INSTITUTO POLITÉCNICO NACIONAL

ESCUELA SUPERIOR DE CÓMPUTO

Materia:

Data Mining

Profesora:

Fabiola Ocampo Botello

Tarea #6:

Regresión Lineal

Equipos #10 y #4:

- Alanís Garduño Mireya
- Ríos Rebollar Víctor Hugo

- García Cruz Octavio Arturo
- Sampayo Hernández Mauro
- Flores Ponce Alan Marcelo

Grupo:

3CV15

EJERCICIO DE REGRESIÓN LINEAL. GRUPO 3CV15

Materia: Minería de datos **Periodo escolar: 2023-2**

Equipo: **4**

Nombre de los integrantes del equipo:

1) Flores Ponce Alan Marcelo

2) García Cruz Octavio Arturo

3) Sampayo Hernández Mauro

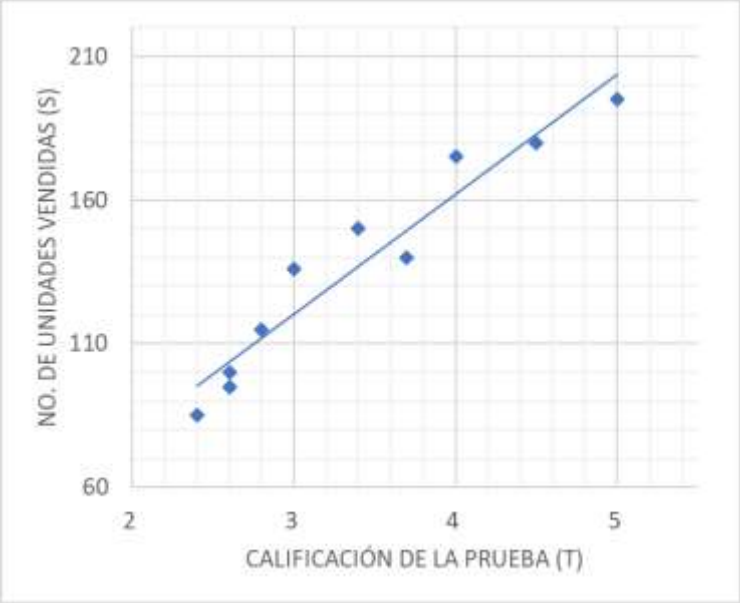
Una compañía aplica a sus vendedores en periodo de capacitación una prueba de ventas antes de salir a trabajar. La administración de la compañía está interesada en determinar la relación entre las calificaciones de la prueba y las ventas logradas por esos vendedores al final de un año de trabajo. Se recolectaron los siguientes datos de 10 agentes de ventas que han estado en el campo un año.
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.

No. Vendedor	Calificación de la prueba (T)	No. De unidades vendidas (s)
1	2.6	95
2	3.7	140
3	2.4	85
4	4.5	180
5	2.6	100
6	5.0	195
7	2.8	115
8	3.0	136
9	4.0	175
10	3.4	150

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.
Las tablas de los cálculos las debe de agregar en este reporte y enviar el archivo respectivo.

1) Generar la gráfica de variables

Calificación de la prueba (T)	**No. de unidades vendidas (S)**
2.6	95
3.7	140
2.4	85
4.5	180
2.6	100
5	195
2.8	115
3	136
4	175
3.4	150



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

Calificación de la prueba (x_i)	No. de unidades vendidas (y_i)	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
2.6	95	-0.8	-42.1	33.68	0.64
3.7	140	0.3	2.9	0.87	0.09
2.4	85	-1	-52.1	52.1	1
4.5	180	1.1	42.9	47.19	1.21
2.6	100	-0.8	-37.1	29.68	0.64
5	195	1.6	57.9	92.64	2.56
2.8	115	-0.6	-22.1	13.26	0.36
3	136	-0.4	-1.1	0.44	0.16
4	175	0.6	37.9	22.74	0.36
3.4	150	0	12.9	0	0
$\sum x_i = 34$	$\sum y_i = 1371$			$\sum (x_i - \bar{x})(y_i - \bar{y}) = 292.6$	$\sum (x_i - \bar{x})^2 = 7.02$

$$\bar{x} = \frac{\sum x_i}{n} = \frac{34}{10} = 3.4 \qquad \bar{y} = \frac{\sum y_i}{n} = \frac{1371}{10} = 137.1$$

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{292.6}{7.02} = 41.681$$

$$b_0 = \bar{y} - b_1 \bar{x} = 137.1 - (41.68)(3.4) = -4.615$$

Ecuación de Regresión: $\hat{y} = -4.615 + 41.681x$

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

Calificación de la prueba (x_i)	**No. de unidades vendidas (y_i)**	$\hat{y} = -4.615 + 41.681x$	**Error** $(y_i - \hat{y}_i)$
2.6	95	103.7556	-8.7556
3.7	140	149.6047	-9.6047
2.4	85	95.4194	-10.4194
4.5	180	182.9495	-2.9495
2.6	100	103.7556	-3.7556
5	195	203.79	-8.79
2.8	115	112.0918	2.9082
3	136	120.428	15.572
4	175	162.109	12.891
3.4	150	137.1004	12.8996
			-0.004

4) Realice los siguientes cálculos (muestre el proceso)

a) Suma de cuadrados debida al error

Calificación de la prueba (x_i)	**No. de unidades vendidas (y_i)**	$\hat{y} = -4.615 + 41.681x$	**Error** $(y_i - \hat{y}_i)$	**Error al cuadrado** $(y_i - \hat{y}_i)^2$
2.6	95	103.7556	-8.7556	76.66053136
3.7	140	149.6047	-9.6047	92.25026209
2.4	85	95.4194	-10.4194	108.5638964
4.5	180	182.9495	-2.9495	8.69955025
2.6	100	103.7556	-3.7556	14.10453136
5	195	203.79	-8.79	77.2641
2.8	115	112.0918	2.9082	8.45762724
3	136	120.428	15.572	242.487184
4	175	162.109	12.891	166.177881
3.4	150	137.1004	12.8996	166.3996802

$$SCE = \sum (y_i - \hat{y}_i)^2 = 961.0652$$

b) Suma total de cuadrados

Calificación de la prueba (x_i)	**No. de unidades vendidas (y_i)**	**Desviación $y_i - \bar{y}$**	**Desviación al cuadrado $(y_i - \bar{y})^2$**
2.6	95	-42.1	1772.41
3.7	140	2.9	8.41
2.4	85	-52.1	2714.41
4.5	180	42.9	1840.41
2.6	100	-37.1	1376.41
5	195	57.9	3352.41
2.8	115	-22.1	488.41
3	136	-1.1	1.21
4	175	37.9	1436.41
3.4	150	12.9	166.41

$$STC = \sum (\boldsymbol{y_i} - \bar{\boldsymbol{y}})^{\boldsymbol{2}} = 13156.9$$

c) Suma de cuadrados debida a la regresión

Calificación de la prueba (x_i)	**No. de unidades vendidas (y_i)**	$\hat{y} = -4.615 + 41.681x$	$(\hat{y}_i - \bar{y})^2$
2.6	95	103.76	1111.85
3.7	140	149.60	156.37
2.4	85	95.42	1737.27
4.5	180	182.95	2102.18
2.6	100	103.76	1111.85
5	195	203.79	4447.56
2.8	115	112.09	625.41
3	136	120.43	277.96
4	175	162.11	625.45
3.4	150	137.10	0.00

$$SCR = \sum (\hat{\boldsymbol{y}}_{\boldsymbol{i}} - \bar{\boldsymbol{y}})^{\boldsymbol{2}} = 12195.89$$

d) El coeficiente de determinación

$$\boldsymbol{r^2} = \frac{\text{SCR}}{\text{STC}} = \frac{12195.89}{13156.9} = 0.92695772$$

e) Exprese el significado del coeficiente de determinación encontrado

Tenemos el 92.69% de bondad de ajuste de la ecuación de regresión estimada.

f) El coeficiente de correlación y su significado

$$\boldsymbol{r_{xy}} = (signo\ de\ b_1)\sqrt{Coeficiente\ de\ determinación} = +\sqrt{0.92695772} = +0.9627345013$$

Existe una correlación positiva, los valores de una variable tienden a incrementarse mientras que los valores de la otra variable descienden, y que por lo tanto existe una relación lineal fuerte entre x y y.

5) Calcule los errores estándar de la estimación

$$\boldsymbol{S_e} = \sqrt{\frac{\boldsymbol{\Sigma(Y - Yi)2}}{\boldsymbol{n - 2}}} = \frac{961.065}{10 - 2} = \sqrt{120.133} = 10.960$$

6) Los intervalos de confianza

Método 1:
Sabiendo que la ecuación de la predicción es:

$$\boldsymbol{\hat{y} = -4.615 + 41.681x}$$

Supongamos 4 interrupciones.

$$\boldsymbol{\hat{y} = -4.615 + 41.681(4)}$$
$$\boldsymbol{\hat{y} = -4.615 + 166.724}$$
$$\boldsymbol{\hat{y} = 162.10}$$

Se estima que la clasificación en la prueba va a tener un valor de 162.10

El error estándar calculado es de $S_e = 10.96$

Para una confianza del 95% es necesario obtener 2 grados de libertad, por lo tanto:

$$\boldsymbol{\hat{y} + 2S_e = 162.10 + (2)(10.96)}$$
$$\boldsymbol{\hat{y} + 2S_e = 162.10 - (2)(10.96)}$$

Por ende, los intervalos de confianza son de 184.02 a 140.18 respecto a la clasificación de la prueba de honestidad

Método 2:
Tomando los valores además de 2 grados de libertad y haciendo uso de Ŷ S la siguiente tabla, e
se harán los cálculos correspondientes.

Grados de Libertad	0.75	0.9	0.95	0.97	0.99	0.995	0.9995
1	1.000	3.078	6.314	12.706	31.821	63.657	636.619
2	0.816	1.886	2.9620	4.303	6.965	9.925	31.598
3	0.765	1.6368	2.353	3.182	4.541	5.841	12.941
4	0.741	1.533	2.132	2.776	3.747	4.604	8.610

Mediante las siguientes fórmulas se van a calcular los intervalos de confianza con una confianza del 0.95 ó 95%

$$\hat{\boldsymbol{y}} + \boldsymbol{t}S_e = \mathbf{162.10} + (\mathbf{2.9620})(\mathbf{10.96})$$
$$\hat{\boldsymbol{y}} + \boldsymbol{t}S_e = \mathbf{162.10} - (\mathbf{2.9620})(\mathbf{10.96})$$

Al resolver las ecuaciones, se obtiene que la clasificación se encontrará 194.563 y 129.636con una confianza del 95%.

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

Se generan las siguientes hipótesis considerando el parámetro β_1

$\boldsymbol{H_0}: \boldsymbol{\rho} = \mathbf{0}$ (No existe relación lineal entre la "calificación de la prueba" y el "no. de unidades vendidas")

$\boldsymbol{H_1}: \boldsymbol{\rho} \neq \mathbf{0}$ (Si la "calificación de la prueba" y el "no. de unidades vendidas" están relacionadas linealmente)

Se plantea el siguiente modelo matemático:

$$\boldsymbol{t}_{(\boldsymbol{n-2})} = \frac{\boldsymbol{r}}{\sqrt{\frac{\mathbf{1}-\boldsymbol{r}^2}{\boldsymbol{n}-\mathbf{2}}}}$$

Sea:

$n = 10$

$r^2 = 0.92695772$

$r = 0.9627345013$

Se tiene:

$$t_{(10-2)} = \frac{0.9627345013}{\sqrt{\frac{1 - 0.92695772}{10 - 2}}}$$

$t_{(8)} = 10.07545$

El valor de $t > 0$, por lo que se puede concluir que **la "calificación de la prueba" y el "no. de unidades vendidas" están relacionadas linealmente**, lo que significa que el modelo es **estadísticamente significativo.**

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

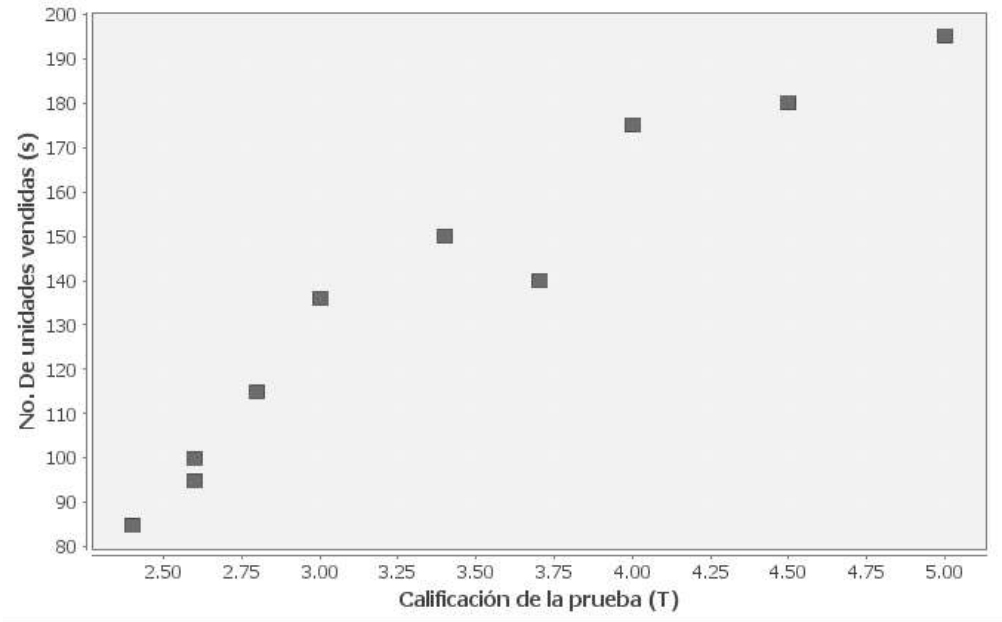


Imagen 1 Gráfica de Variables

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P>|t|
Calificación de la prueba (T)	41.6809	4.1368	10.0757	8.02E-6
Intercept	-4.6151	14.4858	-0.3186	0.7582

R-Squared: 0.927
Adjusted R-Squared: 0.9178

Imagen 2 Resultados de Regresión

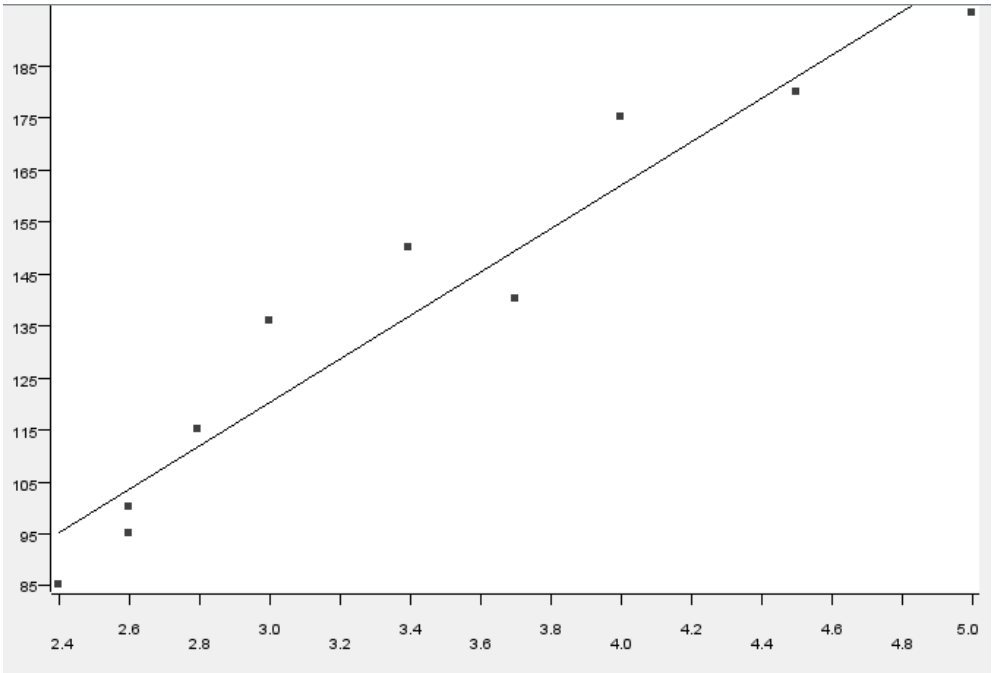


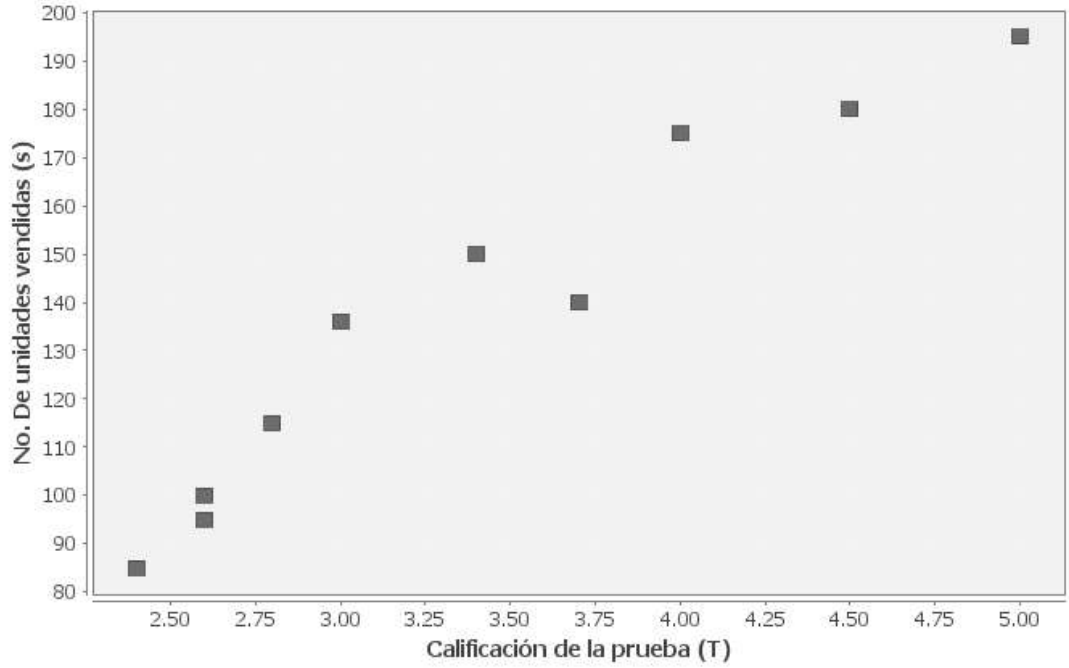
Imagen 3 Ecuación de la Recta

Row ID	B Reject H0	D Test Statistic (W)	D p-Value
Calificación d...	false	0.9214477327954023	0.36918809571677624
No. De unida...	false	0.9467076099252688	0.6297414395733607

Imagen 4 Prueba de Normalidad

Row ID	D Calificación de la prueba (T)	I No. De unidades vendidas (s)	D Prediction (No. De unidades vendidas (s))	D Residuos
Row0	2.6	95	103.755	-8.755
Row1	3.7	140	149.604	-9.604
Row2	2.4	85	95.419	-10.419
Row3	4.5	180	182.949	-2.949
Row4	2.6	100	103.755	-3.755
Row5	5	195	203.789	-8.789
Row6	2.8	115	112.091	2.909
Row7	3	136	120.428	15.572
Row8	4	175	162.109	12.891
Row9	3.4	150	137.1	12.9

Imagen 5 Residuos



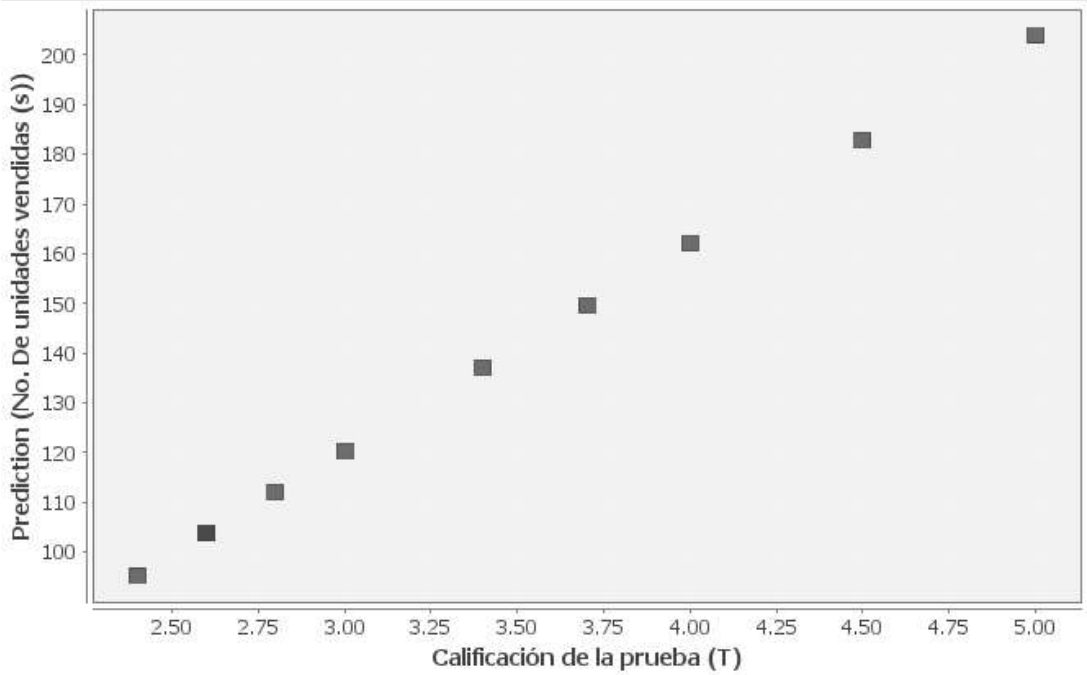


Imagen 6 y 7 Gráficos de Variables y Residuales