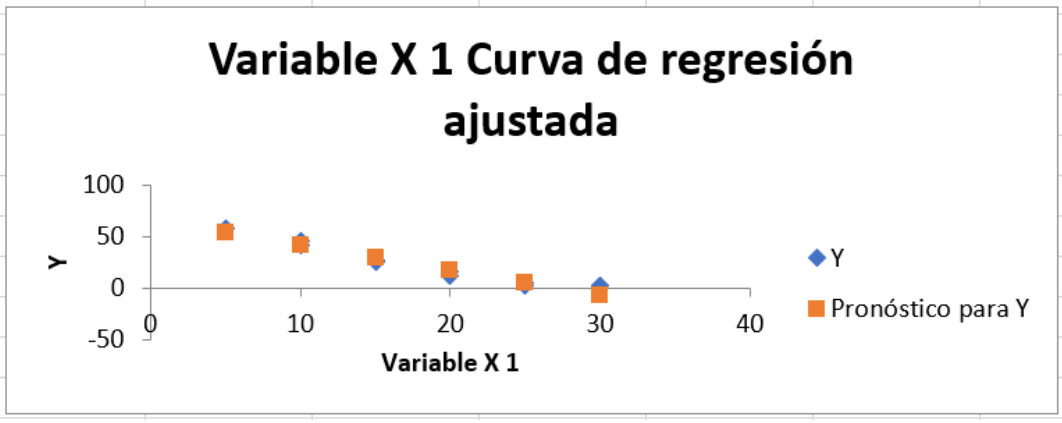
Observación	*Pronóstico para Y*	*Residuos*	*esiduos estándares*
1	53.44444444	4.555555556	1.072138194
2	41.46666667	-0.466666667	-0.109828791
3	41.46666667	3.533333333	0.831560843
4	29.48888889	-2.488888889	-0.58575355
5	29.48888889	-3.488888889	-0.821100958
6	17.51111111	-5.511111111	-1.297025717
7	17.51111111	-1.511111111	-0.355636084
8	5.533333333	-2.533333333	-0.596213435
9	5.533333333	-0.533333333	-0.125518618
10	-6.444444444	8.444444444	1.987378115

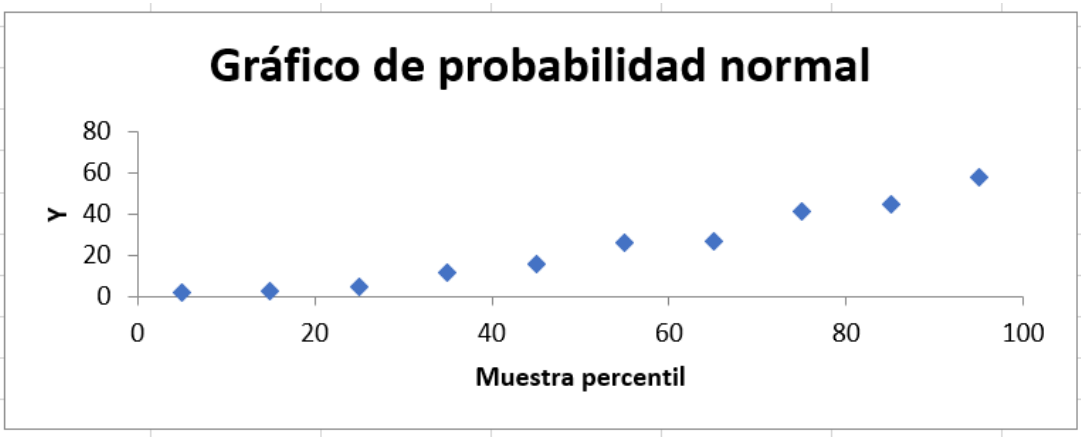
Resultados de datos de probabilidad

Percentil	*Y*
5	2
15	3
25	5
35	12
45	16
55	26
65	27
75	41
85	45
95	58

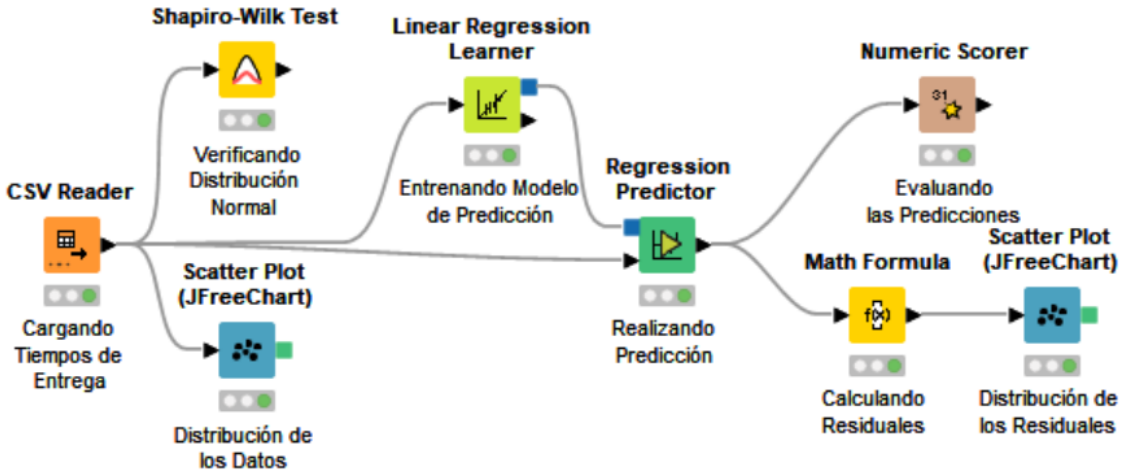
Los coeficientes nos dan los números necesarios para escribir la ecuación de regresión estimada.
Puntuación de hostilidad: -2.3956(interrupciones) + 65.422
Interpretamos que el coeficiente de interrupciones significa que, por cada interrupción, se espera que el nivel de hostilidad disminuya 2.3956 en promedio.
La hostilidad del trabajador se espera que sea cuando las interrupciones seas de cero, esta sea de 65.42.







Modelo en KNIME:

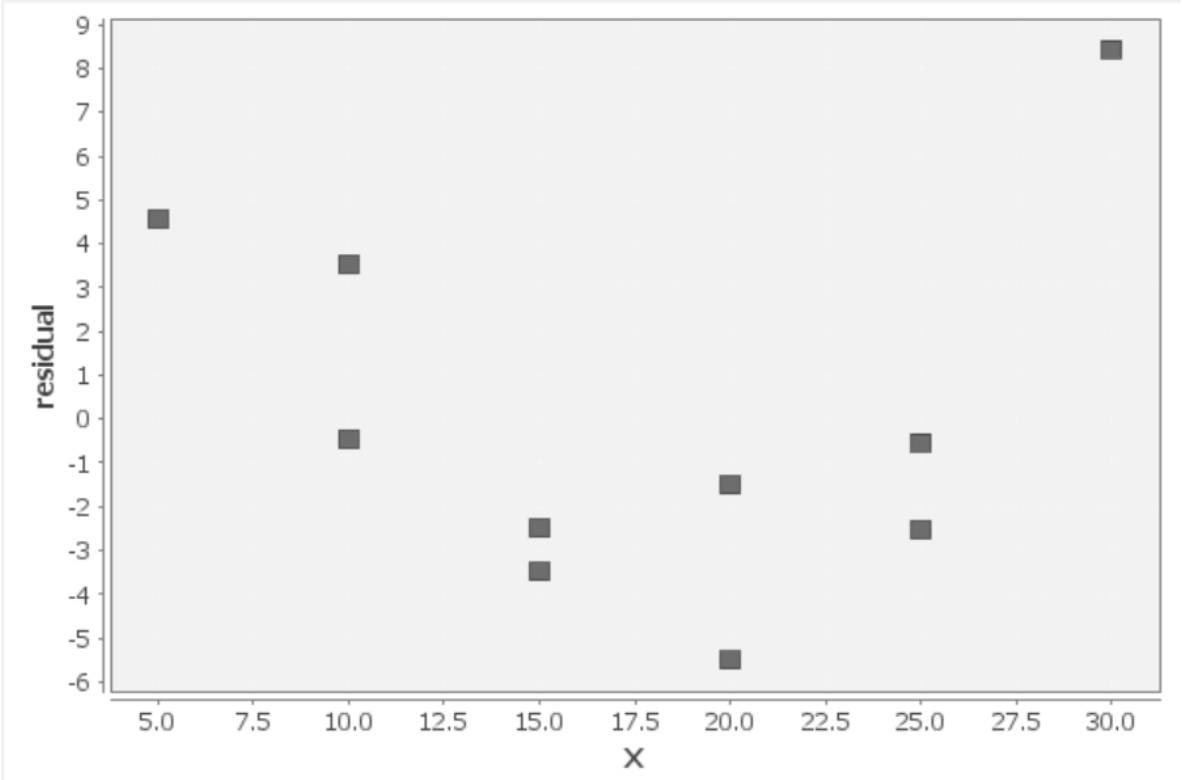


Row ID	S Variable	D Coeff.	D Std. Err.	D t-value	D P>|t|
Row1	X	-2.396	0.19	-12.607	0
Row2	Intercept	65.422	3.618	18.083	0

Row ID	I X	I Y	D Predicti...
Row0	5	58	53.444
Row1	10	41	41.467
Row2	10	45	41.467
Row3	15	27	29.489
Row4	15	26	29.489
Row5	20	12	17.511
Row6	20	16	17.511
Row7	25	3	5.533
Row8	25	5	5.533
Row9	30	2	-6.444

Row ID	I X	I Y	D Predicti...	D residual
Row0	5	58	53.444	4.556
Row1	10	41	41.467	-0.467
Row2	10	45	41.467	3.533
Row3	15	27	29.489	-2.489
Row4	15	26	29.489	-3.489
Row5	20	12	17.511	-5.511
Row6	20	16	17.511	-1.511
Row7	25	3	5.533	-2.533
Row8	25	5	5.533	-0.533
Row9	30	2	-6.444	8.444

Row ID	D Predicti...
R^2	0.952
mean absolut...	3.307
mean square...	16.249
root mean sq...	4.031
mean signed ...	0
mean absolut...	0.612
adjusted R^2	0.952



Row ID	B Reject H0	D Test Statistic (...	D p-Value
Y	false	0.9217713573792...	0.371966059198...