Laboratorio N.4

Introduccion a Los Metodos Estadisticos

Prueba de Hipotesis y Regresion

Diana Carolina Arias Sinisterra Cod. 1528008
Kevin Steven Garcia Chica Cod. 1533173
Cesar Andres Saavedra Vanegas Cod. 1628466

Universidad Del Valle
Facultad De Ingenieria
Estadistica

Diciembre
2017

Índice

Índice de figuras

1. Prueba De Hipotesis

1.1. Situación 2

Dado que los salarios se encuentran distribuidos normalmente con media μ y varianza σ^2 $Salarios\ N(\mu = 13.200, \sigma^2 = 1742400)$ Por lo cual se esta acusando a una compania de pagar por debajo del estándar nacional.
Para contrastar esta situación se ha tomado una muestra de 40 trabajadores sobre los cuales se observa un pagro promedio de $\$12000/hora$.

¿Es suficiente esta evidencia para acusar a la empresa?

Teniendo en cuenta que los salarios se encuentran distribuidos normal
$Salarios\} \rightarrow N(\mu = 13.200, \sigma^2 = 1742400)$

Procedemos a plantear la hipotesis
$H_o : \mu = 13.200$
$H_1 : \mu < 13.200$

Donde se tienen los siguientes datos con un nivel de significancia del 0.05:
$n = 40$
$\bar{x} = 12.000$
$Sx = 1320$
$Z_\alpha = Z_{0.05} = -1.64$

Estadistico de Prueba $Z_p = \frac{\bar{x}-\mu}{\sigma^2/\sqrt{n}}$

$Z_p = \frac{12000-13200}{1320/\sqrt{40}}$
$Z_p = -5.7495$

$Z_p < Z_\alpha\} \rightarrow -5.7495 < -1.64$

Conclusion:
Por lo cual, con una significancia del 5 % se rechaza H_o y se acepta por lo tanto la hipotesis H_1 lo que implica suficiente evidencia para acusar a la empresa de realizar pagos inferiores a $13.200.

1.2. Situación 4

Comuna	n	N. Asistentes
A	60	8
B	50	9

Acorde a los resultados obtenidos por la seccretaria de educacion y consignados en la tabla, se procede a plantear la hipotesis con la cual se pueda decidir si es acertado ubicar la escuela en la comuna A.

Procedemos a plantear la hipotesis
$H_o : P_1 - P_2 = 0$
$H_1 : P_1 - P_2 < 0$

La hipotesis nos lleva a tener que evaluar una diferencia de proporciones, con un coeficiente de significancia de 0.05 y para los cuales se tienen los siguientes datos.
$P_A = 8/50 = 0.16$
$P_B = 9/50 = 0.18$
$n_1 = 50$
$n_2 = 60$
$Z_\alpha = Z_{0.05} = 1.645$

Estadistico de Prueba $Z_p = \frac{(P_A - P_B) - d}{\sqrt{(P_A(1-P_A)/n1) + (P_B(1-P_B)/n2)}}$

$Z_p = \frac{(0.16 - 0.18) - 0}{\sqrt{(0.16(1-0.16)/50) + (0.18(1-0.18)/60)}}$
$Z_p = 0.1444$

$\{Z_p < Z_\alpha\} \rightarrow 0.1444 > 1.645$

Conclusion:
Con base en los resultados obtenidos en los cuales nuestro valor $Z_p = 0.1444$ se encuentra en la region de aceptacion, pese a ser un margen de aceptacion bastante bajo es adecuada la decision de la secretaria de eduacion de ubicar la escuela en la comuna A, es decir aceptamos la hipotesis nula y por lo tanto con una significancia de 5 % podemos afirmar que pese a que la diferencia de proporciones de la poblacion es pequeña podemos aceptar la construccion del centro educativo.

1.3. Situación 6

1.4. Punto C.

De acuerdo con la informacion proporcionada procedemos a realizar el planteamiento de la hipotesis, hipotesis con la cual podremos decidir si el peso medio de los niños es inferior a los 3.0KG.

Antes de plantear la hipotesis es necesario realizar un cambio de unidades, puesto que la informacion de la muestra se encuenta en libras y los analisis respectivos que se requieren se van a llevar a cabo en KG. Para lo cual se realiza una regla de tres para encontrar las equivalencias.

Procedemos a plantear la hipotesis
$H_o : \mu = 6$
$H_1 : \mu < 6$

1.5. Punto D.

De los datos dados en la tabla, tenemos que el percentil 35 es 5.0 libras, esto nos dice que el 35 % de los niños tienen un peso menor o igual a 5.0 libras. Por tanto, de la tabla tenemos que el porcentaje de niños de la muestra, que presentan bajo peso al nacer es del 35 % esto es, $\hat{p} = 0.35$.
Planteando las hipótesis tenemos:
H_0: $p = 0.05$
H_1: $p > 0.05$

tenemos que n=50, por tanto, podríamos considerar la muestra como una muestra grande y por ello, podemos aplicar la prueba con la aproximación a la normal:

$$Z = \frac{\hat{P} - P_0}{\sqrt{\frac{P_0(1-P_0)}{n}}} = \frac{0.35-0.05}{\sqrt{\frac{0.05(0.95)}{50}}} = 9.7333$$

La región de rechazo asignando una confianza del 95 % sera: $Rc = \{Z|Z > 1.645\}$

Conclusión:
Como Z=9.7333 es mucho mayor que 1.645, rechazamos H_0 y concluimos que con una confianza del 95 %, el verdadero porcentaje de niños que presentan bajo peso al nacer en esa comunidad es mayor que 5 %.

1.6. Situación 7

1.7. Punto A.

1.8. Punto B.

2. Regresión

2.1. Situación 1

2.2. Punto A.

- Reducción porcentual del total de sólidos: El total de sólidos disueltos (a menudo abreviado como TDS, del inglés: Total Dissolved Solids) es una medida del contenido combinado de todas las sustancias inorgánicas y orgánicas contenidas en un líquido en forma molecular, ionizada o en forma de suspensión micro-granular (sol coloide). Los TDS (Total dissolved solids) son la suma de los minerales, sales, metales, catiónes o aniones disueltos en el agua. Esto incluye cualquier elemento presente en el agua que no sea (H20) molécula de agua pura y sólidos en suspensión. (Sólidos en suspensión son partículas ó sustancias que ni se disuelven ni se asientan en el agua, tales como pulpa de madera.) En general, la concentración de sólidos disueltos totales es la suma de los catiónes (carga positiva) y aniones (cargado negativamente) iones en el agua.

- Reducción porcentual de demanda bioquímica de oxigeno: La demanda bioquímica de oxígeno (DBO) es un parámetro que mide la cantidad de dioxígeno consumido al degradar la materia orgánica de una muestra líquida.

2.3. Punto B.

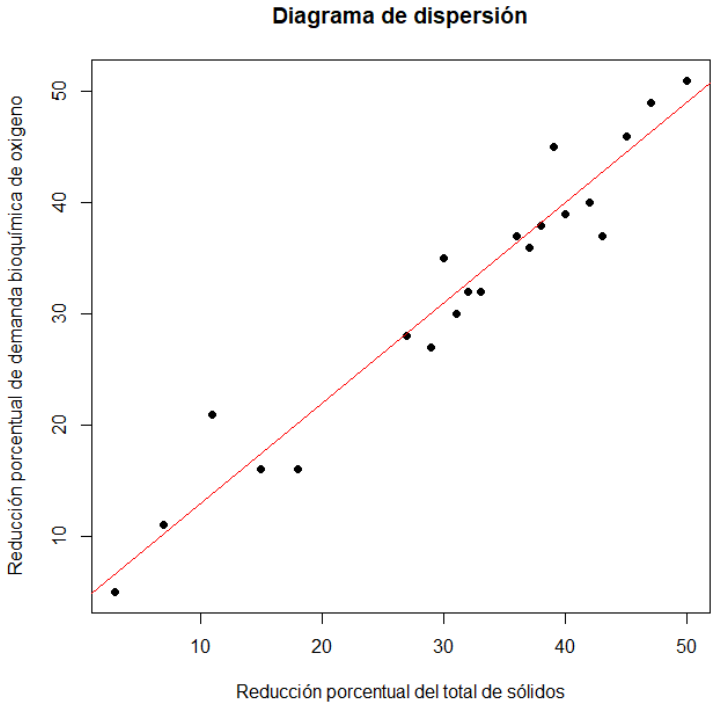

Figura 1: Gráfica de dispersión con recta de ajuste entre las variables X y Y

En esta imagen podemos ver que hay una correlación positiva bastante fuerte, por lo que esperamos que el coeficiente de correlación sea cercano a 1. También podemos ver que las distancias de los datos o los puntos a la recta de ajuste son considerablemente pequeñas, por lo cuál creemos que el R^2 es también cercano a 1, diciéndonos esto, que el modelo ajustado representara en un alto porcentaje la variación total de Y.
Procedemos a calcular los dos valores mencionados:

- Coeficiente de correlación: $\hat{\rho} = r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$

  $\bar{x} = 31.095238$ y $\bar{y} = 31.9523809$, entonces:

$$\hat{\rho} = r = \frac{3202.095238}{\sqrt{3543.809524 \cdot \sqrt{3086.952381}}} = \frac{3202.095238}{3307.502267} = 0.9681$$

- Coeficiente de determinación R^2: $R^2 = r^2 = 0.9372$
  Tal y como esperábamos, el coeficiente de correlación lineal, nos arrojo un resultado de 0.9681 que es bastante alto. Este resultado nos dice que a medida que X aumenta, Y también aumenta en una proporción aproximada de 0.9681. Con respecto al R^2, también nos arrojó

un valor que esperábamos (0.9372) que es bastante alto; este valor nos dice que el 93.72 % de la variabilidad total de la variable Y, es explicada por la variable X. Entonces, según lo anterior, concluimos que estas dos variables tienen una relación positiva demasiado fuerte, es decir, cuando aumenta X, Y también aumentara, y por el R^2 concluimos que X es una buena variable explicativa para Y.

2.4. Punto C.

Para evaluar la asociación lineal entre las variables X y Y, tendremos que plantear las hipótesis sobre ρ de la siguiente manera:

H_0: $\rho = 0$ (No hay correlación lineal)
H_1: $\rho \neq 0$ (Hay correlación lineal)

$T_\rho = \frac{r-\rho}{\sqrt{\frac{1-r^2}{n-2}}}$
Reemplazando los valores, teniendo en cuenta que por el punto anterior $r = 0.9681$, tenemos:
$T_\rho = \frac{0.9681}{\sqrt{\frac{1-(0.9681)^2}{21-2}}} = 16.84139$
Tomando $\alpha = 0.05$, $T_{(0.975;19)} = 2.093$
Como $T_\rho = 16.84139 > 2.093$, rechazamos H_0 y concluimos que con una confianza del 95 %, si existe correlación lineal entre las dos variables.

2.5. Punto D.

El modelo ajustado es de la forma: $Y_i = \beta_0 + \beta_1 X_i + e_i$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2} = \frac{24067 - 21(31.095238)(31.9523809)}{23849 - 21(31.095238)^2} = \frac{3202.095336}{3543.809648} = 0.903574$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 31.9523809 - (0.903574 \cdot 31.095238) = 3.855532$$

El modelo ajustado queda: $Y_i = 3.855532 + 0.903574 X_i + e_i$

INTERPRETACIÓN:
$\hat{\beta}_0 = 3.855532$: Sin tener en cuenta la variabilidad de la reducción porcentual del total de sólidos, se espera una reducción porcentual de demanda bioquímica de oxigeno de 3.855532.

$\hat{\beta}_1 = 0.903574$: Por cada unidad adicional de la reducción porcentual del total de sólidos, se espera que la reducción porcentual de demanda bioquímica de oxigeno aumente en promedio en 0.903574 unidades.

2.6. Punto E.

Los indicadores utilizados para evaluar la bondad de ajuste de un modelo son el ρ y el R^2, como ya los obtuvimos en el punto b, vamos a interpretarlos.

$\rho = 0.9681$: Este valor nos indica que existe una correlación positiva bastante fuerte (casi perfecta), es decir, cuando x aumenta, y aumenta casi en la misma proporción. Cuando tenemos una correlación tan alta entre las dos variables, podemos estar seguros de que el modelo ajustado será un buen modelo (R^2 tendiendo a 1) para explicar Y en términos de X.

$R^2 = 0.9372$: Este valor nos dice que el 93.72 % de la variabilidad total de variable Y(reducción porcentual de demanda bioquímica de oxigeno) es exlicada por la variable X(reducción porcentual del total de sólidos). Lo que en el fondo nos dice que el modelo es bastante bueno, ya que las distancias de los datos o los puntos a la recta de regresión ajustada, son muy pequeñas.

2.7. Punto F.

Para β_0**:** $\langle \beta_0 \rangle_{(1-\alpha)\,\%} = \langle \hat{\beta}_0 \pm t_{(\frac{\alpha}{2}, n-2)} \sqrt{V(\hat{\beta}_0)} \rangle$

$V(\hat{\beta}_0) = \frac{\hat{\sigma^2} \sum_{i=1}^{n} {x_i}^2}{nS_{xx}}$

$\hat{\sigma^2} = \frac{1}{n-2}[S_{yy} - \hat{\beta}_1 S_{xy}] = \frac{1}{21-2}[3086.952381 - (0.903574 \cdot 3202.095238)] = \frac{1}{19}(193.6223784) = 10.1906515$

Entonces, reemplazando: $V(\hat{\beta}_0) = \frac{10.196515 \cdot 23849}{21 \cdot 3543.809524} = 3.265746$

Ahora, con $\alpha = 0.05$:

$$\langle \beta_0 \rangle_{0.95\,\%} = \langle 3.855532 \pm 2.093 \cdot \sqrt{3.265746} \rangle$$

$$(0.073193; 7.6378708)$$

Para β_1**:** $\langle \beta_1 \rangle_{(1-\alpha)\,\%} = \langle \hat{\beta}_1 \pm t_{(\frac{\alpha}{2}, n-2)} \sqrt{V(\hat{\beta}_0)} \rangle$

$V(\hat{\beta}_1) = \frac{\hat{\sigma^2}}{S_{xx}}$
Ya sabemos que $\hat{\sigma^2} = 10.1906515$

Entonces: $V(\hat{\beta}_1) = \frac{10.1906515}{3543.809524} = 0.00287562$

Ahora, con $\alpha = 0.05$:

$$\langle \beta_1 \rangle_{0.95\,\%} = \langle 0.903574 \pm 2.093 \cdot \sqrt{0.00287562} \rangle$$

$$(0.791337; 1.0158107)$$

2.8. Punto G.

ESTIMACIÓN PUNTUAL:
$E[Y|X] = E[\beta_0 + \beta_1 x1 + e] = E[\beta_0] + E[\beta_1 x_1] + E[e]$, sabemos que $E[e] = 0$, por los supuestos del modelo lineal.
Entonces: $E[\hat{Y}|X] = \hat{\beta}_0 + \hat{\beta}_1 x_1$
Por consiguiente: $E[Y|X = 30] = \hat{\beta}_0 + \hat{\beta}_1 \cdot 30 = 3.855532 + 0.903574(30) = 30.962752$

El valor esperado de la reducción de la demanda bioquímica de oxigeno cuando se reduce a 30 % el porcentaje total de sólidos es de 30.962752 %.

ESTIMACIÓN POR INTERVALOS:
No tenemos formula para la estimación or intervalos de la esperanza de Y dado un valor $X = x_0$, esto es $E[Y|X = x_0]$ por lo cual, trataremos de construirlo, como un intervalo para la media, esto seria:
$\langle E[Y|X = x_0]\rangle_{1-\alpha\,\%} = \langle E[Y|\hat{X} = x_0] \pm t_{(\frac{\alpha}{2};n-2)} \cdot \sqrt{V(E[Y|\hat{X} = x_0])}\rangle$

Debemos hallar $V(E[Y|\hat{X} = x_0])$, ya que no la conocemos.

Sabemos que: $E[Y|\hat{X} = x_0] = \hat{\beta}_0 + \hat{\beta}_1 x_0$, como $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

$$E[Y|\hat{X} = x_0] = \bar{Y} - \beta_1 \bar{X} + \hat{\beta}_1 x_0$$

$$E[Y|\hat{X} = x_0] = \bar{Y} + \hat{\beta}_1 (x_0 - \bar{X})$$

Ahora, procedemos a calcular lo que nos interesa:

$$Var(E[Y|\hat{X} = x_0]) = Var(\bar{Y}) + Var[\hat{\beta}_1 (x_0 - \bar{X})]$$

$$= Var(\bar{Y}) + (x_0 - \bar{X})^2 Var(\hat{\beta}_1)$$

como $V(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{S_{xx}}$ y $V(\bar{Y}) = \frac{1}{n}\hat{\sigma}^2$, reemplazando nos queda:

$$Var(E[Y|\hat{X} = x_0]) = \frac{1}{n}\hat{\sigma}^2 + (x_0 - \bar{X})^2 \cdot \frac{\hat{\sigma}^2}{S_{xx}}$$

$$= \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{X})^2}{S_{xx}} \right)$$

Entonces, nuestro intervalo de confianza construido para el valor esperado de Y dado un valor $X = x_0$ quedara:
$\langle E[Y|X = x_0]\rangle_{(1-\alpha)\,\%} = \langle E[Y|\hat{X} = x_0] \pm t_{(\frac{\alpha}{2};n-2)}\sqrt{(\frac{1}{n} + \frac{(x_0-\bar{X})^2}{S_{xx}}) \cdot \hat{\sigma^2}}\rangle$

Ahora, volviendo a nuestro problema, tenemos los siguientes datos: $E[Y|\hat{X} = 30] = 30.962752$, $\hat{\sigma^2} = 10.1906515$, $t_{(\frac{\alpha}{2},19)} = 2.093$, $\bar{X} = 31.095238$ y $S_{xx} = 3543.809524$
Reemplazando en nuestro intervalo, nos queda:
$\langle E[Y|X = 30]\rangle_{95\,\%} = \langle 30.962752 \pm 2.093\sqrt{(\frac{1}{21} + \frac{(30-31.095238)^2}{3543.809524}) \cdot 10.1906515}\rangle$

$$= \langle 30.962752 \pm 2.093(0.694132321)\rangle$$

$$= (29.5099; 32.4155)$$

2.9. Situación 2

2.10. Punto A.

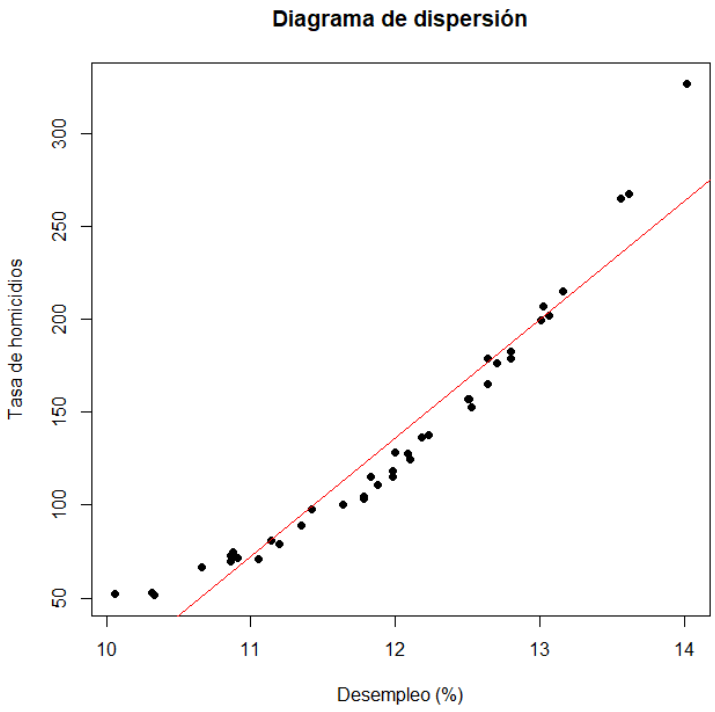

Figura 2: Gráfica de dispersión con recta de ajuste entre las variables X(desempleo) y Y(Tasa de homicidios)

En la imagen anterior, observamos que entre el desempleo(variable X) y la tasa de homicidios(variable Y), existe una correlación positiva bastante alta (por la forma de la distribución de los puntos), es decir, a valores altos de X, le corresponden valores altos de Y. También, si observamos la recta de regresión o de ajuste, podemos ver que las distancias en general de cada punto a la recta no son muy amplias, por lo que creemos que el coeficiente de determinación R^2 también sera bastante alto (cercano a 1).
Para corroborar las hipótesis que tenemos del punto anterior, hallaremos el indice de correlación ρ y el coeficiente de determinación R^2.
Procedemos a calcular los dos valores mencionados:

- Coeficiente de correlación: $\hat{\rho} = r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\cdot\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$

  $\bar{x} = 11.977$ y $\bar{y} = 134.61325$, entonces:

$$\hat{\rho} = r = \frac{2336.75449}{\sqrt{36.65424 \cdot \sqrt{161368.7339}}} = \frac{2336.75449}{2432.046114} = 0.9608$$

- Coeficiente de determinación R^2: $R^2 = r^2 = 0.92317$

Podemos ver con los valores obtenidos que nuestras hipótesis son ciertas. El coeficiente de correlación nos arrojo un resultado de 0.9608, es decir, que existe una correlación positiva bastante fuerte (casi perfecta), esto es, cuando la variable x aumenta en una unidad, la variable y aumenta casi que en la misma proporción. El coeficiente de determinación nos arrojo un resultado bastante alto también, de 0.92317, esto quiere decir que el modelo que ajustemos posteriormente con dicha variable x(desempleo), explicara el 92.317 % de la variación total de y(tasa de homicidios).

Procederemos a ajustar el modelo:
El modelo ajustado es de la forma: $Y_i = \beta_0 + \beta_1 X_i + e_i$

$$\hat{\beta_1} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} {x_i}^2 - n\bar{x}^2} = \frac{66827.2703 - 40(11.977)(134.61325)}{5774.5954 - 40(11.977)^2} = \frac{2336.75449}{36.65424} = 63.7512738$$

$$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x} = 134.61325 - (11.977 \cdot 63.7515738) = -628.935756$$

El modelo ajustado queda: $Y_i = -628.935756 + 63.7512738 X_i + e_i$

2.11. Punto B.

Para estimar la tasa de homicidios cuando disminuya la tasa de desempleo al 11 %, debemos hallar o estimar el valor esperado de Y dado que X tome el valor de 11, esto es, $E[Y|\hat{X} = 11]$:
$E[Y|\hat{X} = 11] = -628.9357563 + 63.7512738(11) = 72.3282555$
En conclusión, la tasa de homicidios cuando el desempleo sea del 11 % es de aproximadamente 72.3282 casos por 100.000 habitantes.

2.12. Punto C.

Para la realización del informe con las conclusiones mas importantes, debemos tener en cuenta la interpretación de los valores encontrados en el literal a.

- $\hat{\beta_0} = -628.9357563$: Sin tener en cuenta la variación la tasa de desempleo, se espera una tasa de homicidios de -628.9357563 casos por cada 100.000 habitantes, obviamente esta variable no toma valores negativos, entonces decimos que cuando el desempleo es 0, el numero de homicidios tambien sera 0.
- $\hat{\beta_1} = 63.7512738$: Cuando la tasa de desempleo aumente en una unidad (en 1 %) la tasa de homicidios aumentara 63.7512 casos por cada 100.000 habitantes. En la interpretación de este coeficiente podemos ver la importancia del desempleo en la tasa de homicidios, ya que, con tan solo una unidad de aumento en el desempleo, la tasa de homicidios aumenta demasiado.

- Con respecto a la correlación entre estas dos variable, vemos que tienen una correlación lineal casi perfecta ($\rho = 0.9608$), esto nos dice que cuando X(tasa de desempleo) aumenta, la variable Y(Tasa de homicidios) tambien aumenta. Y, el coeficiente de determinación R^2, que nos arrojo un resultado de 0.92317, nos dice que el 92.317 % de la variación total de la variable Y(Tasa de homicidios), es explicada por la tasa de desempleo (variable X). Entonces concluimos que esta variable (Desempleo) es muy influyente e importante en la tasa de homicidios.

- Conclusión: Según todo lo anterior, podríamos concluir que para disminuir un poco la problemática de la tasa de homicidios en la comunidad estudiada, se deben enfocar en disminuir primero la tasa de desempleo, ya que una conlleva a la otra (una disminución en la tasa de desempleo causa una disminución considerable en la tasa de homicidios). Entonces una posible solución indirecta a la tasa de homicidios, es por ejemplo, aumentar el numero de empleos.