EJERCICIO DE REGRESIÓN LINEAL

Materia de Minería de datos Periodo escolar: 2022-1

Grupo: 3CV19 Equipo: 6

Nombre de los integrantes del equipo:

1) Castro Cruces Jorge Eduardo
2) Guzman Gutierrez Manuel
3) Medina Granados Alan Alejandro

Introducción

La regresión lineal como sabemos es un proceso o técnica estadística para determinar la relación entre variables. Permite predecir a partir de un muestreo de datos aleatorio. Se adapta a una amplia variedad de situaciones.
De igual manera una regresión se usa para predecir los valores ausentes de una variable basándose en su relación con otras variables de la tabla de datos.
En el siguiente ejercicio se resolverán ejercicios de regresión lineal vistos en clase en las diapositivas de este mismo tema, además de que se utilizarán como una base para la resolución de ellos.
El ejercicio en Knime los datos se sacaron de este mismo ejercicio y dependiendo del tema que nos tocó o se nos asignó en clase.

- ## ¿Qué es la regresión lineal?

Es una técnica estadística para determinar la relación entre variables. Permite predecir a partir de un muestreo de datos aleatorio. Se adapta a una amplia variedad de situaciones. La regresión ajustada con el error cuadrático medio más bajo se elige como el modelo final. Al aplicar el análisis de funciones automáticamente se genera un modelo de regresión lineal de predicción. La precisión del modelo generado depende en gran manera de la cantidad de datos que se manejen, así, la exactitud de la predicción es directamente proporcional al número de datos disponibles
El análisis de la regresión lineal se utiliza para predecir el valor de una variable según el valor de otra. La variable que desea predecir se denomina variable dependiente. La variable que está utilizando para predecir el valor de la otra variable se denomina variable independiente.

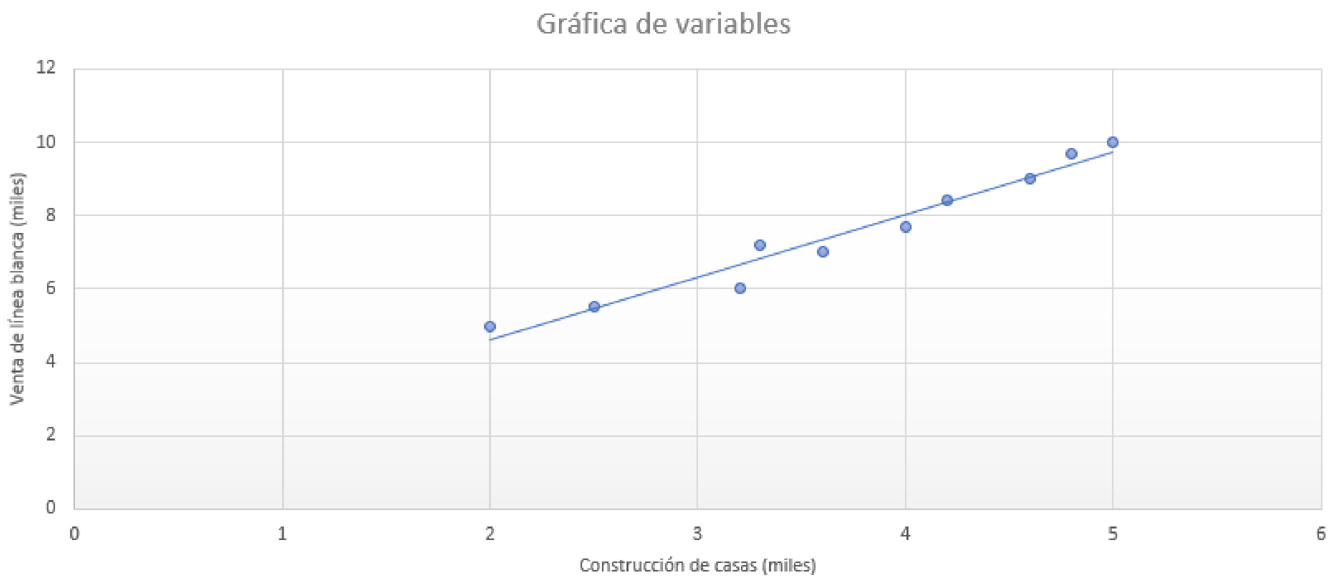
Ejercicio

Las ventas de línea blanca varían dependiendo de la oferta de casas nuevas, cuando se incrementa la venta de casas nuevas también crece el de los electrodomésticos (lavaplatos, lavadoras de ropa, secadoras y refrigeradores). Una empresa compiló los datos que se muestran en la siguiente tabla.
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.

Construcción de casas (miles)	Venta de línea blanca (miles)
2.0	5.0
2.5	5.5
3.2	6.0
3.6	7.0
3.3	7.2
4.0	7.7
4.2	8.4
4.6	9.0
4.8	9.7
5.0	10.0

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.

- ## PRIMER PASO

Construcción de casas (miles)	Venta de línea blanca	XY	x^2
2.0	5.0	10	4
2.5	5.5	13.75	6.25
3.2	6.0	19.2	10.24
3.6	7.0	25.2	12.96
3.3	7.2	23.76	10.89
4.0	7.7	30.8	16
4.2	8.4	35.28	17.64
4.6	9.0	41.4	21.16
4.8	9.7	46.56	23.04
5.0	10.0	50	25
$\boldsymbol{\Sigma X}$ =37.2	$\boldsymbol{\Sigma Y}$ =75.5	$\boldsymbol{\Sigma XY}$ =295.95	$\boldsymbol{\Sigma X^2}$ =147.18

Por lo tanto:

$$\underline{X}= \frac{\Sigma X}{n}= \frac{37.2}{10}= 3.72$$

$$\underline{Y}= \frac{\Sigma Y}{n}= \frac{75.5}{10}= 7.55$$

- **SEGUNDO PASO**

$$b = \frac{\Sigma XY - nXY}{\Sigma X^2 - nX^2} = \frac{295.95-(10)(3.72)(7.55)}{147.18-(10)(3.72)^2} = 1.7155$$

También

$$a = \underline{Y} - b\underline{X}$$
$$a = 7.55 - 1.7155(3.72) = 1.16834$$

La ecuación de regresión Lineal

$$\hat{Y} = a + bX$$

Sustituyendo

$$\hat{Y} = 1.16834 + 1.7155X$$
$$\hat{Y} = 1.16834 + 1.7155(10)$$

$$\hat{Y} = 18.35$$

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

Y	**Ŷ**	**Error individual**
5	- 4.59934	0.40066
5.5	- 5.45709	0.04291
6	- 6.65794	-0.65794
7	- 7.34414	-0.34414
7.2	- 6.82949	0.37051
7.7	- 8.03034	-0.33034
8.4	- 8.37344	0.02656
9	- 9.05964	-0.05964
9.7	- 9.40274	0.29726
10	- 9.74584	0.25416
	Suma=	0

4) Realice los siguientes cálculos (muestre el proceso)

- # a) Suma de cuadrados debida al error

Error Individual	Error al cuadrado
0.40066	0.160528
0.04291	0.001841
-0.65794	0.432885
-0.34414	0.118432
0.37051	0.137277
-0.33034	0.109124
0.02656	0.000705
-0.05964	0.003556
0.29726	0.088363
0.25416	0.064597
Suma =	1.117312

- ## b) Suma total de cuadrados

Recordamos que el valor de la Media = 7.55

Y	Desviación	Desviación al cuadrado
5	-2.55	6.5025
5.5	-2.05	4.2025
6	-1.55	2.4025
7	-0.55	0.3025
7.2	-0.35	0.1225
7.7	0.15	0.0225
8.4	0.85	0.7225
9	1.45	2.1025
9.7	2.15	4.6225
10	2.45	6.0025
	Suma =	27.005

- ## c) Suma de cuadrados debida a la regresión

Suma total de cuadrados - STC
Suma de cuadrados debido al error - SCE
Suma de cuadrados debido a la regresión - SCR

$$STC = SCE + SCR$$

Despejando

$$SCR = STC - SCE = 27.005 - 1.117312 = 25.887687$$

- ## d) El coeficiente de determinación

r(al cuadrado) = SCR/STC = 25.887687/27.005 = 0.958625

- ## e) Exprese el significado del coeficiente de determinación encontrado

Se concluye que 95.86% de la variabilidad en la venta de línea blanca se explica por la relación lineal que existe entre la construcción de casas y las ventas de línea blanca.

- ## f) El coeficiente de correlación y su significado

rxy= +raiz(0.958625) = +0.979094

está en un valor entre -1 a 1, esto quiere decir que hay una fuerte relación entre x y y de manera positiva.

5) Calcule los errores estándar de la estimación

Tenemos que:

$$Se = \sqrt{\frac{\Sigma(Y-\hat{Y})^2}{n-2}} = \sqrt{\frac{1.117312^2}{10-2}} = \sqrt{1.5605} = 0.39503$$

Por lo tanto, el error estándar de la estimación es de **0.39503** miles de venta de línea blanca.

6) Los intervalos de confianza

Tomando en cuenta nuestra ecuación de estimación encontrada:

$$\hat{Y} = 1.16834 + 1.7155X$$

Si se considera la construcción de 2 mil casas, tenemos:

$$\hat{Y} = 1.16834 + 1.7155(2)$$
$$\hat{Y} = 4.59 \text{ miles de venta de linea blanca}$$

Supongamos que se desea tener una confianza del **68%** de que la venta de linea blanca está dentro de ± 1de desviación estándar de la desviación de $\hat{Y}$. Los intervalos de confianza son:

$$\hat{Y} + 1Se = 4.59 + (1)(0.39503) = \mathbf{4.99437} \leftarrow$$
$$\textit{límite superior del intervalo de predicción}$$

Y

$$\hat{Y} - 1Se = 4.59 - (1)(0.39503) = \mathbf{4.1949} \leftarrow$$
$$\textit{límite inferior del intervalo de predicción}$$

Si en lugar de esto decidimos que estamos seguros aproximadamente el **95.5%** del tiempo de que la venta real de linea blanca estará dentro de ± 2 errores estándar de la estimación de $\hat{Y}$. Entonces nuestro intervalo de confianza sería de la siguiente manera:

$$\hat{Y} + 2Se = 4.59 + (2)(0.39503) = \mathbf{5.38} \leftarrow$$
$$\textit{límite superior del intervalo de predicción}$$

Y

$$\hat{Y} - 2Se = 4.59 - (2)(0.39503) = \mathbf{3.799} \leftarrow$$
$$\textit{límite inferior del intervalo de predicción}$$

Ahora, si necesitamos tener una seguridad del **97.5%** de que las ventas reales de linea blanca caerán en el intervalo de estimación, utilizamos los valores de la tabla T correspondientes a la columna 0.975 y la fila de dos grados de libertad, siendo t = 4.303

Entonces:

$$\hat{Y} + tSe = 4.59 + (4.303)(0.39503) = \mathbf{6.2898} \leftarrow$$
$$\textit{límite superior del intervalo de predicción}$$

Y

$$\hat{Y} - tSe = 4.59 - (4.303)(0.39503) = \mathbf{2.8902} \leftarrow \textit{límite inferior del intervalo de predicción}$$

7) <u>Aplique la prueba t para determinar si el modelo es estadísticamente significativo</u>

La prueba t consiste en PRUEBA DE t DE SIGNIFICANCIA PARA LA REGRESIÓN LINEAL
Se generan las hipótesis considerando el parámetro β1

$$H0 : \beta\ 1 = 0$$
$$Ha: \beta 1 \neq 0$$

ESTADÍSTICO DE PRUEBA

$$t = b1/sb1$$

REGLA DE RECHAZO

Método de valor p: Rechazar H 0 si valor p <= α

Método de valor crítico: Rechazar H 0 si t <= t α /2 o si t >= t α /2

Donde se toma de la distribución tα/2 con n-2 grados de libertad

Se toma el p valor del resultado que proporcionó el Knime .

Coefficients and Statistics - 0:7 - Linear Regression Learner (Regresion Lineal)

File Edit Hilite Navigation View

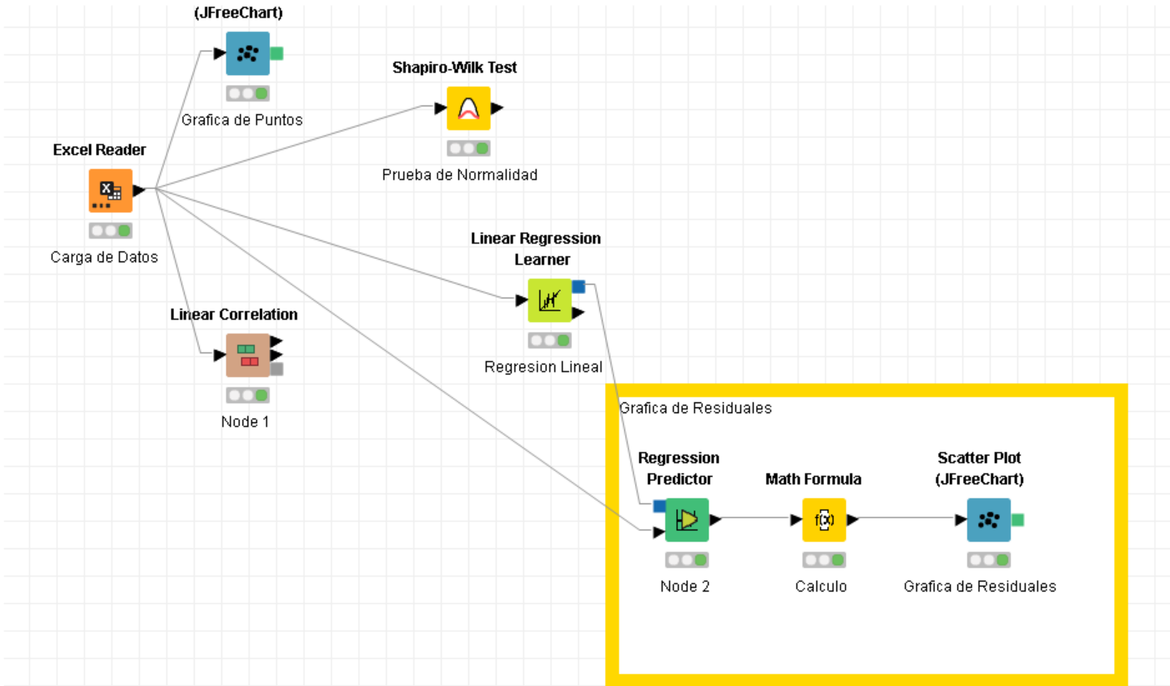
Table "Coefficients and Statistics" - Rows: 2 | Spec - Columns: 5 | Properties | Flow Variables

Row ID	S Variable	D Coeff.	D Std. Err.	D t-value	D P>|t|
Row1	Construcción de casas (miles)	1.716	0.126	13.615	0
Row2	Intercept	1.168	0.483	2.416	0.042

Si α = 0.042 como nivel de significancia, p valor < α se rechaza H0 y se concluye que hay una relación estadísticamente significativa entre el incremento de las casas de la zona y la compra de línea blanca.

La probabilidad de que los valores observados se deban al azar es 0.

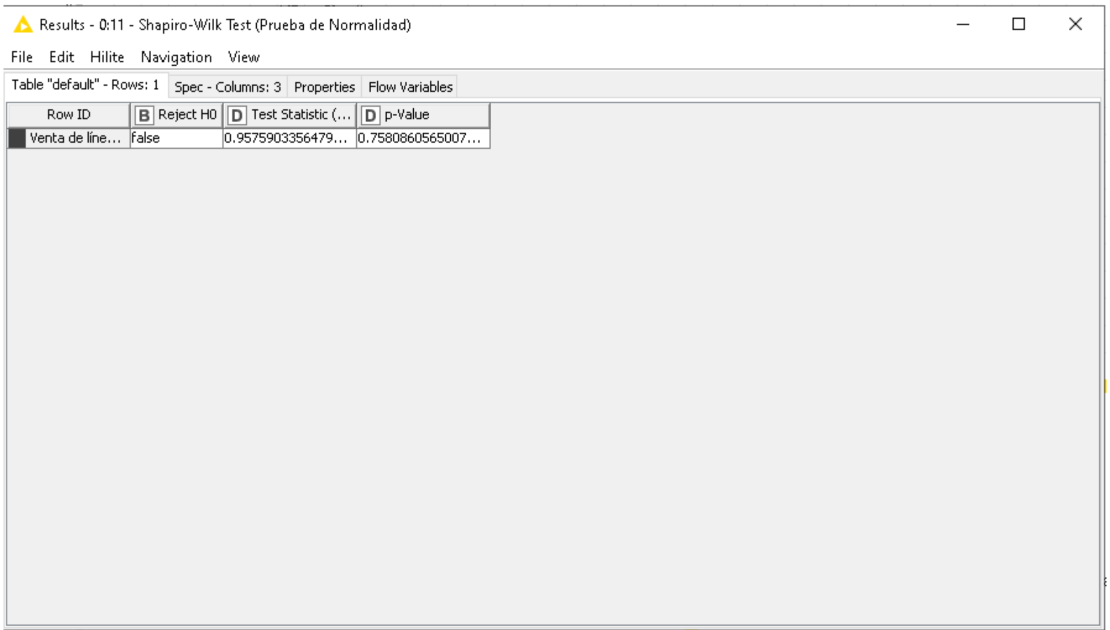
8) Genere la ecuación de recta en el Knime incorporando prueba de normalidad y gráfico de residuales.



- # Ecuación de la Recta

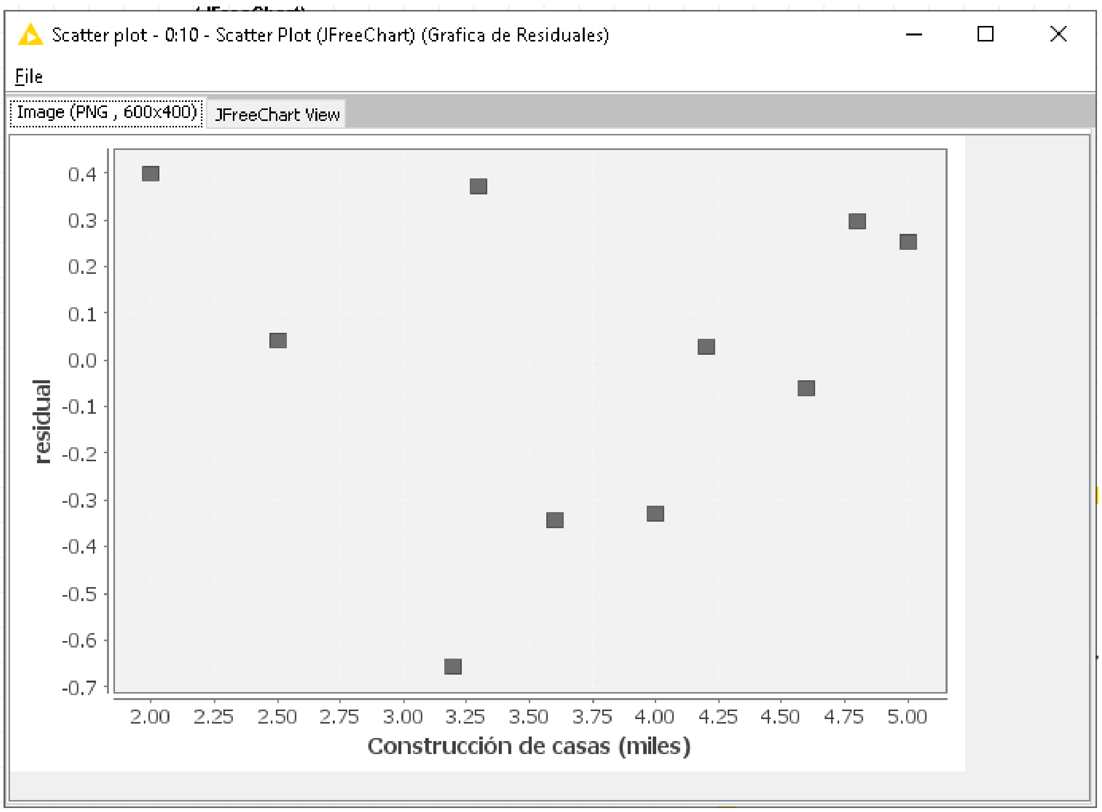
Row ID	Construcción de casas (miles)	Venta de línea blanca (miles)		xi- x'	yi - y'	(xi - x')(yi - y')	(xi- x')2
1	2	5		-1.72	-2.55	4.386	2.9584
2	2.5	5.5		2.5	5.5	13.75	6.25
3	3.2	6		3.2	6	19.2	10.24
4	3.6	7		3.6	7	25.2	12.96
5	3.3	7.2		3.3	7.2	23.76	10.89
6	4	7.7		4	7.7	30.8	16
7	4.2	8.4		4.2	8.4	35.28	17.64
8	4.6	9		4.6	9	41.4	21.16
9	4.8	9.7		4.8	9.7	46.56	23.04
10	5	10		5	10	50	25
	37.2	75.5				290.336	146.1384
	x'	y'					
	3.72	7.55		b1	1.98671944		
				b0	0.15940369		
				Ecuacion de la Recta			
				$y° = 0.1594 + 1.9867x$			

- **Prueba de Normalidad**

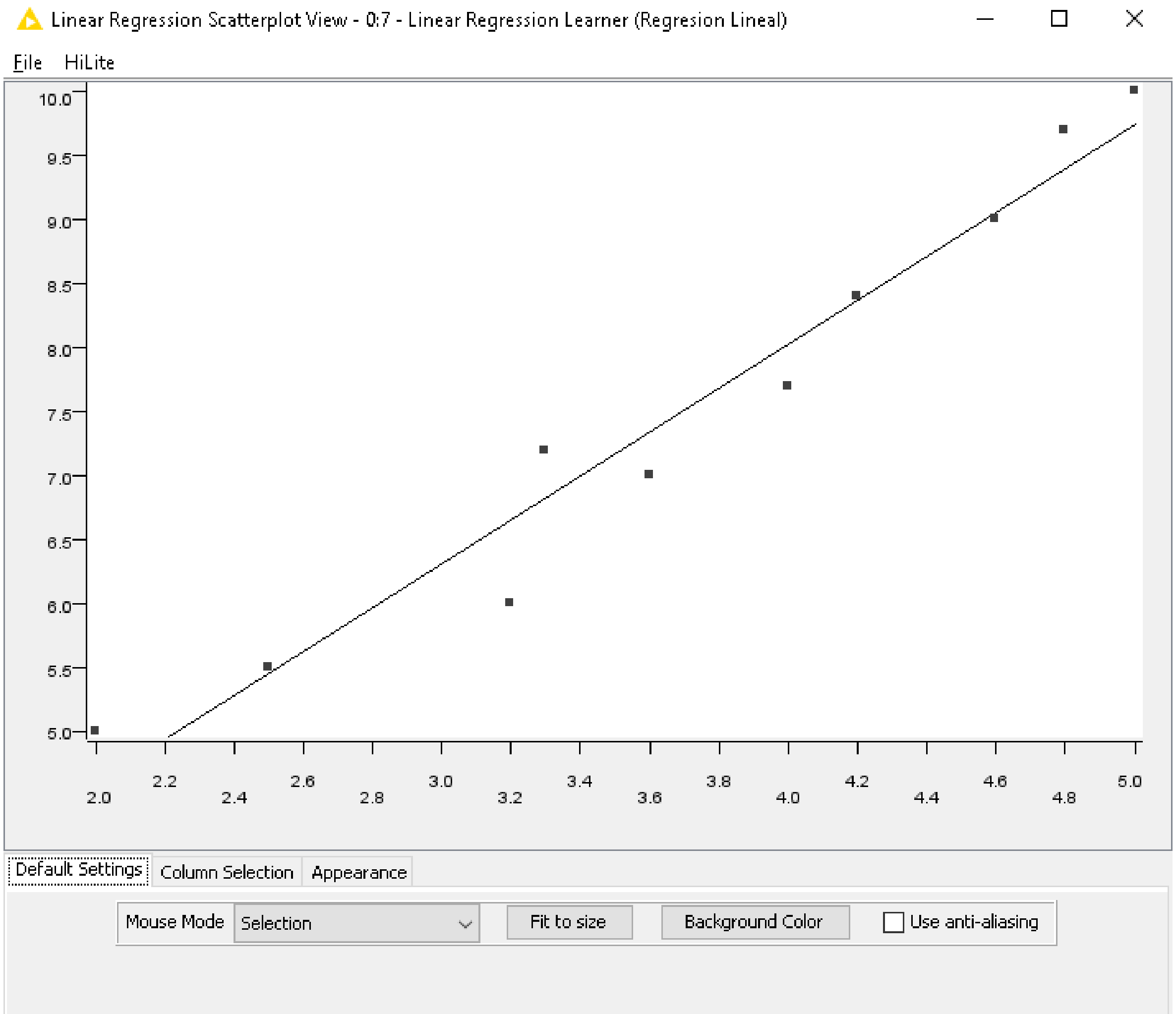


Row ID	B Reject H0	D Test Statistic (...	D p-Value
Venta de líne...	false	0.9575903356479...	0.7580860565007...

- **Grafica de Residuales**



- **Grafica de Regresión Lineal**



9) Conclusiones

En esta tarea se logró el objetivo principal, que fue determinar la relación entre dos variables, y así poder predecir su comportamiento en base a un modelo de regresión lineal.

En este caso, el ejercicio nos menciona dos variables:

- Construcción de casas (variable independiente)
- Venta de línea blanca (variable dependiente)

Primeramente, generamos la gráfica con los valores de las variables, con ayuda de Excel, y pudimos notar que cuentan con una ligera tendencia a elevarse conforme crece la variable independiente (Construcción de casas) lo que nos indica que cuenta con una pendiente positiva.

Posteriormente, realizamos los cálculos paso a paso para generar la ecuación de regresión; Tabulamos los valores de las variables para poder calcular su sumatoria y calculamos la pendiente en base a dos valores promedio, y efectivamente nos mostró una tendencia positiva.

Por último, nos gustaría comentar que gracias a la prueba t realizada pudimos concluir que: La probabilidad de que los valores observados se deban al azar es 0.