Universidad de Chile
Facultad de Ciencias Agronómicas

PRACTICO REGRESIÓN LINEAL SIMPLE

El Medidor Portátil de Clorofila, es un equipo manual que permiten evaluar un índice (CCI) construido a partir del cociente de la transmitancia de dos longitudes de onda, previamente determinadas través de las hojas y medidas por este equipo. Se cree que este índice puede ayudar a predecir el contenido de clorofila en las plantas, por lo que se procedió a evaluar la relación existente entre los resultados entregados por el Medidor Portátil de Clorofila (quien entrega los resultados en Unidades CCI) con los resultados del contenido de clorofila total entregados por la metodología tradicional del laboratorio (medición entregada en *mg* / cm^2), ambas variables medidas en la misma hoja.

Ìndice CCI (Equipo manual)	Corofila Total (Laboratorio)
42,3	526,310
40,0	488,093
41,4	506,299
29,9	454,636
10,2	220,067
6,1	157,918
16,7	281,353
4,7	121,391
33,6	486,144
18,3	299,082
23,8	388,218
12,0	259,486
18,8	336,444
45,1	570,412
26,1	398,102

1. ¿Cuál es la variable dependiente (y) y cual es la variable independiente (x) en este caso?

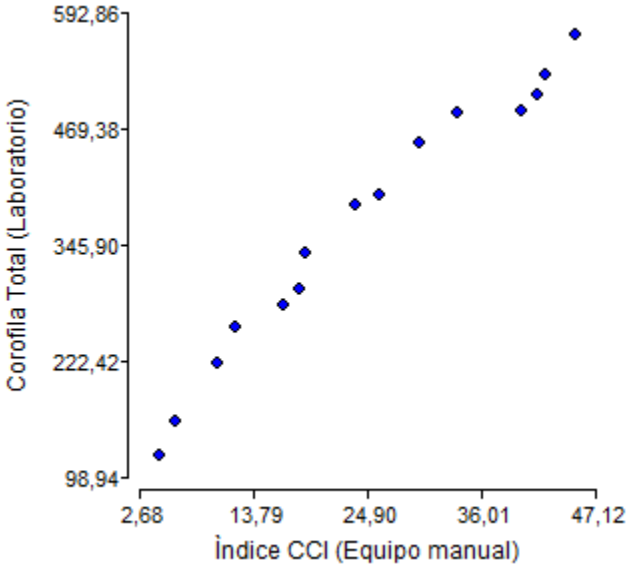
Lo que nos interesa estimar es el contenido real de clorofila en función de la medición que nos entrega el equipo CCI, por lo tanto, la variable dependiente en este caso es la Clorofila total (laboratorio) y la variable independiente es el índice CCI. Este modelo que propondremos es conocido como modelo de regresión lineal simple, ya que queremos predecir 1 variable dependiente en función de una variable independiente.

2. Realice un diagrama de dispersión, asignando la variable respuesta al eje “y” y la variable independiente al eje “x”.

La razón de realiza un gráfico de dispersión entre ambas variables es determinar si visualmente la clorofila total y el índice CCI se ajustan a un modelo de regresión lineal simple. Para que ambas variables sean buenas candidatas a ser modeladas de acuerdo a esa metodología, debería observar una tendencia rectilínea de las observaciones, ya sea de manera directamente proporcional (si aumenta el índice CCI, aumenta la clorofila total) o inversamente proporcional (si aumenta el índice CCI, disminuye la clorofila total). Lo anterior es fundamental en un análisis de regresión, ya que si no observo esa tendencia en los datos el ajuste realizado sobre ellos no sería el adecuado.

PRACTICO REGRESIÓN LINEAL SIMPLE

Para realizar lo anterior debemos ir a gráficos-diagrama de dispersión:
En el eje y: agregamos la variable dependiente (clorofila total).
En el eje x: agregamos la variable independiente (índice CCI).



Los resultados del gráfico sugieren que a medida que aumentan los valores del índice CCI, aumentan los valores de clorofila total y que este aumento es de tendencia rectilineal. Por lo tanto, ambas variables serían buenas candidatas a ser modeladas mediante un modelo de regresión lineal simple y que una recta estimada, que pase sobre los puntos es una simplificación bastante buena.

3. Identifique el modelo que relaciona la variable dependiente con la variable independiente.

El modelo que se puede proponer en este caso es un modelo de regresión lineal simple, en el cual la variable dependiente es la clorofila total y la variable independiente es el índice CCI.

4. Especifique matemáticamente el modelo lineal a utilizar, especificando cada uno de sus terminos.

$$y_{ij} = \beta_0 + \beta_1 x_i + \varepsilon_{ij} \qquad (1)$$

Donde:
y_{ij}: corresponde a la variable dependiente, bajo el i-ésimo nivel de x en la j-ésima unidad experimental.
x_i: corresponde al i-ésimo nivel de variable independiente.
β_0: corresponde al parámetro que representa la ordenada al origen.
β_1: corresponde al parámetro que representa la pendiente.

PRACTICO REGRESIÓN LINEAL SIMPLE

ε_{ij}: corresponde al error experimental, el cual se asume independiente, con media 0 y varianza σ^2.

El modelo podría haber sido especificado como:

$$y_{ij} = \alpha + \beta x_i + \varepsilon_{ij} \quad (2)$$

Simplemente utilizando otras letras. De aquí en adelante utilizaremos el formato 1. Cabe destacar que la presentación de este modelo se realiza a través de letras griegas, ya que se considera que se está trabajando con la población completa de los datos y no con una muestra de ella. La inclusión de un componente aleatorio ε indica una variación no explicada (aleatoria) en el modelo, lo cual expone que a pesar de ajustar una recta entre ambas variables, será imposible que esa recta pase a través de todos los pares de puntos. Como no contamos con los datos de la población, pero si con una pequeña muestra (15 pares de datos) deberemos "estimar" esos parámetros.

Una vez estimado el modelo, este será especificado de la siguiente manera[1]:

$$\hat{y} = \hat{\beta}_o + \hat{\beta}_1 x \quad (3)$$

Donde:
$\hat{y}$: corresponde al valor predicho de la clorofila.
$\hat{\beta}_o$: corresponde a la estimación de la ordenada al origen.
$\hat{\beta}_1$: corresponde a la estimación de la pendiente.
x: corresponde a la variable independiente, en este caso el índica CCI.

Para poder estimar los parámetros a partir de los datos observados, utilizaremos una técnica conocida como "Estimación por mínimos cuadrados". El método de los mínimos cuadrados es una de las alternativas que existen al momento de estimar los parámetros del modelo que debemos proponer. Como somos incapaces de trabajar con toda la población de datos, trabajaremos en base a nuestra muestra aleatoria con n observaciones (15 datos), procediendo a estimar, a través de estas muestras los parámetros del modelo teórico. Llamaremos $\hat{\beta}_0$ a la estimación de la ordenada al origen (β_0) y $\hat{\beta}_1$ a la estimación de la pendiente (β_1), tal como se resume en la fórmula 3. Luego la curva estimada se puede escribir de la siguiente forma:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \quad (4)$$

[1] En algunos textos puede aparecer el modelo estimado con letras latinas de la siguiente manera: $\hat{y} = b_o + b_1 x$ o incluso $\hat{y} = a + bx$. En ambos casos, junto con la ecuación 3 los modelos son equivalentes. En esta guía usaremos los valores griegos con "sombrero" para el modelo estimado y no la letra griega "a secas" las cuales representan los parámetros del modelo teórico.

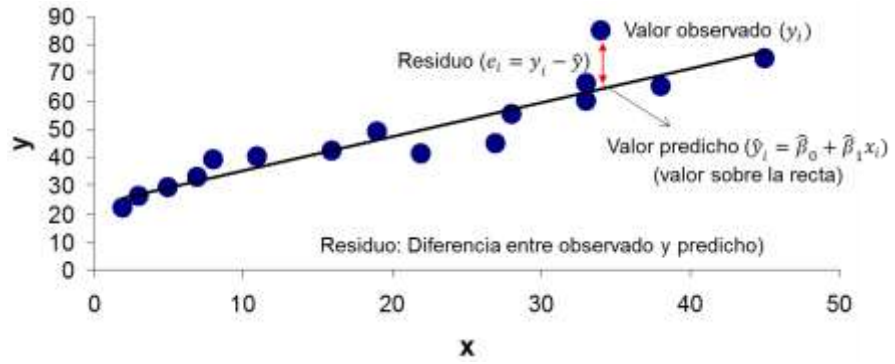
PRACTICO REGRESIÓN LINEAL SIMPLE

Donde $\hat{y}_i$ es el valor predicho del modelo estimado para un cierto valor de x_i. Si evaluamos la discrepancia que existe entre lo que observamos y lo que predice nuestro modelo, dado un valor de la variable independiente *x*, podemos representar esa expresión de acuerdo a la siguiente forma:

$$e_i = y_i - \hat{y} \qquad (5)$$

Donde e_i corresponde al i-ésimo residuo del modelo.

Graficamente los valores observados, esperados (predichos) y residuos se puede resumir de la siguiente manera:



Los valores observados son todos los puntos en el gráfico. En este caso se muestra como referencia de un valor observado el punto sobre la flecha roja, su valor predicho, el que está bajo la flecha roja y el residuo la diferencia de ambos valores.

5. Corra el modelo en Infostat y en la ventana Diagnóstico guarde los Residuos y los Predichos. Identifique en el gráfico los valores observados, los predichos y los residuos.

Para realizar lo anterior debemos ir a *Estadísticas-Regresión lineal*:
En *variable dependiente*: agregamos la variable clorofila total.
En *variable independiente*: agregamos la variable índice CCI.
Luego en *Diagnóstico* marcamos *Residuos* y *predichos*. Aprovecharemos de marcar en *Graficar…*
REstud vs predichos (está marcada por defecto)
Ajuste (está marcada por defecto)
QQ-plot
Luego *Aceptar.*

PRACTICO REGRESIÓN LINEAL SIMPLE

(Nótese que por defecto se utiliza como técnica de estimación el método de los mínimos cuadrados).

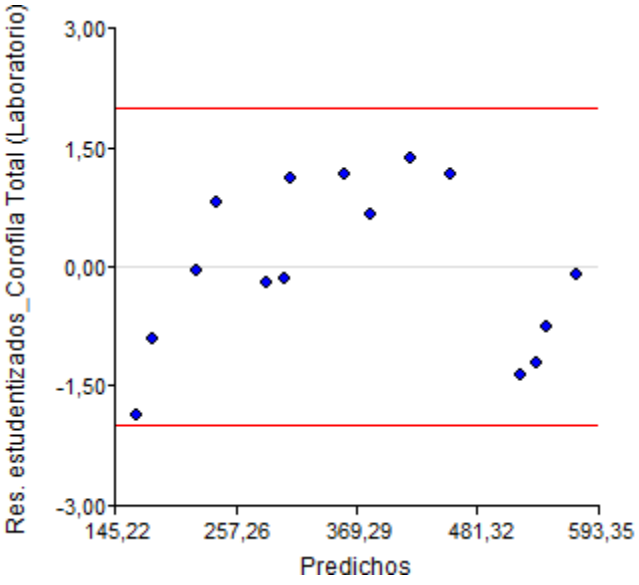
6. Verifique los supuestos del modelo indicando detalladamente las hipótesis que interesa contrastar en cada supuesto.

Los supuestos del modelo, escritos en orden de importancia corresponden a:

a) Independencia de los errores.
b) Homogeneidad de varianza.
c) Normalidad de los errores.

La importancia de su verificación radica en que la eficiencia de la estimación por minimos cuadrados se basa en que se cumplan estos supuestos de manera simultánea. Para asegurarnos de que se cumpla el primer supuesto, debemos asignar los valores de la variable independiente de forma aleatoria a las unidades experimentales. En el caso de no poder asignar esos valores aleatoriamente (como es nuestro caso, ya que se basa en la selección de hojas que arrojan diferentes valores de CCI y clorofila) debemos asigurarnos que la muestra se realice aleatoriamente, tratando de abarcar una superficie lo suficientemente representativa.

El segundo supuesto asume que, independiente de que aumenten los valores predichos, la variabilidad de los residuos debe mantenerse constante. Lo anterior se puede verificar mediante un gráfico de residuos versus predichos (ambos valores obtenidos previamente en el punto 5) o utilizar el gráfico de Residuos estudentizados (*REstud*) *vs* predichos (también obtenidos previamente en el punto 5). La interpretación de este grafico se realiza en función de la dispersión que tienen los puntos. No debería observar ningún patrón en particular, sino una nube de puntos completamente al azar.



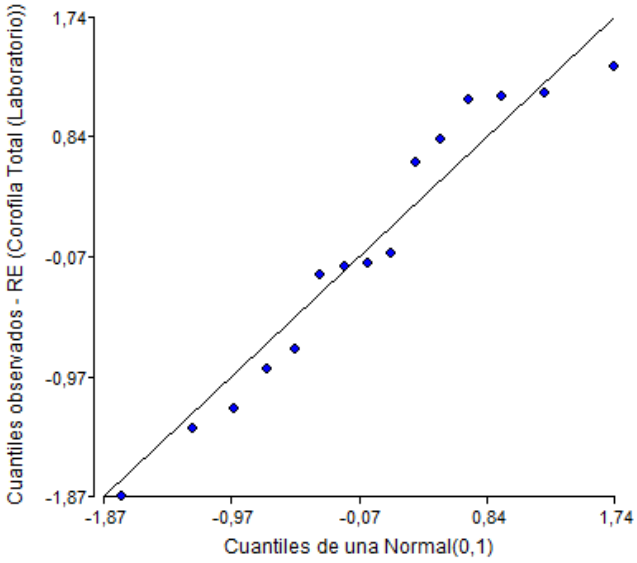
PRACTICO REGRESIÓN LINEAL SIMPLE

Aparentemente no se observa un patrón regular en la nube de puntos, por lo tanto podemos asumir que el supuesto se cumple.

Las hipótesis a contrastar corresponden a:

$$H0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2 = \cdots = \sigma_i^2$$
$$HA: al\ menos\ 1\ \sigma_i^2 \neq \sigma_j^2$$

La verificación del supuesto de normalidad de los errores se podría realizar a través de la construcción de un histograma, pero en este caso utilizando los residuos del modelo. Si el supuesto se satisface, deberíamos observar la forma de una distribución normal con centro en cero. Lamentablemente, cuando hay pocas observaciones esta estrategia no es la más adecuada, por lo que la alternativa consiste en construir un gráfico QQ plot normal. Utilizaremos el QQ plot normal obtenido en el punto 5.



Lo que debemos observar en este gráfico, si se cumple el supuesto de distribución normal, es que los puntos, que corresponden a los residuos estén lo más cercano a la recta. En este caso aparentemente se cumple lo anterior, por lo que podemos suponer que el supuesto de normalidad de los errores se cumple.

Las hipótesis a contrastar corresponden a:

$H0: \varepsilon \sim Normal$ (Los errores se distribuyen de forma nomal)
$HA: \varepsilon\ no \sim Normal$ (Los errores no se distribuyen de forma normal)

PRACTICO REGRESIÓN LINEAL SIMPLE

7. Estimar los parámetros del modelo e interprételos.

Utilizaremos los resultados obtenidos en el punto 5.

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	118,20	14,61	86,63	149,76	8,09	<0,0001		
Índice CCI (Equipo manual)..	10,08	0,52	8,95	11,21	19,26	<0,0001	371,01	1,00

La estimación de la ordenada al origen $\hat{\beta}_0$ es 118,2 (const) y la estimación de la pendiente $\hat{\beta}_1$ es 10,08 (Índice CCI), tal como se verifica en el ejercicio resuelto "a mano" en el punto 4.

La interpretación de la ordenada al origen es: si los valores de la variable independiente son iguales a 0 (lo que mido con el equipo me entrega valores iguales a 0) el contenido de clorofila en promedio será 118,2 mg/cm^2. Este resultado sólo es valido si en la muestra he medido valores iguales a cero. Revisando los datos, el valor más pequeno medido corresponde a 4,7, por lo que la interpretación de la ordenada al origen en este caso no es válida. Sólo consideraremos esta estimación como un valor de ajuste de modelo.
La interpretación de la pendientes es: por cada unidad cci[2] que aumenta la variable independiente (CCI) el contenido de clorofila aumenta en promedio 10,08 mg/cm^2. Esta estimación es válida para valores entre el mínimo de la variable independiente (4,7) y el máximo (45,1).

8. Escriba la ecuación de la recta ajustada. Realice una estimación del contenido de clorofila total en laboratorio a una cantidad de Indice CCI de 90. ¿Es válida esta estimación? Si su respuesta es afirmativa o negativa justifique claramente por qué.

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$
$$\hat{y}_i = 118{,}2 + 10{,}08 x_i$$
$$\hat{y} = 118{,}2 + 10{,}08 \times 90$$
$$\hat{y} = 1025{,}4 \text{ mg/cm}^2$$

La estimación anterior no es valida, ya que la estimación de un predicho sólo es valida en la medida que se utilicen valores entre el mínimo de la variable independiente (4,7) y el máximo (45,1). 90 está fuera de ese rango y no tengo la capacidad de saber el comportamento del modelo a valores mayores que 45 o menores a 4,7..

9. Interprete agronómicamente el valor estimado de la pendiente.

[2] Es importante destacar la unidad de medición que utiliza la variable independiente (kg, cm, etc). En este caso la unidad de medición del equipo CCI es "unidades cci".

PRACTICO REGRESIÓN LINEAL SIMPLE

Por cada unidad cci que aumenta la variable independiente (CCI) el contenido de clorofila aumenta en promedio 10,08 mg/cm^2. Esta estimación es válida para valores entre el mínimo de la variable independiente (4,7) y el máximo (45,1).

10. Interpretación agronómicamente el valor estimado de la ordenada al origen. ¿Es válida esta interpretación en el contexto en que se montó el experimento? Justifique.

La interpretación de la ordenada al origen es: si los valores de la variable independiente son iguales a 0 (lo que mido con el equipo me entrega valores iguales a 0) el contenido de clorofila en promedio será 118,2 mg/cm^2. Este resultado sólo es valido si en la muestra he medido valores iguales a cero. Revisando los datos, el valor más pequeno corresponde a 4,7, por lo que la interpretación de la ordenada al origen en este caso no es válida. Sólo consideraremos esta estimación como un valor de ajuste de modelo.

11. ¿Es significativo el modelo ajustado para predecir la variable dependiente en función de la variable independiente? (ns 5 %?). Platee las hipótesis correspondientes, pruébelas y concluya.

La primera pregunta que surge al analizar los datos y obtener las predicciones de los parámetros es si la relación lineal que estamos observando y que podemos estimar a través de un modelo lineal es producto del "azar" o si realmente la utilización de la variable "x" (CCI) nos ayuda a predecir a la variable "y" (clorofila) mediante la construcción de un modelo. Si la relación fuese por azar, sería equivalente a que en un modelo teórico, independiente aumenten o disminuyan los valores de la variable "x", la variable "y" se mantiene constante. La opción más conservadora, conocida como hipótesis nula (H_0) nos hace suponer que la variable CCI no nos ayuda a predecir la clorofila, lo cual se puede expresar de la siguiente manera[3]:

$$H_0: \beta_1 = 0$$

Si la pendiente teórica es igual a cero, significa construir una curva paralela al eje x, en cuyo caso la predicción de la variable "y", dado un valor de "x" siempre sería en promedio la misma. Visualmente, la tendencia que observariamos en nuestro gráfico de dispersión sólo sería por azar.

La alternativa a la opción conservadora, es que la variable x si ayuda a predecir la variable y. Eso significa que gráficamente la pendiente estimada puede ser positiva o negativa, lo cual se puede expresar como:

$$H_a: \beta_1 \neq 0$$

Para poder proponer esta prueba estadística debemos conocer la distribución de un estadístico determinado bajo hipótesis nula. El estadístico:

[3] En las pruebas de hipótesis se usan las letras griegas "sin sombrero" ya que en las pruebas de hipótesis nos referimos a los parámetros del modelo y no a sus estimadores.

PRACTICO REGRESIÓN LINEAL SIMPLE

$$t = \frac{\hat{\beta}_1}{\sqrt{\widehat{var}(\hat{\beta}_1)}} = \frac{\hat{\beta}_1}{EE_{\hat{\beta}_1}}$$

Tiene distribución t-student con $n - p$ grados de libertad, donde n es el número de observaciones (15) y p es el número de parámetros estimados (en este caso 2, la ordenada al origen y la pendiente).

Coeficientes de regresión y estadísticos asociados

Coef	Est.	E.E.	LI(95%)	LS(95%)	T	p-valor	CpMallows	VIF
const	118,20	14,61	86,63	149,76	8,09	<0,0001		
Índice CCI (Equipo manual)..	10,08	0,52	8,95	11,21	19,26	<0,0001	371,01	1,00

Los errores estándares para para $\hat{\beta}_1$ es 0,52. Luego el estadístico t, es $10{,}08/0{,}52 = 19{,}26$. El valor obtenido (19,26), positivo y negativo (dada la prueba de hipótesis bilateral), deberá ser contrastado con los obtenidos de una t tabulada $t_{n-p,1-\alpha/2}$. Se rechazará la hipótesis nula en la medida que el t calculado (19,26) sea mayor al percentil de una $t_{n-p,1-\alpha/2}$, que en nuestro caso corresponde a 2,16, o menor (-19,26) que $t_{n-p,\alpha/2}$ que en nuestro caso corresponden a -2,16 Lo cual se cumple en ambos casos, por lo tanto rechazamos H0[4]

Los softwares estadísticos entregan, junto a la estimación de los parámetros, errores estándares y valores t, un valor conocido como $p - valor$ el cual se utiliza como herramienta al momento de aceptar o rechazar la hipótesis nula. Este valor se interpreta de la siguiente manera:

"Si la hipótesis nula fuese verdadera (H_0: $\beta_1 = 0$), o sea, la pendiente fuese 0, la probabilidad de obtener una estimación de la pendiente, positiva o negativa igual a 10,08, es menor que 1 en 10000 (p valor < 0,0001)". Esto bajo los criterios de estadística inferencial dice que la pendiente estimada es estadísticamente distinta de cero. Luego el modelo propuesto es estadísticamente significativo. El criterio utilizado habitualmente dice aceptar H0 a valores de p > 0,05 y rechazar H0 a valores de p < 0,05. Si conocemos este valor, no es necesario decidir aceptar/rechazar H0 en función del t calculado y el t de tabla.

12. ¿El ajuste realizado permite darle una interpretación agronómica a la pendiente?

Si, dada la significancia del modelo. Ver punto 9.

13. ¿El ajuste realizado permite darle una interpretación agronómica al intercepto?

A pesar de la significancia del modelo, el intercepto no tiene interpretación agronómica. Ver punto 10.

[4] Rechazar la hipótesis nula no significa que ésta sea falsa. Significa que con los antecedentes con los que cuento, la evidencia nos sugiere decidir rechazar la hipótesis nula, teniendo una probabilidad al tomar esta decisión de equivocarnos (α=0,05).