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 1

1.2. Situación 2

1.3. Punto A.

1.4. Punto B.

1.5. Situación 3

1.6. Punto A.

1.7. Punto B.

1.8. Situación 4

1.9. Situación 5

1.10. Punto A.

1.11. Punto B.

1.12. Punto C.

1.13. Situación 6

1.14. Punto C.

1.15. Punto D.

1.16. Situación 7

1.17. Punto A.

1.18. Punto B.

2. Regresion

2.1. Situación 1

2.2. Punto A.

- Reduccion porcentual del total de solidos: 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.

- Reduccion porcentual de demanda bioquimica 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.

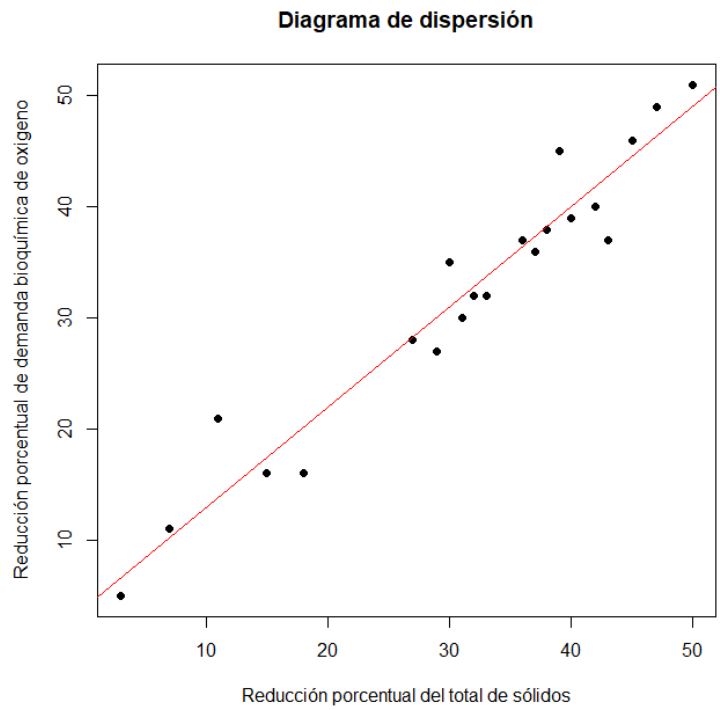

Figura 1: Gráfica de dispersión con recta de ajuste entre las variables X y Y

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 esperabamos, 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 tambien aumenta en una proporción aproximada de 0.9681. Con respecto al R^2, también nos arrojó

un valor que esperabamos (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 tambien aumentara, y por el R^2 concluimos que X es una buena variable explicativa para Y.

2.4. Punto C.

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

H_0: $\rho = 0$ (No hay correlacion lineal)
H_1: $\rho \neq 0$ (Hay correlacion 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 correlacion 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: $Y_i = 3.855532 + 0.903574X_i + e_i$

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

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

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

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.

2.11. Punto B.

2.12. Punto C.