Actividades y Lecturas complementarias Capítulo 10

Finalidad

Se plantean una serie de actividades para que los estudiantes generen datos. Se propone resolver un rompecabezas del mapa de la República Mexicana, se le puede dar el enfoque administrativo económico. Estas ideas se pueden extender a otros países o América Latina. Varias actividades se plantean con temas relacionados con la economía y habilidades de percepción. Con la información generada se puede construir un curso de estadística.
Se proponen una serie de lecturas que, principalmente, se pueden obtener por internet. Ellas se irán actualizando.

Actividades y Lecturas complementarias Capítulo 10

La relación entre dos variables es como sigue:

1. Cuantitativa contra cuantitativa:

El diagrama de dispersión es la herramienta clave para comparar dos variables cuantitativas.

2 Cualitativa contra cuantitativa:

Describir los valores de la variable cualitativa y graficar la variable cuantitativa para cada grupo. Éste se puede hacer usando puntos o diagramas de caja si el tamaño de la muestra es mayor o igual a 5. También pueden utilizarse diagramas de barra.

3 Cualitativa contra cualitativa

Usar una tabla de contingencia o una tabla cruzada de conteos, donde se escribe en cada celda el número de veces que las variables coinciden. La relación puede explorarse mediante proporciones.

Procedimiento para obtener el modelo de regresión lineal.

1. Obtener los datos

Variable independiente	X	$X_1, X_2, ..., X_n$
Variable dependiente	Y	$Y_1, Y_2, ..., Y_n$

2. Trazar el diagrama de dispersión

3. Calcular las dos sumas $\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i$. Realizar los cálculos para encontrar las siguientes cantidades

$$\begin{aligned}
\bar{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i,\ \bar{Y}=\frac{1}{n}\sum_{i=1}^{n} Y_i, \\
S_{xy} &= \sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) = \sum_{i=1}^{n} X_iY_i - n\overline{XY} \\
S_{xx} &= \sum_{i=1}^{n}(X_i-\bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2, \\
S_{yy} &= \sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2
\end{aligned}$$

4 Obtener los valores estimados de mínimos cuadrados para el intercepto $\widehat{\beta}_0$ y la pendiente $\widehat{\beta}_1$:

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\overline{X}, \quad \widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$$

5 Escribir la línea de regresión de mínimos cuadrados:

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X.$$

6 Encontrar el valor predicho Y en X_i, usando la línea de regresión:

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_i.$$

7 Determinar el coeficiente de correlación: $r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$.

Inferencia estadística sobre el parámetro ρ : $H_0 : \rho = 0$ **vs** $H_1 : \rho \neq 0$.

La prueba para verificar la hipótesis y construir el intervalo de confianza es:

Prueba $t - Student$ para el coeficiente de correlación ρ	$t = \frac{r-\rho}{S(r)}$

Donde t tiene una distribución t de $Student$con $n-2$ grados de libertad.

El error estándar del coeficiente de correlación r	$S(r) = \sqrt{\frac{1-r^2}{n-2}}$

Los puntos críticos para verificar la hipótesis $H_0 : \rho = 0$, $H_1 : \rho \neq 0$, son: $r_{c_i} = \rho + t(n-2, \alpha/2)S(r)$ y $r_{cd} = \rho + t(n-2, (1-\alpha/2))S(r)$. Los puntos críticos en forma estándar son: $t_{ci} = t(n-2, \alpha/2)$, $t_{cd} = t(n-2, (1-\alpha/2))$.

Así cómo para el intervalo del (1-alfa) % de confianza para ρ. $L_i \leq \rho \leq L_f$ Donde $L_i = r + t(gl, \alpha/2)S(r)$, y $L_d = r + t(gl, 1-\alpha/2)S(r)$, $S(r)$ es el error estándar.

Metodología para probar la hipótesis: $H_0 : \beta_1 = 0$ **vs** $H_1 : \beta_1 \neq 0$

1. Dado que la hipótesis alternativa tiene el signo $\neq$, indica que tienen dos valores uno para izquierda generado para $\alpha/2$, el valor crítico es $t_{ci} = t(gl, \alpha/2)$. El otro, a la derecha con $1-\alpha/2$ y valor crítico es $t_{cd} = t(gl, (1-\alpha/2))$. Se usa la distribución $t - Student$ ya que, el tamaño de muestras es pequeño.

2. Los estadísticos de prueba derivados para esta situación son:

$$\begin{array}{ll} \text{Izquierda} & \widehat{\beta}_{1ci} = \beta_1 + t(gl, \alpha/2)ES(\widehat{\beta}_1) \\ \text{Derecha} & \widehat{\beta}_{1cd} = \beta_1 + t(gl, (1-\alpha/2))ES(\widehat{\beta}_1) \end{array}$$

   No se rechaza la hipótesis nula, si el valor $\widehat{\beta}_{1m}$ estimado de la muestra está entre estos valores, es decir $\widehat{\beta}_{1ci} \leq \widehat{\beta}_{1m}\widehat{\beta}_{1cd}$, en caso contrario se rechaza.

$$\begin{array}{ll} \textbf{Error estándar para el estadístico } \widehat{\beta}_1 & ES(\widehat{\beta}_1) = \widehat{\sigma}\sqrt{\frac{1}{S_{xx}}} \\ \textbf{Error estándar para el estadístico } \widehat{\beta}_0 & ES(\widehat{\beta}_0) = \widehat{\sigma}\sqrt{\frac{1}{n} + \frac{X^2}{S_{xx}}} \end{array}$$

Procedimiento alternativo usando el estadadístico estandarizado

La prueba anterior es similar con el estadístico $t - Student$, y en atención a la información que proporciona la muestra, éste tiene la forma:

$$t_m = \frac{(\widehat{\beta}_{1m} - \beta_1)}{\widehat{\sigma}(\widehat{\beta}_1)}$$

Se distribuye como una $t - Student$ con $n-2$ gl. Se realiza la prueba de hipótesis planteada comparando éste valor los puntos críticos t_{ci} y t_{cd}. Entonces, no se rechaza la hipótesis nula $H_0 : \beta_1 = 0$, si t_m está entre estos puntos y de nuevo, se rechaza en caso contrario.

Procedimiento alternativo con el $valor - p$

Finalmente se calcula la probabilidad para obtener el $valor - p$ y comparar éste con el nivel de significancia $\alpha/2$, para la prueba bilateral, $H_1 : \beta_1 \neq 0$, o con α para la prueba mayor $H_1 : \beta_1 > 0$ o menor $H_1 : \beta_1 < 0$. Estas probabilidades, cuando se aplica la distribución $t - Student$, se obtienen mediante la expresión:

$$\begin{array}{l} valor - p = P(t \geq t_m), \text{ o} \\ valor - p = P(t \leq t_m) \end{array}$$

Intervalos de confianza para los parámetros del modelo de regresión

1. Intervalo de confianza para β_1. Se puede encontrar un intervalo de confianza para el parámetro β_1 usando la distribución t. Un intervalo de confianza para β_1 con un nivel de

$100(1-\alpha)\,\%$ se obtiene mediante la expresión:

$$L_{1i} = \widehat{\beta}_1 + t(gl, \alpha/2)\widehat{\sigma}(\widehat{\beta}_1) = \widehat{\beta}_1 + t(gl, \alpha/2)\widehat{\sigma}\sqrt{\tfrac{1}{Sxx}}$$
$$L_{1d} = \widehat{\beta}_1 + t(gl, (1-\alpha/2))\widehat{\sigma}(\widehat{\beta}_1) = \widehat{\beta}_1 + t(gl, (1-\alpha/2))\widehat{\sigma}\sqrt{\tfrac{1}{Sxx}}$$

donde $t(gl, \alpha/2)$ es el punto correspondiente a la distribución t para $gl = n-2$ y $\alpha/2$.

2. Intervalo de confianza para β_0. Con un nivel de significancia del $100(1-\alpha)\,\%$:

$$L_{0i} = \widehat{\beta}_0 + t(gl, \alpha/2)\widehat{\sigma}(\widehat{\beta}_0) = \widehat{\beta}_0 + t(gl, \alpha/2)\widehat{\sigma}\sqrt{\tfrac{1}{n} + \tfrac{\overline{X}^2}{Sxx}}$$
$$L_{0d} = \widehat{\beta}_0 + t(gl, (1-\alpha/2))\widehat{\sigma}(\widehat{\beta}_0) = \widehat{\beta}_0 + t(gl, (1-\alpha/2))\widehat{\sigma}\sqrt{\tfrac{1}{n} + \tfrac{\overline{X}^2}{Sxx}}$$

Coeficiente de determinación

R^2 es un índice para evaluar el porcentaje de los datos que son explicados por el modelo y se expresa por:

$$R^2 = \frac{SC_{regresión}}{SC_{total}} = \frac{SC_{total} - SC_{error}}{SC_{total}} = 1 - \frac{SC_{error}}{SC_{total}}$$

Una expresión operativa para calcular el coeficiente se plantea en la siguiente ecuación,

$$\frac{SC_{regresión}}{SC_{total}} = \frac{S_{xy}^2/S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}}$$

Puesto que $0 \leq SC_{regresión} \leq SC_{total}$, entonces

$$0 \leq R^2 \leq 1.$$

Análisis de la varianza y la prueba de hipótesis sobre el modelo de regresión:

Fuentes de Variación	Suma de Cuadrados	GL	Cuadrado Medio	Razón de varianza
Modelo (R)	$SC_{regresión}$	1	$\frac{SC_{regresión}}{1}$	$RV = \frac{SCR/1}{SCE/(n-2)}$
Residual (E)	SC_{error}	$n-2$	$\frac{SC_{error}}{n-2}$	
Total (T)	SC_{total}	$n-1$		

H_0 : No existe relación lineal entre X y Y
H_1 : Sí existe relación lineal entre X y Y :

RV es una variable aleatoria que tiene una distribución F con 1 y $n-2$ grados de libertad respectivamente, para decidir si los datos confirman la hipótesis nula se compara RV con el valor $F_c = F(1, n-2, \alpha)$.

$$Si\ RV > F_c = F(1, n-2, \alpha),\ \text{se rechaza la hipótesis } H_0$$

Parte 2. Temas selectos de regresión

Regresión lineal simple: notación matricial

Una aplicación para el curso de álgebra lineal. La finalidad de este apartado es mostrar mediante un ejemplo el procedimiento de mínimos cuadrados empleando la notación matricial , esta es importante por la facilidad de generalizar la estimación de los parámetros en un modelo de regresión con más de una variable independiente X.

Ejemplo 10.1E

La administración quiere establecer tiempos de garantía para ello realiza una prueba para las denominadas baterías. Se ha probado que el tiempo de vida de un acumulador se puede predecir (por ejemplo) midiendo la carga que se le proporciona a la batería (en voltios). Se realiza una prueba de vida acelerada, en este caso los acumuladores se sometieron a ciertas condiciones ambientales de calor (entre otras posibles pruebas). Los datos para seis acumuladores son:

corriente X	corriente Y
17.9	245
23.6	220
30.9	215
56.1	211
61	161
77	135

El objetivo es encontrar la mejor relación lineal entre las variables X y Y. El estimador de mínimos cuadrados es:

$$\widehat{\beta}' = (X'X)^{-1}X'Y,$$

donde:

$$X' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 17.9 & 23.6 & 30.9 & 56.1 & 61 & 77 \end{pmatrix}$$

Aplicando el procedimiento de mínimos cuadrados se obtienen los siguientes resultados:

$$X'X = \begin{pmatrix} 6 & 266.5 \\ 266.5 & 14629.39 \end{pmatrix} \qquad X'y = \begin{pmatrix} 1187 \\ 48274.1 \end{pmatrix}$$

la inversa:

$$(X'X)^{-1} = \begin{pmatrix} 0.873 & -0.016 \\ -0.016 & 0.00036 \end{pmatrix}$$

finalmente se tiene:

$$\widehat{\beta} = (268.593, -1.593)$$

el modelo estimado es:

$$\widehat{Y} = 268.593 - 1.593x$$

Análisis estadístico:

Hipótesis

$$H_o : \beta_1 = 0$$
$$H_1 : \beta_1 \neq 0$$

El estadístico, usando la distribución t

$$t_c = \frac{\widehat{\beta}_1 - \beta_1}{ES(\widehat{\beta}_1)} \sim t(n-1, \alpha)$$

Cálculos

$$ES(\widehat{\beta}_1) = (S^2(X'X)^{-1})^{1/2} = 0.3543$$

donde S^2 es un estimador de la varianza σ^2, y se expresa por:

$$S^2 = \frac{(Y - \widehat{Y})'(Y - \widehat{Y})}{n - p} = 350.4384$$

$$t_c = \frac{-1.593 - 0}{0.3543} = -4.5$$

el valor de $t(n - p, \alpha) = t(4, 0.025) = -2.78$.

El coeficiente de determinación:

$$R^2 = 1 - \frac{(Y - \widehat{Y})'(Y - \widehat{Y})}{(Y - \overline{Y})'(Y - \overline{Y})} = 1 - \frac{1401.9653}{8488.83} = 83.5$$

el coeficiente de correlación:

$$r = \left(\frac{(X - \overline{X})'(Y - \overline{Y})}{(X - \overline{X})'(X - \overline{X})(Y - \overline{Y})'(Y - \overline{Y})} \right) = -0.913$$

Las siguientes tres tablas muestran el resumen estadístico del análisis de este ejemplo.

Tabla 10.1E. Resumen estadístico.

parámetro	estimación	error estd.	t_c	p
β_o	268.593	17.494	15.353	0.0001
β_1	-1.593	0.354	-4.496	0.0108

Tabla del análisis de la varianza: ANDEVA

Tabla 10.2E. Resumen del análisis de varianza.

fuente de variación	suma de cuadrados	gl	cuadrado medio	razón	p
modelo	7086.868	1	7086.868	20.220	0.0108
residual	1401.965	4	350.497		0.0108
total	8488.833	5			0.0108

.

Tabla 10.3E. Análisis de la varianza en general.

fuente de variación	suma de cuadrados	gl	cuadrado medio	razón	p
modelo	$(\widehat{Y}-\overline{Y})'(\widehat{Y}-\overline{Y})$	$p-1$	$\frac{(\widehat{Y}-\overline{Y})'(\widehat{Y}-\overline{Y})}{p-1}$	$\frac{CM_{regresin}}{CM_{residual}}$	
residual	$(Y-\widehat{Y})'(Y-\widehat{Y})$	$N-p$	$\frac{(\widehat{Y}-\overline{Y})'(\widehat{Y}-\overline{Y})}{p-1}$		
total	$(Y-\overline{Y})'(Y-\overline{Y})$	$N-1$			

Inferencia por intervalo de confianza:

Intervalo de confianza para la pendiente β_1 :

$$\widehat{\beta}_1 = \pm\, t(N-p,\alpha)S\sqrt{\frac{1}{(X-\overline{X})'(X-\overline{X})}}$$

Intervalo de confianza para una media de la variable de respuesta μ^* :

$$\widehat{y}^* \pm t(N-p,\alpha)S\sqrt{\frac{1}{N}+\frac{(X^*-\overline{X})^2}{(X-\overline{X})'(X-\overline{X})}}$$

Donde:

$$\widehat{Y}^* = \widehat{\beta}_o + \widehat{\beta}_1 X^*$$

Intervalo de confianza para predecir un valor de la variable de respuesta:

$$\widehat{Y}^* \pm t(N-p,\alpha)S\sqrt{1+\frac{1}{N}+\frac{(X^*-\overline{X})^2}{(X-\overline{X})'(X-\overline{X})}}$$

Modelo de regresión múltiple

Modelo. Si con la finalidad de explicar un fenómeno o proceso se incorporan nuevas variables al modelo lineal simple, entonces se tiene el modelo de regresión múltiple , el cual se representa mediante la siguiente expresión :

$$y_i = \beta_0 + \beta_1 X_{1_i} + \beta_2 X_{2_i} + ... + \beta_k X_{k_i} + \varepsilon \tag{1}$$

donde β_0 es una contante, los, β_i, $i = 1, 2, ..., k$, son los parámetros del modelo, las X_{ji} $j = 1, 2, ...n$, $i = 1, ..., k$ son las variables independientes que describen las k características de los n individuos u objetos, Y_i es la variable de respuesta, se tienen n observaciones sobre los individuos, ε es una variable aleatoria.

La ecuación 1 es un modelo hipotético con el cual se tratará de explicar los resultados de una situación real, en general la idea es bosquejar o construir un modelo que nos va a describir e interpretar un fenómeno, para ello planteamos el siguiente procedimiento :

1. Proponer un modelo, esto incluye la selección de las variables que aparecen en el modelo.
2. Estimar los parámetros $\beta_0, \beta_1, \beta_2$,...,$\beta_k$ del modelo.
3. Especificar la distribución de probabilidad de la variable aleatoria ε, estimar la varianza σ^2 de la distribución.
4. Verificar la utilidad del modelo.
5. Usar el modelo para predecir valores particular de Y, dentro del rango de valores de la variable independiente.

Nota: El modelo 1 es lineal en los parámetros y en las variables independientes, en este caso el modelo recibe el nombre de modelo lineal de primer orden. Más adelante se estudiarán otros modelos.

Ajuste del modelo

Supuestos estadísticos sobre el modelo. Con el objeto de estimar los parámetros del modelo es necesario hacer algunas suposiciones sobre la variable aleatoria ε:

1. Para cualquier conjunto de valores de $X_1, X_2, X_3, ..., X_k$, la variable ε se distribuye como una distribución de probabilidad normal con media cero y varianza σ^2. En símbolos $\varepsilon \sim N(0, \sigma^2)$, σ^2 es constante.
2. Los errores aleatorios son independientes en el sentido probabilístico.

Proceso de estimación. El objetivo es nuevamente estimar los parámetros del modelo con k variables independientes que influirán en la respuesta Y. Con la idea de fijar ideas sobre el procedimiento de estimación por mínimos cuadrados , se considerará la situación de que únicamente existen dos variables que explican la respuesta. Este modelo es de la forma siguiente:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i, \tag{2}$$

el procedimiento es análogo al que se presentó para el caso de una variable, puesto que la finalidad es optimizar la siguiente expresión:

$$\sum_{i=1}^{n} \varepsilon_i \varepsilon_i = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}))^2 \tag{3}$$

Así, los estimadores de mínimos cuadrados obtenidos por este planteamiento es la solución que se obtiene al resolver simultáneamente las ecuaciones que resultan de la optimización.

Estimación por mínimos cuadrados

Si el número de variables independientes aumentan, las ecuaciones simultáneas que se obtienen al optimizar la expresión 1 también aumentan, así la solución para estimar los parámetros se va complicando algebráicamente. Una alternativa matemática que facilita el proceso de estimación por mínimos cuadrados es la utilización de la representación matricial, esto da lugar a realizar operaciones con matrices como el producto, inversión y la transposición. A continuación se presenta este planteamiento.

Planteamiento empleando la notación matricial

Es de la mayor utilidad operativa y para la comprensión la representación del modelo de regresión en términos de matrices, así las respuestas se expresan mediante un vector Y de dimensión $n \times 1$, n es el número de observaciones que se realizan a un sujeto u objeto. Las variables independientes se representan por la matriz X de dimensión $n \times r$ donde el número de columnas r describe al término constante y las k variables (las k características de interés para explicar un fenómeno).

El vector de parámetros se indica por β de dimensión $r \times 1$. Por ε el vector aleatorio $n \times 1$.

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ . \\ . \\ . \\ Y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & X_{11} & X_{12} & . & . & . & X_{1k} \\ 1 & X_{21} & X_{22} & . & . & . & X_{2k} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & X_{n1} & X_{n2} & . & . & . & X_{nk} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ . \\ \beta_k \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_n \end{bmatrix}$$

de esta manera el modelo se puede escribir en forma matricial como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon = X\beta + \varepsilon \tag{4}$$

Estimación de los parámetros del modelo

Al inicio describimos que el método de mínimos cuadrados nos permite estimar los parámetros del modelo. Por medio del procedimiento de mínimos cuadrados estimamos los parámetros del modelo, el planteamiento estadístico es como sigue:

$$\text{minimizar } \varepsilon'\varepsilon = (y - X\beta)'(y - X\beta) \tag{5}$$

La solución de esta minimización queda representada por la siguiente expresión:

$$\widehat{\beta} = (X'X)^{-1} X'Y \tag{6}$$

Los elementos del vector $\widehat{\beta}$ contienen los estimadores de los parámetros del vector β, y con ello se puede construir el modelo de regresión estimado.

Ejemplo 10.2E

Un economista agrícola tiene interés en evaluar el rendimiento (Y) de un grano en unas parcelas, debido al promedio de precipitación anual de lluvia X_1 y a la temperatura promedio anual X_2, para ello considera ocho años. Estime el modelo de regresión, evalúe la importancia estadística de cada variable para explicar la respuesta. Se puede notar que el modelo que se propone es un plano, así se busca el plano de regresión que mejor se ajuste a los datos descritos por el estudio. Los datos observados se presentan en la siguiente Tabla

año	X_1 :nivel de lluvia	X_2 :temperatura promedio	y prod.
1981	39	20	55
1982	37	26	65
1983	47	19	80
1984	37	27	75
1985	39	24	70
1986	38	21	50
1987	40	23	60
1988	41	22	65

Solución mediante el uso de CalEst

La finalidad es presentar la solución usando el módulo de regresión múltiple en el *CalEst*. La descripción del análisis estadístico y la justificación del procedimiento de estimación se plantea en el siguiente apartado.
En la figura 10.1E se muestra la *estimación de los parámetros* y la *inferencia estadística* de éstos, para realizar esta última se estima el error estándar de los estimadores y se calcula el estadístico de prueba t de Student y finalmente se indica el valor del nivel de significancia descriptivo p. El valor de p es la probabilidad de la distribución t con los grados de libertad correspondientes a la izquierda si el estadístico es negativo o a la derecha si el estadístico es positivo. Si $p < \alpha$ se rechaza, la hipótesis correspondiente al parámetro β_i, $i = 1, 2$, en caso contrario no se rechaza la hipótesis.

Estimación de los parámetros del modelo

Los valores de los parámetros estimados son:

$$\widehat{\beta} = (\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2) = (-177.439, 3.863, 3.908)$$

Por lo tanto el modelo es:

$$y = -177.439 + 3.863X_1 + 3.908X_2$$

Prueba de hipótesis: parámetros del modelo

Con esta información se puede hacer *inferencia sobre los parámetros*, y de esa manera averiguar si tanto el nivel de lluvia como la temperatura de manera individual tienen efecto estadísticamente significativo sobre el rendimiento por hectárea. Para ello se plantean las siguientes *hipótesis* :

$$H_0 : \beta_1 = 0 \quad \text{contra} \quad H_1 : \beta_1 \neq 0 \qquad \text{y}$$

$$H_0 : \beta_2 = 0 \quad \text{contra} \quad H_1 : \beta_2 \neq 0$$

Ver los resultados reportados en la figura 10.1E. Cada una de estas hipótesis se prueban con el estadístico t de Student, el planteamiento y el cálculo de éstas es como sigue:

$$t_1 = \frac{\widehat{\beta}_1 \ - \beta_1}{\widehat{\sigma}(\widehat{\beta}_1)} = \frac{3.863 - 0}{0.725} = 5.331$$

y

$$t_2 = \frac{\widehat{\beta}_2 \ - \beta_2}{\widehat{\sigma}(\widehat{\beta}_2)} = \frac{3.908 - 0}{0.834} = 4.687$$

Como ambos estadísticos son mayores que el valor $t(\alpha/2, gl = 5) = 2.571$ y con un $\alpha = 0.05$, se tiene que ambas variables contribuyen de manera significativa en el efecto de rendimiento de producción del grano. La figura 10.2E indica el valor del estimador, su error estándar y el intervalo de confianza del 95 %, en este caso para los parámetros. Finalmente, se indica el coeficiente de correlación entre la respuesta y cada una de las variables de entrada. Asimismo se indican los coeficientes de determinación y el error estándar.

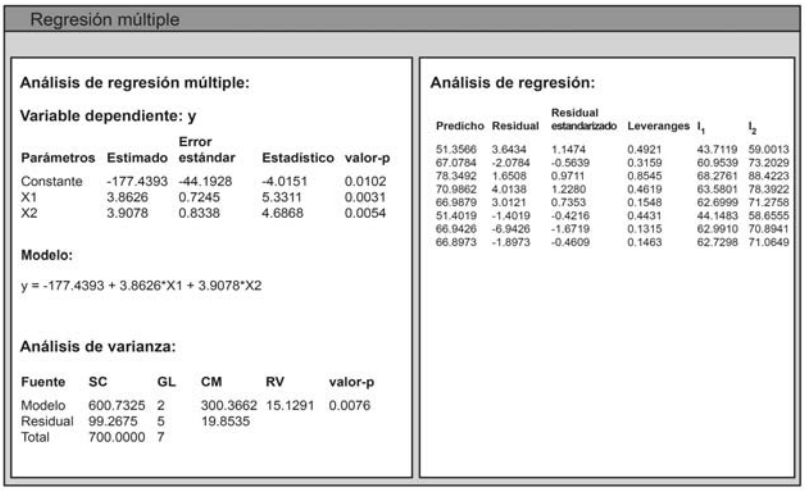

Figura 10.1E Cuadro de la izquierda reporta la estimación de los parámetros y el análisis de la varianza. A la derecha se presenta información descriptiva e intervalos de confianza.

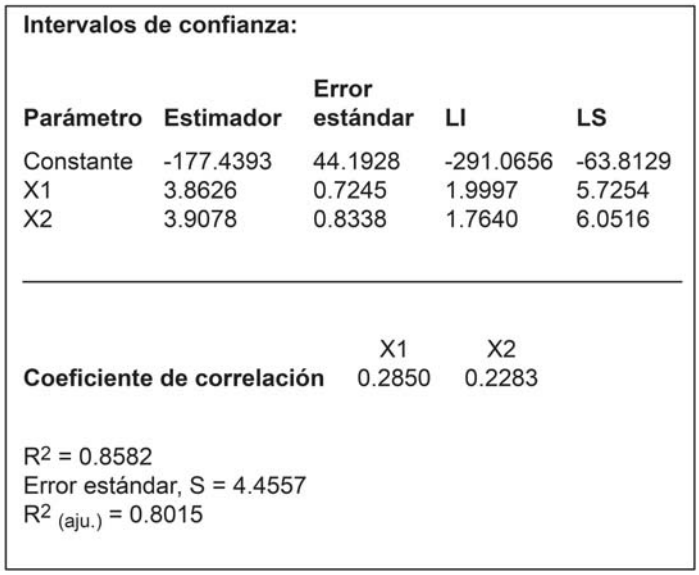

Figura 10.2E Completa el análisis de regresión múltiple, las correlaciones son entre (y y X_1) y (y y X_2).

Análisis de residuales

Las gráficas de la figura 10.3E describen el análisis de residuales.

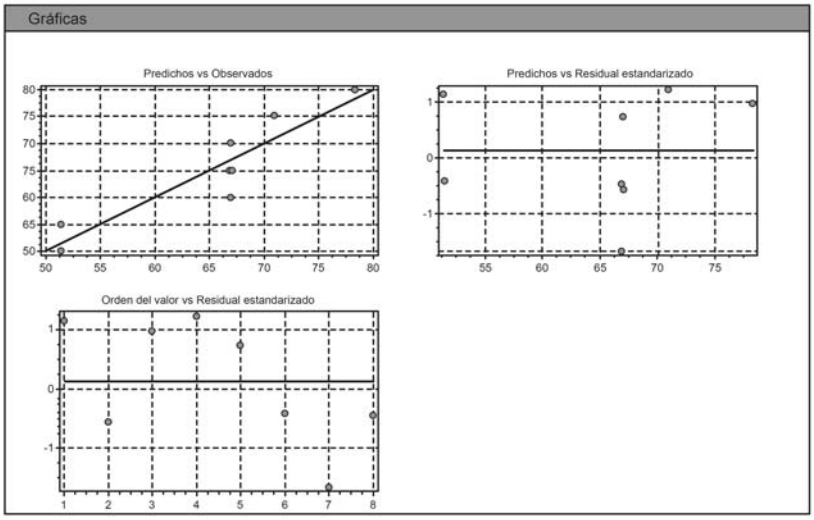

Figura 10.3E Análisis gráfico de los residuales.

Análisis e interpretación del modelo

Se ha construido el modelo de regresión del ejemplo. Una inferencia importante en el análisis de regresión es evaluar la significancia del modelo, tal situación se plantea mediante la hipótesis :

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$

$$H_1 : \beta_i \neq 0 \text{para alguna } i, i = 1, ..., k$$

Si se rechaza la hipótesis H_0 se concluye que el modelo es estadísticamente significativo, así el modelo es de utilidad para predecir valores de Y dentro del rango de las variables independientes. El procedimiento para probar esta hipótesis se resume en una tabla conocida como análisis de la varianza y se abrevia ANDEVA, la cual consiste en dividir la suma de cuadrados total en dos componentes, una que concierne al modelo y la otra, la que no explica el modelo; esta última corresponde a la suma de cuadrados de la discrepancia que existe en el valor observado y el ajustado por el modelo. En general a la discrepancia se le reconoce como *residual*. Finalmente se tiene que: suma de cuadrados total = suma de cuadrados del modelo + suma de cuadrados de residuales; en símbolos:

$$\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{n} \left(Y_i - \widehat{Y}\right)^2 \tag{7}$$

Notación: los términos de la ecuación 7, se conocen como la suma de cuadrados, el término del primer miembro se denomina el *total* y se denota por SCT, éste es: $\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$, el primer término del segundo miembro se conoce como *la suma de cuadrados que se refiere al modelo* y se denota por SCM y es: $\sum_{i=1}^{n} \left(\widehat{Y}_i - \overline{Y}\right)^2$, finalmente el otro término del segundo miembro se

denomina como *la suma de cuadrados no explicada por el modelo* o *suma de cuadrados del error*, ésta se denota por SCE, es decir: $\sum_{i=1}^{n}\left(Y_i-\widehat{Y}\right)^2$. Con esta información y los grados de libertad se construye la tabla 10.4E del *análisis de la varianza*:

Tabla 10.4E Descripción de la tabla de ANDEVA

Fuente de variación	Suma de cuadrados	gl	Cuadrados medios	Razón	P
Modelo	$\sum_{i=1}^{n}\left(\widehat{Y}_i-\overline{Y}\right)^2$	$k-1$	$\sum_{i=1}^{n}\left(\widehat{Y}_i-\overline{Y}\right)^2/(n-k)$	$F^*=\frac{CMmodelo}{CMerror}$	p
Error	$\sum_{i=1}^{n}\left(Y_i-\widehat{Y}\right)^2$	$n-k$	$\sum_{i=1}^{n}\left(Y_i-\widehat{Y}\right)^2/(n-k)$		
Total	$\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2$	$n-1$			

Los cuadrados medios que aparecen en la cuarta columna de la tabla 10.4E se obtienen dividiendo la suma de cuadrados por los grados de libertad. Luego se divide el cuadrado medio del modelo entre el cuadrado medio del error, el valor resultante viene de una distribución de probabilidad F con ν_1 y ν_2 grados de libertad. Éste permite probar la hipótesis nula sobre igualdad de parámetros, es decir:

$$F_c=\frac{CM_{modelo}}{CM_{error}}$$

Si F_c es mayor que un valor de la distribución F para un $\alpha=0.05$ con los grados de libertad apropiados, entonces se dice que los datos no apoyan la hipótesis nula H_0 y se rechaza. También se compara el valor de α con el valor de la probabilidad p, es decir si $\alpha>p$ se rechaza H_0.

Análisis de la varianza con respecto a la representaci ón matricial

En forma esquemática, el procedimiento para contrastar la *hipótesis global* se muestra en la tabla 10.5E. En la columna 5 se presenta el *valor calculado del estadístico*.

Tabla 10.5E Descripción del análisis de la varianza considerando la notación matricial.

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios	F_c	valor P
Regresión	$g-1$	$SC_{reg}=(\mathbf{y}'\mathbf{X}\widehat{\beta}-\overline{y}_\bullet^2/n)$	$\frac{SC_{reg}}{g-1}$	$\frac{CM_{reg}}{CM_{error}}$	
Error	$n-g$	$SC_{error}=\mathbf{y}'\mathbf{y}-\mathbf{y}'\mathbf{X}\widehat{\beta}$	$\frac{SC_{error}}{n-g}$		
Total	$n-1$	$SC_{total}=(\mathbf{y}'\mathbf{y}-\overline{y}_\bullet^2/n)$			

ANDEVA para el ejemplo 10.2E

La hipótesis que se probará para este ejemplo es:

$$H_0:\beta_1=\beta_2=0 \text{ contra la alternativa } H_1:\beta_i\neq 0, \text{para alguna } i=1,2.$$

En el cuadro Análisis de la varianza se muestra el análisis de la varianza, de ahí se puede concluir que se rechaza la hipótesis nula, ya que el valor de p es menor que $\alpha=0.05$.

Estimación de la varianza

La *varianza* σ^2 se estima por:

$$\widehat{\sigma}^2 = \frac{SCE}{n-k} = \frac{\sum \left(Y_i - \widehat{Y}\right)^2}{n-k}$$

la varianza σ^2 se estima así $\widehat{\sigma}^2$ es:

$$\widehat{\sigma}^2 = \frac{SCE}{n-k} = \frac{\sum (Y - \widehat{Y})^2}{n-k} = \frac{99.268}{7-2} = 19.854$$

y el error estándar es $\widehat{\sigma} = \sqrt{\widehat{\sigma}^2} = \sqrt{19.854} = 4.4578$.

Coeficiente de determinación

Otros resultados estadísticos para evaluar el modelo de regresión se siguen a partir del *coeficiente de determinación* , el cual se obtiene por:

$$R^2 = \frac{SCM}{SCT} = 1 - \frac{SCE}{SCT} = 1 - \frac{99.268}{700.0} = 0.858,$$

donde la $SCT = \sum \left(\widehat{Y}_i - \overline{Y}\right)^2 = 700.0$, entonces este coeficiente indica que aproximadamente el 86 % de los datos son explicados por el modelo. Por otro lado, el coeficiente de determinación ajustado por los grados de libertad es:

$$R^2_{ajs} = 1 - \frac{(n-1)SCE}{(n-k)SCT} = 1 - \frac{(7-1)99.268}{(7-2)700.0} = 0.802$$

Considerando esta información proporcionada por los datos se tiene que el modelo explica el 80 %. Se puede notar que estos coeficientes son aproximadamente iguales si el número de datos observados aumenta de manera considerable.

Coeficiente de determinación R^2: notación matricial

El reporte sobre la bondad del modelo se completa mediante el cálculo del *coeficiente de determinación*, el cual indica qué porcentaje de los datos son explicados por el modelo, éste se estima por la siguiente expresión:

$$R^2 = \frac{SC_{reg}}{SC_{total}} = 1 - \frac{SC_{error}}{SC_{total}} = 1 - \frac{\mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\widehat{\beta}}{\mathbf{y}'\mathbf{y} - \overline{y}^2_{\bullet}/n}$$

con la información presentada en la tabla del ANDEVA se obtiene el valor estimado de R^2.

Evaluación del modelo

La construcción del modelo y la estimación de los parámetros se basan en el supuesto de que la variable aleatoria ε tiene una distribución de probabilidad normal, entonces es importante verificar este supuesto, también se debe observar si se cumple la homogeneidad de la varianza. Estos aspectos relevantes se analizan descriptivamente mediante técnicas gráficas como se

muestran en la figura 10.4E.

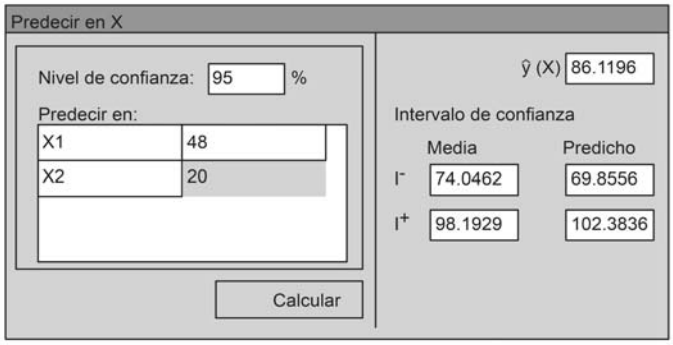

Figura 10.4E Intervalos de confianza para el modelo y la predicción.

Intervalos de confianza para el modelo y la predicci ón

En el análisis del modelo de regresión es importante estimar mediante *intervalos de confianza a los parámetros del modelo* y a la respuesta media en un valor específico de X, y mediante un *intervalo de predicción* para la respuesta individual para un valor determinado de la variable X.

Intervalo de confianza para β_1.

$$\widehat{\beta}_1 \pm \ t(n-g,\alpha/2)ES(\widehat{\beta}_1) \pm \ t(n-g,\alpha/2)S\sqrt{m_{11}}$$

donde m_{11} es el segundo elemento de la diagonal en la matriz $(\mathbf{X}'\mathbf{X})^{-1}$.

Intervalo de confianza para la media de la variable de respuesta μ^* en $\mathbf{x}_0$.

$$\widehat{Y}^* \pm t(n-g,\alpha/2)S\sqrt{(1,\mathbf{x}_0)\,(\mathbf{X}'\mathbf{X})^{-1}\,(1,\mathbf{x}_0)'}$$

donde $\widehat{Y}^* = \mathbf{x}_0\widehat{\beta}$, el vector de parámetros $\widehat{\beta}' = (\widehat{\beta}_0, \widehat{\beta}_1)$ y el vector $(1, \mathbf{x}_0)$, donde $\mathbf{x}_0$ representa a los niveles de los factores donde se requiere estimar la media de Y.

Intervalo de confianza para un valor de la variable de respuesta (predicho).

$$\widehat{Y}^* \pm t(n-g,\alpha/2)S\sqrt{1+(1,\mathbf{x}_0)(\mathbf{X}'\mathbf{X})^{-1}(1,\mathbf{x}_0)'}$$

Del ejemplo tratado se puede estimar el *intervalo de confianza* correspondiente al modelo y al de la predicción para un valor de las variables de entrada, considere que: $X_1 = 48$ y $X_2 = 20$, se tienen los resultados que muestra la figura 10.4E. Estos intervalos se obtienen usando el *CalEst*.

Parte 3. Estrategias de aprendizaje

1. Esta práctica está relacionada con el conocimiento de nuestra memoria a corto plazo.

Utilice la lista de palabras que aparece abajo, escríbalas en una tarjeta de manera clara. Preséntela a 10 amigos o conocidos de la siguiente manera:

1. *a*) Seleccione un amigo y muéstrele la tarjeta durante 15 segundos, deje pasar un minuto. Después, dígale que tiene 2 minutos para decirle las palabras que recuerde y anótelas.

   b) Repita la operación anterior con el siguiente amigo, pero ahora déjelo que la vea durante 20 segundos.

   c) Continúe de esa manera hasta completar los 10 amigos. Cada vez dele más tiempo (en cada caso 5 segundos más) para que lea la tarjeta.

Las variables registradas son "Tiempo de lectura de la tarjeta en segundos" y "Número de palabras recordadas". Utilice un cronómetro o un reloj para medir el tiempo. Complete la tabla:

Segundos	15	20	25	30	35	40	45	50	55	60
Número de palabras										

Lista de palabras:

Melón	Coche	Piedra	Cama
Llave	Silla	Local	Mesa
Joven	Lluvia	Casa	Cine
Abuelo	Pelota	Reloj	

1. *a*) 1) Dibuje el diagrama de dispersión. ¿Qué observa?

   2) Agrupe su información con la de 3 compañeros y elaboren otro diagrama de dispersión con la información de todos. ¿Qué observan ahora?

   3) Trace las medianas en cada eje para sus datos. ¿Qué observa?

2 Construya el artefacto, al que denominamos helicóptero, que aparece en la figura 10.5E. La finalidad es soltarlo desde cierta altura para que caiga girando. Para ello, tome el artefacto con las alas estiradas y déjelo caer desde una altura de 2 metros y registre el tiempo (ésta es una variable) con un reloj o un cronómetro. La otra variable consistirá en aumentar la longitud de la base del artefacto un centímetro (cm) en 10 ocasiones. Tome un centímetro como longitud inicial. Las dimensiones del artefacto son:

Características	Dimensiones
Longitud de ala	9 centímetros
Ancho de ala	3 centímetros
Longitud de cuerpo	1 centímetro
Longitud de la base	1 centímetro (inicial)
Ancho de cuerpo	2 centímetros

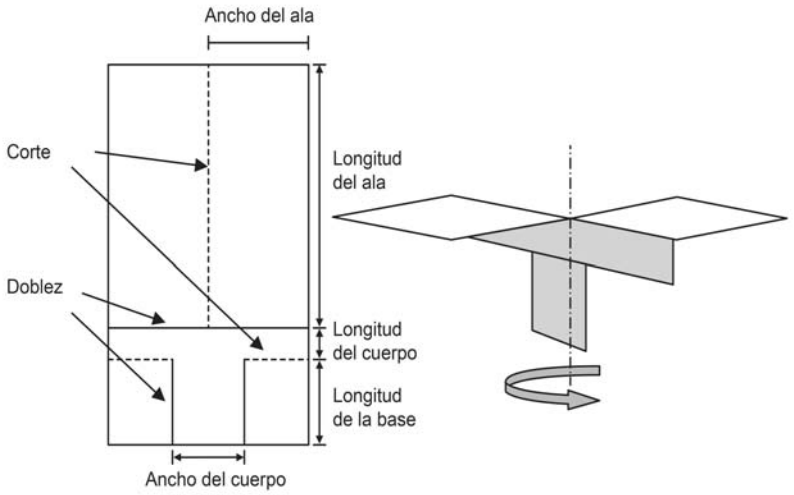

Figura 10.5E Descripción del helicóptero.

Longitud (cm)	1	2	3	4	5	6	7	8	9	10
Tiempo										

1. *a*) 1) Construya el diagrama de dispersión. ¿Qué puede decir de la relación entre estas variables?

   2) Calcule el coeficiente r.

   3) Repita el experimento dos veces más y haga un nuevo diagrama de dispersión. Trace las medianas para cada eje. ¿Qué observa?

Longitud (cm)		1	2	3	4	5	6	7	8	9	10
Tiempo	1										
	2										
	3										