Unidad 2: Ajuste de modelos.

2.0 Contenidos y objetivos.

Contenidos:

2.0	*Contenidos, objetivos y palabras clave*	*41*
2.1	*Introducción. Generalidades.*	*42*
2.2	*Ajuste de un modelo lineal simple por el método de mínimos cuadrados.*	*43*
2.2.1	*Ejercicios resueltos, paso a paso.*	*46*
2.2.2	*Ejercicios propuestos.*	*54*
2.3	*Modelos reducibles a lineales.*	*55*
2.3.1	*Ejercicios resueltos, paso a paso.*	*57*
2.3.2	*Ejercicios propuestos.*	*59*
2.4	*Comparación de modelos lineales y transformables a lineales.*	*60*
2.4.1	*Ejercicios resueltos, paso a paso.*	*61*
2.4.2	*Ejercicios propuestos.*	*70*
2.5	*Modelo parabólico.*	*71*
2.5.1	*Ejercicios resueltos, paso a paso.*	*72*
2.5.2	*Ejercicios propuestos.*	*75*
2.6	*Modelo de regresión múltiple.*	*75*
2.6.1	*Ejercicios resueltos, paso a paso.*	*76*
2.6.2	*Ejercicios propuestos.*	*78*

Objetivos:

- *Establecer el grado de correlación entre dos variables.*
- *Calcular los coeficientes de regresión y correlación.*
- *Determinar el mejor modelo que se ajusta a los datos e interpretar el error estándar de estimación*

Palabras Clave:

Diagrama de dispersión
Coeficiente de correlación
Variable independiente
Recta de regresión lineal
Error estándar de estimación.
Modelo lineal
Calidad del ajuste lineal
Covarianza
Coeficiente de regresión lineal
Variable dependiente
Estimación puntual.
Coeficiente de determinación
Modelo no lineal

2.1. Introducción. Generalidades.

- ✓ Si se miden **dos variables** sobre un **mismo sujeto** y se sospecha (después de observar la nube de puntos experimental en un **diagrama de dispersión**) que éstas están relacionadas mediante una relación de tipo lineal de la forma $y = a + bx$, se puede lograr una ecuación que permita relacionar los puntos experimentales con el modelo utilizado de manera de explicar, con algún grado de error, los valores de la **variable de respuesta** Y (de valores experimentales y), sobre la base de valores predeterminados de la **variable predictora** X (de valores experimentales x). Esto permite solucionar, al menos con un porcentaje manejado de error, muchos problemas de distintos ámbitos de aplicación.

- ✓ El análisis de regresión es útil para averiguar la forma probable de las relaciones entre las variables, y el objetivo final, cuando se emplea este método de análisis, es predecir o estimar el valor de una variable que corresponde a un valor dado de la predictora.

- ✓ Para el modelo de regresión simple son importantes dos variables numéricas X e Y. A la variable X se le conoce por lo general como **variable independiente**, ya que con frecuencia se encuentra bajo el control del investigador; es decir, los valores de x de la variable X pueden ser seleccionados por el investigador para obtener uno o más valores y de Y como respuesta a valores de x.

- ✓ El análisis de la **correlación** permite medir la intensidad de la relación entre las variables.

- ✓ Una medida estadística que puede indicar la posible relación (o asociación lineal) entre las variables es la **covarianza entre** X **e** Y**,** denotada por $Cov(X,Y)$ (en la muestra su valor será denotado por $Cov(x,y)$).

- ✓ La **ecuación de regresión de la población** describe la **relación real** entre la variable dependiente Y y la variable independiente X.

- ✓ Al método que se utiliza regularmente para obtener los estimadores de los valores a y b en la recta de regresión $y = a + bx + e$ se le conoce como **método de mínimos cuadrados**, y a la recta resultante de reemplazar a y b por sus estimadores $\hat{a}$ y $\hat{b}$ se le conoce como **recta de mínimo cuadrados** o **recta de regresión mínimo cuadrática**.

- ✓ En la gráfica cartesiana de la recta $y = a + bx$, y es un valor sobre el eje vertical Y, x un valor sobre el eje horizontal X, a es el punto de intersección de la recta con el eje vertical, y b es la **razón de cambio** que indica la cantidad en que cambia la variable Y por cada unidad de cambio en la variable X**.**

- ✓ Una vez logradas las **estimaciones** de a y de b, denotadas por $\hat{a}$ y $\hat{b}$, se construye la **recta de regresión de** y **sobre** x:

$$\hat{y} = \hat{a} + \hat{b}x$$ (**modelo de predicción** de valores de la variable Y).

- ✓ Un **valor estimado de** Y, para un valor determinado de X, por ejemplo, $X = x_0$, se obtiene al reemplazar en la recta de regresión el valor de x_0, es decir $\hat{y}$ (o bien y_{est}, en la notación de otros autores) es tal que $\hat{y} = \hat{a} + \hat{b}x_0$.

- ✓ Se llama **error de estimación puntual** (e_i) para un valor dado $x = x_0$, a la diferencia entre el valor observado (real) y (también anotado como y_{obs}), y el valor estimado $\hat{y}$. Es decir, $e_i = y_{obs} - \hat{y}$, o bien $e_i = y - \hat{y}$, para un valor dado $X = x_0$.

- ✓ Una vez **ajustada** la recta de regresión a la **nube de observaciones experimentales**, es importante disponer de una medida que mida la **bondad del ajuste** realizado y que permita decidir si el ajuste lineal es suficiente o se deben buscar modelos alternativos. Tal medida es llamada **Coeficiente de correlación lineal de Pearson** entre las variables X e Y.

- ✓ El **error estándar de estimación** (*e.e.e.*) permite también medir la calidad del ajuste, pero, términos del error cometido con las **predicciones** $\hat{y}$, cuando son comparadas con los valores y de la muestra.

- ✓ El error estándar de estimación se calcula con la finalidad de medir la **confiabilidad** de la ecuación de la estimación.

- ✓ Un modelo que no es lineal recibe el nombre de **modelo no lineal.**

- ✓ El modelo parabólico o cuadrático está dado por $y = a + bx + cx^2$ donde a, sigue siendo la intersección con el eje de ordenadas Y, y b y c están relacionados con la **pendiente** y la **tasa de variación** de la curva.

- ✓ Cuando se usa **más de una variable independiente** para predecir los valores de una variable dependiente, el proceso se llama **análisis de regresión múltiple**

$$\underbrace{y}_{\text{Varieble de respuesta}} = \underbrace{(b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k)}_{\text{Modelo Matemático}} + \underbrace{e}_{\text{Error}}$$

donde y es la variable que se quiere predecir; $x_1, x_2, \ldots y\ x_k$ son k variables con valores conocidos; $b_0, b_1, b_2 \ldots b_k$ son **coeficientes numéricos** que deben ser determinados a partir de la información conocida y

$$y_{\text{modelo matem.}} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

es el modelo matemático supuesto para explicar la relación.

2.2. Ajuste de un modelo lineal simple por el método de mínimos cuadrados.

Resumen de conceptos y propiedades:

- ✓ La covarianza es una medida estadística que puede indicar la relación o asociación lineal entre las variables X e Y. Está definida por

$$Cov(X,Y)=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)}{n}$$

donde el producto $\left(X_i-\bar{X}\right)\left(Y_i-\bar{Y}\right)$ permite auscultar la dispersión de los puntos bidimensionales $\left(X_i,Y_i\right)$, $i=1,2\ldots n$, que representan a las observaciones en el plano bidimensional, con respecto al centro de gravedad $\left(\bar{X},\bar{Y}\right)$.

Si $Cov(X,Y)>0$, existe una posible **asociación lineal directa** entre X e Y.

Si $Cov(X,Y)<0$, existe una posible **asociación lineal indirecta** entre X e Y.

Si $Cov(X,Y)=0$, lo único que se puede afirmar es que **la relación entre** X e Y **no es lineal**.

También se puede expresar como

$$\boxed{Cov(X,Y)=\overline{XY}-\bar{X}\cdot\bar{Y}}$$

Para datos numéricos presentados en forma de tabla se tiene:

$$Cov(X,Y)=\frac{\sum_{j=1}^{m}\sum_{i=1}^{k}n_{ij}\left(X_i-\bar{X}\right)\left(Y_j-\bar{Y}\right)}{n}=\overline{XY}-\bar{X}\cdot\bar{Y}$$

donde, para $i=1,2,...,k$ y $j=1,2,...,m$:

n_{ij} : Frecuencia conjunta de los pares observados (x_i,y_j).

$\bar{X}$: Media marginal de la variable X.

$\bar{Y}$: Media marginal de la variable Y.

$\overline{XY}$: Media de los productos de valores X_i y Y_j, para todo i y j.

- ✓ En el **modelo lineal simple** $y=a+bx+e$ son importantes los siguientes indicadores y ecuaciones:

a. $r_{X,Y} = \dfrac{Cov(X,Y)}{S_X S_Y}$ (**coeficiente de correlación lineal** de Pearson entre las variables X e Y).

b. $\hat{b} = \dfrac{Cov(X,Y)}{S_X^2}$ y $\hat{a} = \overline{Y} - \hat{b} \cdot \overline{X}$ (estimadores mínimo cuadráticos de a y b).

c. $\hat{y} = \hat{a} + \hat{b} \cdot x$ (**recta de predicción de Y dado $X = x$**).

d. $\hat{y}_{x=x_0} = \hat{y}_0 = \hat{a} + \hat{b} \cdot x_0$ (**Ecuación de estimación** de Y para un valor dado $X = x_0$).

e. En particular, $\hat{y}_i = \hat{a} + \hat{b} \cdot x_i$, $i = 1,2,\ldots,n$ para los datos de la muestra de tamaño n, y para un valor fijo $X = x_i$.

✓ El **error de estimación puntual** en la muestra, se define por $\hat{e}_i = y_i - \hat{y}_i$, $i = 1,2,\ldots,n$, donde y_i es el valor observado de la variable Y en el punto (x_i, y_i) e $\hat{y}_i$ es su valor estimado.

✓ Los indicadores más usados **para medir calidad de ajuste** son los siguientes:

a. Correlación lineal de Pearson:

$$r_{X,Y} = \frac{Cov(X,Y)}{S_X S_Y} = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{S_X S_Y}$$

Permite medir la calidad del ajuste de un modelo **lineal** a la trayectoria de la nube de puntos experimentales. Este indicador hereda las cualidades de la covarianza (obsérvese que $S_X S_Y > 0$).

Propiedades

- $-1 \le r_{X,Y} \le 1$.
- $r_{X,Y} = r_{Y,X}$.
- **Si** $r_{X,Y} > 0$ la **relación lineal** entre X e Y es **directa**.
- $r_{X,Y} < 0$, la **relación lineal** entre X e Y es **indirecta**.
- Si $|r_{X,Y}| \approx 1$ la trayectoria de los puntos experimentales **se ajusta <u>aproximadamente</u> a un modelo lineal simple**.
- Si $|r_{X,Y}| = 1$ la trayectoria de los puntos experimentales se ajusta **perfectamente** a un modelo lineal simple.

- Si $|r_{X,Y}| \approx 0$ la trayectoria de los puntos experimentales **no se ajusta a la de un modelo lineal simple**.

b. **Error estándar de estimación**: $e.e.e. = \sqrt{\dfrac{\sum_{i=1}^{n} e_i^2}{n}} = S_e$

El **error estándar de estimación** permite medir la calidad del ajuste de cualquier modelo supuesto para explicar la trayectoria de la nube de puntos experimentales. Este indicador en las mismas unidades de la variable de respuesta.

c. **Coeficiente de determinación** R^2:

$$R^2 = \frac{S_{\hat{y}}^2}{S_y^2} \cdot 100\% = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} .100\%$$

El coeficiente de determinación se expresa en porcentaje. Se interpreta como el porcentaje de variabilidad explicada de la variable dependiente y, debida al modelo utilizado (en este caso un modelo lineal simple).

Si el modelo es lineal $R^2 = r_{xy}^2 100\%$.

2.2.1 Ejercicios resueltos, paso a paso:

Ejemplo 1.

(**Aplicación Ciencias de la Salud**)

En un estudio llevado a cabo en Italia, 10 pacientes con Hipertrigliceridemia se sometieron a una dieta baja en grasas y alta en carbohidratos para investigar si había alguna relación entre estas variables. La tabla muestra los valores antes de comenzar la dieta.

Paciente	Nivel de colesterol (mmol/l)	Nivel de triglicéridos (mmol/l)
1	5,12	2,3
2	6,18	2,54
3	6,77	2,95
4	6,65	3,77
5	6,36	4,18
6	5,9	5,31
7	5,48	5,53
8	6,02	8,83
9	10,34	9,48
10	8,51	14,2

a) Construya un diagrama de dispersión para estos datos.

b) ¿Existe evidencia de relación lineal entre los niveles de colesterol y triglicéridos antes de la dieta?

c) Estime el nivel de triglicéridos cuando el nivel de colesterol es de 6,10 mmol/l

d) ¿Cuál es el porcentaje de variación de nivel de triglicéridos explicada por el nivel de colesterol?

Esquema de desarrollo.

Paso 1. Leer problema:

- ✓ Se han hecho $n=10$ mediciones representadas por pares (x_i, y_i) en la tabla.
- ✓ Se estudian dos variables:

  X : Nivel de colesterol en el paciente.

  Y : Nivel de triglicéridos en el paciente.
- ✓ Se investiga alguna relación entre estas variables.

Paso 2. Identificar las variables:

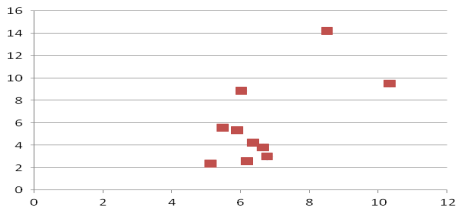
Variable independiente: X : Nivel de colesterol.

Variable dependiente: Y : Nivel de triglicéridos .

Paso 3. Desarrollar la respuesta mencionando los fundamentos y responder.

a. **Graficar los puntos** dados: Los valores del nivel de colesterol se identifican en el eje de abscisas X , y el nivel de triglicéridos (variable dependiente) en el eje de ordenadas Y . La nube de puntos experimentales se muestra en la gráfica de más abajo.

b. **Identificar el indicador que permite establecer la relación lineal** entre las dos variables:



Si la nube de puntos quiere ajustarse a un modelo lineal simple de la forma $y=a+bx$, se tiene:

Coeficiente de correlación lineal calculado en términos de los valores experimentales (x, y) con valores (x_i, y_i), $i=1,2,...,n$, puede ser escrito y calculado como:

$$r_{x,y}=\frac{Cov(x,y)}{S_x S_y}=\frac{\overline{xy}-\overline{x}\cdot\overline{y}}{S_x S_y}=0.65$$

Por el signo de $r_{x,y}$ (positivo) se deduce que existe una relación directa entre el nivel de colesterol y el nivel de triglicéridos antes de la dieta (crece x , crece y ; es decir, aumenta el nivel de colesterol, aumenta el nivel de triglicéridos).

Resp.: El valor del coeficiente de correlación lineal de Pearson es 0.65. No existe una evidencia fuerte de que un modelo lineal pueda ser un buen modelo para predecir la relación entre el nivel de colesterol y el nivel de triglicéridos. En todo caso tienen una tendencia a relacionarse de manera directamente proporcional.

c. **Estime el nivel de triglicéridos** cuando el nivel de colesterol es de 6,10 mmol/l.

Para efectuar una estimación es necesario haber decidido antes el modelo que se va a ajustar a los datos experimentales. Si fuera de la forma $y = a + bx$ (una línea recta) entonces se puede estimar los valores de a y b. Así,

$$\hat{b} = \frac{Cov(x, y)}{S_x^2} = 1.59 \qquad \hat{a} = \bar{y} - \hat{b} \cdot \bar{x} = -4.78$$

Por lo tanto, la ecuación de predicción es:

$$\hat{y} = \hat{a} + \hat{b} \cdot x \quad \Rightarrow \quad \boxed{\hat{y} = (-4.78) + (1.59) \cdot x}$$

En particular si $x = 6.1$:

$$\hat{y} = (-4.78) + (1.59)(6.1) = 4.92$$

Resp.: Si el nivel de colesterol es de $x = 6,1$ mmol/l, se estima que si la relación fuese lineal el valor estimado del nivel de triglicéridos es 4,92 mmol/l.

d. Porcentaje de variabilidad del nivel de triglicéridos explicada por el nivel de colesterol:

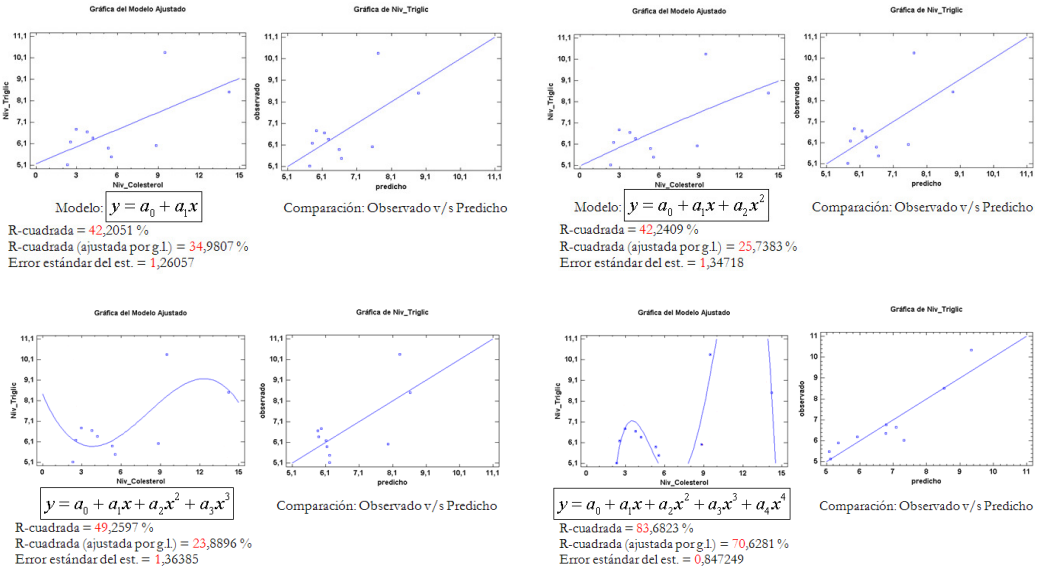
Identificar el indicador:

El indicador que se define como porcentaje de variación de nivel de triglicéridos explicada por el nivel de colesterol es el coeficiente de determinación R^2. Por lo tanto, se calcula

$$R^2 = \frac{S_{\hat{y}}^2}{S_y^2} 100\% = \frac{\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y} \right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y} \right)^2} .100 = 0.42\%$$

Resp.: El porcentaje de variación de nivel de triglicéridos explicada por el nivel de colesterol es de 0,42%. Bastante bajo. Debería buscarse un modelo que haga mejores predicciones. El modelo lineal simple es deficiente.

Observación: Si se observa bien, la gráfica de los puntos experimentales, ésta parece sugerir un comportamiento parabólico con el eje principal girado en algún ángulo. Si se ajustara un modelo de esa naturaleza, se podría obtener mejores resultados de calidad de modelo y de estimaciones. Se deja como tarea de investigación para el estudiante analizar la aplicación de un modelo parabólico con eje rotado.



Las gráficas muestran distintos modelos polinómicos ajustados a los puntos experimentales de este ejemplo. Arriba a la izquierda se muestra el ajuste de un modelo lineal simple y a la derecha uno parabólico (polinomial de grado 2). Pareciera que ambos son lo mismo, pero si se observa bien, en las gráficas de comparación los valores observados y predichos del nivel de triglicéridos están más cerca de la recta de igualdad. Además, si se lee el valor de R^2 se observará que en el modelo lineal será levemente menor (42,20 v/s 42,24). Lo mismo ocurre con el error estándar de estimación que en el modelo polinomial de grado 2 es de 0,847, menor que en el modelo lineal simple (1,363).

Las gráficas de abajo muestran ajustes a polinomios de grados 3 y 4. Puede verse que en estos casos disminuye el error estándar de estimación y aumenta el valor de R^2, lo cual significa que el ajuste a los puntos experimentales es mucho mejor, como se visualiza también en las gráficas de comparación.

No es la intención de estas notas llegar a un análisis muy acabado y por lo tanto se dejará este comentario hasta aquí para que el estudiante que esté motivado investigue sus dudas en libros de análisis de regresión lineal.

Ejemplo 2.

(**Aplicación en Ciencias de la Ingeniería y Ciencias de la Salud**)

Los datos de la tabla siguiente relacionan la solubilidad del nitrato de sodio Y ($NaNO_3$) con la temperatura del agua T (en °C). A la temperatura indicada T, Y partes de nitrato de sodio se disuelven en 100 partes de agua obteniendo

T	0	4	10	15	21	29	36	51	68
Y	66,7	71	76,3	80,6	85,7	92,9	99,4	114	125

Para estimar la solubilidad del nitrato de sodio se sugiere un modelo lineal de la forma $Y = mT + b$.

a. Usando el método de mínimos cuadrados obtenga los estimadores de los parámetros.

b. ¿Qué puede decir de la calidad del ajuste obtenido?

c. ¿Cree Ud. que a mayor temperatura existe mayor solubilidad del nitrato de sodio?

d. Determine el error estándar de estimación con el modelo ajustado y obtenga un intervalo de longitud dos errores estándar de estimación, centrado en la estimación de la solubilidad del nitrato de sodio cuando la temperatura es de 25°C.

Esquema de solución

Paso 1. Leer problema:

- ✓ Se entregan datos de $n = 9$ puntos bidimensionales de la forma (t, y).
- ✓ Se propone un modelo de la forma $Y = b + mT$ (modelo lineal simple).
- ✓ Se pide:
  - Encontrar una estimación de b ($\hat{b}$).
  - Encontrar una estimación de m ($\hat{m}$).
  - Investigar si las variables tienen una relación directa.
  - Determinar el error estándar de estimación con el modelo logrado.
  - Determinar un intervalo de longitud $2e.e.e$, centrado en la estimación de Y cuando $T = 25\,°C$.

Paso 2. Identificar las variables:

Variable independiente:

T: Solubilidad del agua.

Variable dependiente:

Y: Solubilidad del nitrato de sodio.

Paso 3. Desarrollar la respuesta mencionando los fundamentos y responder.

Se agrega a la tabla una fila que muestre los valores de los productos $T \cdot Y$ necesarios para cálculos posteriores. Se destaca también el cálculo de $\overline{T}$ (PromT), $\overline{Y}$ (PromY), $\overline{TY}$ (PromTY), S_T (ST) y S_Y (SY):

T	0	4	10	15	21	29	36	51	68
Y	66,7	71	76,3	80,6	85,7	92,9	99,4	113,6	125,1
TY	0	284	763	1209	1799,7	2694,1	3578,4	5793,6	8506,8

promT	26
promY	90,1444
promTY	2736,51

ST	21,2394
SY	18,5112

a. De la tabla obtenemos los valores de los parámetros:

$$\hat{m} = \frac{Cov(T,Y)}{S_T^2} = \frac{\overline{TY} - \overline{T}\cdot\overline{Y}}{S_T^2} = \frac{2736.5111 - 26\cdot 90.1444}{(21.2394)^2} = \frac{392.7565}{(21.2394)^2} = 0.8706$$

$$\hat{b} = \overline{Y} - \hat{m}\cdot\overline{T} = 90.1444 - (0.8706)\cdot(26) = 67.5088$$

Resp.: Los estimadores de m y b son $\hat{m} = 0.8706$ y $\hat{b} = 67.5088$.

b. Puesto que el modelo que se está ajustando es un modelo lineal de la forma $Y = b + mT$, el coeficiente de correlación lineal entre T e Y es

$$r_{T,Y} = \frac{392.7565}{21.239\cdot 18.5112} = 0.999$$

Resp.: El valor del coeficiente de correlación lineal entre T e Y es 0,999. Muy cercano a 1 y, por tanto, este modelo es casi perfecto para estimar los valores de Y sobre la base de los valores de X.

c. La covarianza permite averiguar si las variables T e Y tienen una relación directa. Por lo tanto, se calcula

$$Cov(T,Y) = \overline{TY} - \overline{T}\cdot\overline{Y} = 392.87 > 0$$

Resp.: Puesto que $Cov(T,Y)$ resulta positiva, la relación entre T e Y es directa, a mayor temperatura existe mayor solubilidad del nitrato de sodio.

d. Como el modelo **es lineal**, para hallar el error estándar estimado calculamos como sigue:

$$e.e.e. = s_y\sqrt{1 - r_{xy}^2} = 18.5112\cdot\sqrt{1 - (0.999)^2} = 0.8680$$

Para determinar el intervalo centrado en la estimación de y cuanto $t = 25$ °C, y con una longitud dos errores estándar se procede como sigue:

$$\left[\hat{y}_{t=25} - (e.e.e.), \hat{y}_{t=25} + (e.e.e.)\right]$$

$$\left[\left(0.8706\cdot 25+67.5088\right)-(e.e.e.),\left(0.8706\cdot 2.5+67.5088\right)+(e.e.e.)\right]$$

$$\left[89.2738-82.76,89.2738+82.76\right]$$

$$\left[6.5138,89.2738\right]$$

Resp.: El intervalo centrado en la estimación de la solubilidad del nitrato de sodio cuando la temperatura es de 25ºC es $\left[6.5138,89.2738\right]$.

Ejemplo 3.

(**Aplicación en Ciencias de la Ingeniería**)

Se decide investigar la relación entre dos variables: ingresos, I, y compras futuras con tarjetas de créditos, C. El ingreso es tomado de las solicitudes de tarjetas de crédito y las compras a crédito durante el primer mes de posesión de la tarjeta.

Una muestra aleatoria de 10 tenedoras de tarjetas de créditos, produce los datos siguientes (en miles de pesos):

Ingreso (I)	176	168	202	138	213	159	193	122	185	153
Compras (C)	448	556	844	427	811	398	447	154	534	313

a. ¿Es el modelo lineal un buen modelo para estimar las compras a crédito en base a los ingresos? Justifique su respuesta con un indicador estadístico y escriba el modelo ajustado con los parámetros estimados.
b. Determinar un intervalo centrado en la estimación de las compras para un ingreso de $240 000 de longitud 3 errores estándar de estimación.
c. ¿Cuáles el porcentaje de variación de las compras totales a créditos explicada por el ingreso a través de este modelo lineal ajustado?

Esquema de solución.

Paso 1. Leer problema:

- ✓ Se investiga la relación entre dos variables: Ingresos de una persona y sus compras con una tarjeta de crédito.
- ✓ Se dispone de $n=10$ datos bidimensionales de la forma (i,c).
- ✓ Se pide:

- Determinar si un modelo lineal es un buen modelo para efectuar predicciones del gasto con tarjeta de crédito.
- Un intervalo centrado en la estimación de las compras para un ingreso de $240 000, de longitud 3 errores estándar de estimación.
- El porcentaje de variación de las compras totales a créditos explicado por el ingreso a través de este modelo lineal ajustado.

Paso 2. Identificar las variables:

Variable independiente:

I : El ingreso de la persona.

Variable dependiente:

D : Compras efectuadas con la tarjeta de crédito.

Paso 3. Desarrollar la respuesta mencionando los fundamentos y responder.

a. La calidad del modelo lineal se puede calcular con el coeficiente de regresión lineal

$$r_{I,C} = 0.862,$$

Resp.: El coeficiente de correlación lineal entre el ingreso de la persona y las compras que efectúa con su tarjeta de crédito es de0.862. Indica que el modelo lineal es un modelo relativamente bueno para estimar las compras sobre la base de los ingresos.

b. Para encontrar el intervalo se necesita primero estimar los parámetros del modelo lineal $C = a + bI$. Se obtiene $\widehat{b} = 6.294$, $\widehat{a} = -582.400$. El modelo de predicción es $C = -582.400 + 6.294 \cdot I$

Como el intervalo está centrado en la estimación de las compras para un ingreso de $240.000 y debe tener longitud de 3 errores estándar de estimación.

Para $I = 240 \Rightarrow \hat{C} = 928.098$.

Con los datos de la tabla se calcula el error de estimación (¿cuánto vale $e.e.e.$?) y con su valor se establece el intervalo pedido:

$$I.C = \left[(928.098) - (1.5) \cdot (e.e.e.),\ (928.098) + (1.5) \cdot (e.e.e.)\right]$$

$$= \left[776.741,\ 1079.456\right].$$

Resp.: El intervalo centrado en la estimación de las compras para un ingreso de $240.000 y que tiene una longitud de 3 errores estándar de estimación es $\left[776.741,\ 1079.456\right]$ (en miles de pesos).

c. Coeficiente de determinación: $R^2 = r_{x,y}^2 = 0.74$ (¿Cómo se logró este valor?)

Resp.: El porcentaje de variación de las compras totales a crédito explicado por el ingreso a través de este modelo lineal ajustado es de 0.74.

2.8 Ejercicios Propuestos.

.

1. En la siguiente tabla, **x** es la fuerza de tensión aplicada a una varilla de acero, en cientos de kilos y t es la deformación (estiramiento) resultante en milésima de centímetro.

x	1	2	3	4	5
t	2	4	5	6	8

   Suponiendo una regresión lineal de y en x determine:
   a) La recta de mínimos cuadrados para los datos
   b) El error estándar de estimación
   c) ¿Qué porcentaje de la variación en las ventas mensuales se explica por la temperatura máxima promedio?

2. Los siguientes datos representan la estatura X(en cms.) y la circunferencia, Y(en cms.) de la cabeza de 10 bebes al momento de nacer

X	47	48	50	50	51	52	52	52	54	50
Y	35	34	33	35	34	36	36	37	38	35

   Para analizar estos datos se propone el modelo lineal.

$$Y = a + b \cdot X$$

   a) Hallar la recta de mínimos cuadrados
   b) ¿Cuál es el error de estimación, al estimar la circunferencia de la cabeza de un bebe de 51 cms.?
   c) Determine el error estándar de estimación
   d) Calcular e interpretar R^2

3. La tabla siguiente muestra la duración de cada erupción (x) del famoso géiser Old Faithful (viejo fiel) y el intervalo de tiempo transcurrido (y) ambos en minutos, hasta la siguiente erupción(Fuente Parque Nacional de Yellowstone)

x	y	x	y	x	y
1,8	56	2,82	73	4,27	90
1,82	58	3,13	6	4,3	84
1,88	60	3,27	77	4,3	89
1,9	60	3,65	77	4,43	84
1,92	60	3,7	82	4,43	89
1,93	56	3,78	79	4,47	86
1,98	59	3,83	85	4,53	89
2,03	60	3,87	81	4,55	86
2,05	57	3,88	80	4,6	88
2,37	61	4,1	89	4,6	92

a) ¿A mayor duración de las erupciones del Old Faithful mayor es el intervalo transcurrido entre ellas?

b) Suponga que se observa una erupción que dura 2 minutos y 40 segundos, ¿cuándo estima que se producirá la próxima? , use el modelo lineal.

2.3. Modelos reducibles a lineales.

Resumen de conceptos y propiedades:

✓ **Modelo Exponencial**: $y = a \cdot b^x$.

El modelo exponencial se convierte en un modelo lineal definiendo las nuevas variables $Y^* = \log(y)$, $a^* = \log(a)$ y $b^* = \log(b)$, pues en este caso, $Y^* = a^* + b^* x$, después de haber aplicado logaritmo decimal y obtenido: $\log(y) = \log(a) + \log(b) \cdot x$.

Así, el coeficiente de correlación lineal estará dado por

$$r_{(x,\log(y))} = \frac{Cov(x,\log(y))}{S_x \cdot S_{\log(y)}} = \frac{\overline{x\log(y)} - \bar{x} \cdot \overline{\log(y)}}{S_x \cdot S_{\log(y)}}$$

Los estimadores de a y b serán:

$$\log(b) = \frac{Cov(x, \log(y))}{S_x^2} \Rightarrow \boxed{\hat{b} = \text{antilog}\left(\frac{\text{cov}(x, \log(y))}{S_x^2}\right)}$$

$$\log(a) = \overline{\log(y)} - \hat{b} \cdot \overline{x} \Rightarrow \quad \boxed{\hat{a} = \text{antilog}\left(\overline{\log(y)} - \hat{b} \cdot \overline{x}\right)}$$

El modelo de predicción será: $\hat{y} = \hat{a} \cdot \hat{b}^x$.

Para estimar un valor de y dado un valor $x_0 \Rightarrow \hat{y} = \hat{a} \cdot \hat{b}^{x_0}$.

El error de estimación puntual está dado por $\hat{e}_i = y_i - \hat{y}_i$, $i = 1, 2, ..., n$.

El error estándar de estimación deberá calcularse con los errores $\hat{e}_i = y_i - \hat{y} = y_i - \hat{a} \cdot \hat{b}^{x_i}$, para cada valor de i. Así,

$$e.e.e. = \sqrt{\frac{\sum_{i=1}^{n} \hat{e}_i^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{a} \cdot x_i^{\hat{b}})^2}{n}} = S_e$$

- ✓ **Modelo Potencial:** $y = a \cdot x^b$.

El modelo potencial se convierte en un modelo lineal definiendo las nuevas variables $Y^* = \log(y)$, $a^* = \log(a)$, $b^* = b$ y $x^* = \log(x)$, pues en este caso, $Y^* = a^* + b^* \cdot x^*$. después de haber aplicado logaritmo decimal y obtenido: $\log(y) = \log(a) + b \cdot \log(x)$

Así, el coeficiente de correlación lineal estará dado por

$$r_{(\log(x), \log(y))} = \frac{Cov(\log(x), \log(y))}{S_{\log(x)} \cdot S_{\log(y)}} = \frac{\overline{\log(x)\log(y)} - \overline{\log(x)} \cdot \overline{\log(y)}}{S_{\log(x)} \cdot S_{\log(y)}}$$

Y las estimaciones de a y b serán:

$$\hat{b} = \frac{Cov(\log(x), \log(y))}{S_{\log(x)}^2}$$

$$\log(a) = \overline{\log(y)} - \hat{b} \cdot \overline{\log(x)} \Rightarrow \quad \boxed{\hat{a} = anti\log(\overline{\log(y)} - \hat{b} \cdot \overline{\log(x)})}$$

El modelo de predicción será: $\hat{y} = \hat{a} \cdot x^{\hat{b}}$.

Para estimar un valor de y dado un valor $x_0 \Rightarrow \hat{y} = \hat{a} \cdot x_0^{\hat{b}}$

El error estándar de estimación deberá calcularse con los errores $\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{a} \cdot x_i^{\hat{b}}$, para cada valor de i. Así,

$$e.e.e. = \sqrt{\frac{\sum_{i=1}^{n} \hat{e}_i^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{a} \cdot x_i^{\hat{b}})^2}{n}} = S_e$$

2.3.1 Ejercicio resuelto, paso a paso:

Ejemplo .

(**Aplicación en Ciencias de la Ingeniería**)

La tabla muestra valores experimentales de presión P de una masa de gas para distintos valores de volumen, V. De acuerdo con los principios de la termodinámica debe existir entre las variables una relación de la forma $PV^{\gamma} = C$, donde γ y C son constantes.

Volumen V(pul^3)	54,3	61,8	72,4	88,7	118,6	194
Presión P(lb/pul^2)	61,2	49,5	37,6	28,4	19,2	10,1

a. Determine el valor de las constantes γ y C.
b. Escribir el modelo de predicción de la presión.
c. Estimar P cuando $V = 100$ pulgadas cúbicas.

Esquema de solución:

Paso 1. Leer problema:

- ✓ Se entrega una muestra de $n = 6$ datos de una variable bidimensional (V, P) con valores (v_i, p_i), $i = 1, 2, ..., 6$.
- ✓ Se da una relación entre V y P en la forma $PV^{\gamma} = C$, que deberá ser transformada en un modelo lineal para aplicar los conocimientos que se tienen de este modelo.
- ✓ Se pide:
  - Determine el valor de las constantes γ y C.
  - Escribir el modelo de predicción de la presión P.
  - Estimar p cuando $v = 100$ pulgadas cúbicas.

Paso 2. Identificar variables:

Variable independiente: V : Volumen del gas (pul^3)

Variable dependiente: P : Presión del gas (lb/pul^2)

Paso 3. Desarrollar la respuesta mencionando los fundamentos y responder.

Aplicando logaritmo decimal:

$$PV^{\gamma} = C \quad \Rightarrow \quad P = \frac{C}{V^{\gamma}} \quad \Rightarrow \quad P = CV^{-\gamma} \quad \Rightarrow \quad \log P = \log C - \gamma \log V$$

Se define $P^* = \log P$, $a^* = \log C$, $b^* = -\gamma$ y $V^* = \log V$ para obtener el modelo lineal

$P^* = a^* + b^* V^*$ deducido de

$$\underbrace{\log P}_{P^*} = \underbrace{\log C}_{a^*} + \underbrace{(-\gamma)}_{b^*}\underbrace{\log V}_{V^*}$$

V	P	logV	logP	P_{est}	e = P - Pest
54,3	61,2	1,73479983	1,78675142	58,5250	2,6750
61,8	49,5	1,79098848	1,69460520	48,8024	0,6976
72,4	37,6	1,85973857	1,57518784	39,0752	-1,4752
88,7	28,4	1,94792362	1,45331834	29,3814	-0,9814
118,6	19,2	2,07408469	1,28330123	19,5396	-0,3396
194	10,1	2,28780173	1,00432137	9,7907	0,3093

Modelo: logP = logC - γ logV	
Coeficiente de correlación lineal	0,99861
Coeficiente de determinación R^2	0,99722
R^2 ajustado	0,99652
Intercepción = **a***	4,20335
Razón de cambio (pendiente) = **b***	-1,40420

Si se calcula los valores de los logaritmos para lograr los estimadores mínimo cuadráticos de a^* y b^*, se obtiene las tablas que se muestran a la derecha.

$$a^* = \log C = 4.20335$$

$$\Rightarrow \boxed{C = anti\log(4.20335) = 15971.8071}$$

$$-\gamma = -1.40420 \Rightarrow \boxed{\gamma = 1.40420}$$

Los estimadores de C y de γ permiten plantear la ecuación de regresión:

$$\hat{P} = \hat{C}V^{-\hat{\gamma}} \quad \Rightarrow \quad \boxed{\hat{P} = (15971.8071) \cdot V^{-1.40420}}$$

Con esta ecuación se puede lograr los valores de las estimaciones de P ($\hat{p}$ o p_{est}, que es lo mismo) que se muestran en la tabla superior junto al error cometido en cada caso experimental.

Resp.:

a. Los estimadores mínimo cuadráticos de C y γ son: $C = 15971.8071$ y $\gamma = 1.40420$.

b. El modelo de predicción de la presión es $\boxed{\hat{P} = (15971.8071) \cdot V^{-1.40420}}$.

c. La estimación de la presión, $\hat{p}$, cuando $v = 100$ pulgadas cúbicas es 24.8282. En efecto,

$$\hat{P} = (15971.8071) \cdot V^{-1.40420} = (15971.8071) \cdot (100)^{-1.40420} = 24.8282 \text{ pul}^3.$$

2.8 Ejercicios Propuestos.

1. La presión de la atmósfera decrece con la altitud. Al nivel del mar la presión media es una atmósfera (1,033227kg/cm²). La tabla muestra los valores de p a varias alturas (en km).

h	0	5	10	15	20	25
p	1	0,55	0,25	0,12	0,06	0,02

Para ajustar los datos se proponen los siguientes modelos:

$$i)\ e^{p-a} = h^b \qquad ii)\ p = a + bh$$

a. ¿Cuál es el modelo que mejor ajusta la serie de datos?
b. Con el modelo elegido estimar la presión a una altura de 13 km

2. En una experiencia de laboratorio se lograron los siguientes resultados de la variable h de cada valor de la variable d.

d	2	3	4	5	6	7
h	1.25	1.07	1.03	1.02	1.01	1.00

a. Determine cuál de los siguientes modelos es más adecuado para efectuar una regresión de h sobre d

$$i)\ h = e^{a+bd} \qquad 2i)\ h = ab^d$$

b. De acuerdo al modelo escogido obtenga los estimadores mínimos cuadráticos de a y b.
c. Calcular el error estándar de estimación con el modelo escogido
d. Calcular e interpretar el coeficiente de variación

3. Se pretende estimar los valores de producción Y (en miles de toneladas) de cierto material, en función del ,tiempo transcurrido X (en años) usando datos:

Tiempo (x)	1	2	3	4	5	6
Compras (y)	296	28	264	236	2	156

Para realizar las estimaciones se proponen los siguientes modelos

$$i)\ Y = A \cdot X^{2B} + 3 \qquad 2i)\ \frac{1 - BY}{Y} = \frac{A}{X}$$

a. Fundamentando con indicadores estadísticos pertinentes indique qué modelo de los propuestos escogería usted para efectuar mejores predicciones de la producción.
b. Determine el estimador mínimo cuadrático de A y de B en el modelo resultante de la predicción de la producción.
c. ¿Cuál es el error de estimación contenido en la predicción de la producción cuando han transcurrido 5 años?
d. ¿Cuál es el porcentaje de variación de la variación de la producción explicada por el tiempo trascurrido?

4. Se tomaron las medidas de longitud (L, centímetros) y peso (W, gramos) de 16 unidades de una especie de pez, el *Nemipterus marginatus*. Se muestran los datos en la tabla adyacente.

L	8,1	9,1	10,2	11,9	12,2	13,8	14,8	15,7
W	6,3	9,6	11,6	18,5	26,2	36,1	40,1	47,3

L	17	18	18,7	19	20,6	21,9	22,9	23,5
W	66	69	76,4	82,5	107	120	169	173

Se desea ajustar un modelo que explique la longitud en función del peso. Se postulan los siguientes tres modelos:

$$L = a + bW; \qquad L = a \cdot b^{W}; \qquad L = aW^{b}$$

a. ¿Cuál es el mejor modelo para ajustar esos datos? Justifique su respuesta, con el indicador adecuado?
b. Determine los parámetros a y b del modelo elegido.
c. Estime según el modelo seleccionado la longitud para un pez que pese 150 gramos.
d. Con el modelo elegido Determine el error estándar de estimación (e.e.e.).
e. Determine R^2.

5. Los valores siguientes corresponden al efecto del gasto cardiaco x medida en (lts/min) sobre el consumo metabólico de oxígeno y medida en (ml/min.)

x	0	1	2	3	4
y	4,3	9,89	22,75	53,32	120,34

a. Determine cuál de los siguientes modelos es más adecuado para estimar el consumo metabólico de oxígeno:

$$i)\ y = a + bx \qquad 2i)\ y = ab^{x} \qquad 3i)\ y + ax^{-b} = 4$$

b. De acuerdo al modelo escogido estime los parámetros a y b.
c. Determine el error estándar de estimación con el modelo escogido y obtenga un intervalo de longitud cuatro errores estándar de estimación para estimar el consumo metabólico de oxígeno en un paciente cuyo efecto del gasto cardiaco es de 7lts/min.

2.4. Comparación de modelos lineales y transformables a lineales.

Ajustar modelos matemáticos a nubes de puntos experimentales de manera que las predicciones de valores de las variables de respuesta tengan una aceptación sin lugar a dudas, es el objetivo de la temática de este capítulo. El estudio de ajustes por separado no tiene ninguna importancia si de paso no se demuestra que es el mejor

modelo para explicar la relación entre las variables. Y esto debe contemplar las características de las variables (conocidas por quienes trabajan frecuentemente con ellas) como para garantizar que si la nube de puntos amplía su dominio a rangos más amplios de valoración permitida de las variables predictoras, el modelo sigue vigente y aceptado. Debe o observarse que la mayor calidad del modelo que se acepte se muestra para valores de las variables predictoras en el rango de variación observada en los datos experimentales. Fuera de ese ámbito el modelo puede ser dudoso y de riesgosa aplicación.

Siempre que se encuentre un "buen modelo", puede aparecer otro mejor que disminuya el error de predicción. Si el modelo aceptado tiene vigencia en el tiempo, será siempre "el mejor entre los que se tiene a disposición". Es muchos casos esto basta, pero siempre es así. Para aprender a comparar modelos se ha introducido este apartado para comparar y decidir, de una forma sencilla, cuál merece ser llamado, en este nivel de de estudio, "el mejor".

No habrá resumen de conceptos y propiedades porque todo ha sido expuesto ya en los apartados anteriores.

2.4.1 Ejercicios resueltos, paso a paso:

Ejemplo 1.

(**Aplicación en Ciencias de la Salud**)

Los datos de la tabla representan la estatura X (en cms.) y la circunferencia, Y (en cms.), de la cabeza de bebés recién nacidos.

X	47	48	50	50	51	52	52	52	54	50
Y	35	34	33	35	34	36	36	37	38	35

10

a. ¿Puede afirmarse con base en los pares experimentales que "A mayor estatura mayor circunferencia de la cabeza de los bebes"?.

b. Para analizar estos datos se proponen los siguientes modelos:

$$i)\ Y = a + b \cdot X \qquad ii)\ Y = a \cdot b^{X} \qquad iii)\ Y = a \cdot X^{b}$$

¿Cuál es el mejor modelo para representar estos datos?

c. En cada modelo determine las estimaciones de los valores de los parámetros a y b, y escriba el modelo de predicción.

d. En "el mejor modelo" (el elegido por usted):

   i. ¿Cuál es el error de estimación de la circunferencia de la cabeza de un bebé que tiene 51 cms. de estatura?
   ii. Determine el error estándar de estimación (e.e.e.).

e. Si se propone un modelo cuadrático de la forma $Y = a_0 + a_1 X + a_2 X^2$, hallar su ecuación de predicción.

Esquema de solución.

Paso 1. **Leer el problema detenidamente hasta comprenderlo**.

- ✓ Se da un total de $n = 10$ puntos experimentales de la forma (x_i, y_i) que son observaciones de la variable bidimensional (X, Y).
- ✓ Se dan cuatro modelos matemáticos (incluyendo el de la pregunta d).
- ✓ Se pide comparar los tres primeros modelos y hallar la ecuación de `predicción del cuarto.

Paso 2. **Identificar los datos que entrega el problema**:

Variable independiente (X): Estatura de los bebés (en cms.).

Variable dependiente (Y): Circunferencia de la cabeza del bebé.

	X	Y	XY
	47	35	1645
	48	34	1632
	50	33	1650
	50	35	1750
	51	34	1734
	52	36	1872
	52	36	1872
	52	37	1924
	54	38	2052
	50	35	1750
SUMA	506	353	17881
PROMEDIO	50,6	35,3	1788,1

Paso 3. **Desarrollar la respuesta mostrando los fundamentos y responder**:

a. Un indicador estadístico adecuado para auscultar relación directa o indirecta entre variables es la covarianza entre las variables:

$$Cov(x, y) = \overline{xy} - \bar{x} \cdot \bar{y}$$
$$= 1788{,}1 - 50{,}2 \cdot 35{,}3 = 1{,}92 > 0$$

Podemos concluir, por tanto, que a mayor estatura mayor circunferencia de la cabeza de los bebés.

Estos mismos cálculos se pueden realizar con una calculadora científica.

Para el ingreso de los datos: MODE REG LIN,
Limpiar la memoria de la calculadora: SHIFT MODE SCL =
Ingreso de datos:

$$\begin{bmatrix} x_1, y_1 \to M^+ \\ \vdots \\ \vdots \\ \vdots \\ x_n, y_n \to M^+ \end{bmatrix} \Rightarrow \begin{bmatrix} 47,35 \to M^+ \\ \vdots \\ \vdots \\ \vdots \\ 50,35 \to M^+ \end{bmatrix}$$

Para obtener el valor de $\overline{x}$: SHIFT 2 1 = 50,6
Para obtener el valor de $\overline{y}$: SHIFT 2 $\rightarrow$ 1 = 70,6

Para obtener el valor de $\overline{xy}$, la calculadora no da la información directamente. Nos entrega la suma $\sum_{i=1}^{n} x_i y_i$, presionando SHIFT 1 $\rightarrow$ 3 = 17881. Este resultado se divide por el total de datos n, obteniendo el valor de $\overline{xy}$ =1788,1.

Luego

$$Cov(x, y) = \overline{xy} - \overline{x} \cdot \overline{y} = 1788,1 - 50,6 \cdot 35,3$$

.

	x	y	xy	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$
	47	35	1645	12,96	0,09
	48	34	1632	6,76	1,69
	50	33	1650	0,36	5,29
	50	35	1750	0,36	0,09
	51	34	1734	0,16	1,69
	52	36	1872	1,96	0,49
	52	36	1872	1,96	0,49
	52	37	1924	1,96	2,89
	54	38	2052	11,56	7,29
	50	35	1750	0,36	0,09
SUMA	506	353	17881	38,4	20,1
PROMEDIO	50,6	35,3	1788,1	3,84	2,01
	$\overline{x}$	$\overline{y}$	$\overline{xy}$	S^2_x	S^2_y
	S_x	1,95959179	S_y	1,41774469	

Resp.: Puesto que $Cov(x, y) > 0$ se deduce que a mayor estatura mayor circunferencia de la cabeza de los bebés.

b. Para decidir por el modelo más adecuado para relacionar las variables se utilizará el coeficiente de correlación. No obstante como a priori no se sabe cuál será el elegido, cada modelo es conveniente también dejar constancia de su modelo de predicción:

❖ **Modelo:** $y = a + bx$:

Como el primer modelo propuesto es el lineal luego

$$r_{xy} = \frac{Cov(x, y)}{S_x S_y} \cdot 100\% .$$

Esto nos ahorra de hacer este trabajo (si no usa calculadora adecuada)

Los mismos datos ingresados en la calculadora nos sirven para calcular r_{xy},

$$\boxed{r_{xy} = \frac{Cov(x, y)}{s_x s_y} = \frac{1,92}{1,960 \cdot 1,4177} == 0,6911}$$

Es conveniente estimar los valores de los parámetros a y b, por si el modelo lineal resulta ser el mejor.

$$\hat{b} = \frac{Cov(x, y)}{S_x^2} = \frac{1,92}{3,84} = 0,5 \quad ; \quad \hat{a} = \overline{y} - \hat{b}\overline{x} = 35,3 - 0,5 \cdot 50,6 = 10$$

Usando la calculadora:

Para obtener r_{xy}: SHIFT 2 $\rightarrow\rightarrow$ 3 = 0,69109, es decir $r_{xy} = 0.6911$

Valor de $\hat{b}$: SHIFT 2 $\rightarrow\rightarrow$ 2 =0,5, es decir $\hat{b} = 0.5$

Valor de $\hat{a}$: SHIFT 2 $\rightarrow\rightarrow$ 1 =10,
es decir $\hat{a} = 10$

Modelo lineal de predicción:

$$\boxed{\hat{y} = 10 + 0.5x}$$

x	y	log(y)	xlog(y)	$(\log y_i - \overline{\log y})^2$
47	35	1,54406804	72,5711981	1,128E-05
48	34	1,53147892	73,510988	0,00025433
50	33	1,51851394	75,925697	0,00083594
50	35	1,54406804	77,2034022	1,128E-05
51	34	1,53147892	78,1054248	0,00025433
52	36	1,5563025	80,92773	7,8781E-05
52	36	1,5563025	80,92773	7,8781E-05
52	37	1,56820172	81,5464897	0,0004316
54	38	1,5797836	85,3083142	0,00104697
50	35	1,54406804	77,2034022	1,128E-05
suma	506	353	15,4742662	783,230376	0,00301458
promedio	50,6	35,3	1,54742662	78,3230376	0,00030146
D. estándar	1,95959	1,41774			0,01736255

❖ **Modelo** $y = ab^x$

Aplicando logaritmo decimal, obtenemos

$$\log(y) = \log(a) + \big(\log(b)\big) \cdot x$$

Su coeficiente de correlación es, entonces, de la forma

$$r_{x,\log(y)} = \frac{\overline{x\log(y)} - \overline{x} \cdot \overline{\log(y)}}{s_x s_{\log y}}$$

Con base en la tabla de puntos experimentales se construye, por conveniencia una tabla como la adyacente.

$$\boxed{r_{x,\log(y)} = \frac{(78,3230376) \cdot (50,6)}{(1,95959179) \cdot (0,01736255)} = 0,68}$$

Usando calculadora:

Limpiar la memoria: SHIFT MODE SCL

Ingreso de datos:

$$\begin{array}{|l|} \hline x_1, \log y_1 \ \rightarrow M^+ \\ \vdots \\ \vdots \\ \vdots \\ x_n, \log y_n \ \rightarrow \ M^+ \\ \hline \end{array} \quad \Longrightarrow \quad \begin{array}{|l|} \hline 47, \log 35 \ \rightarrow M^+ \rightarrow n = 1 \\ \vdots \\ \vdots \\ \vdots \\ 50, \log 35 \ \rightarrow M^+ \rightarrow n = 10 \\ \hline \end{array}$$

Valor de $r_{X,\log Y}$: $Shift\,2 \rightarrow\rightarrow 3 = 0.68$, se obtiene $r_{X,\log Y} = 0.68$

Valor de $\hat{b}$: $Shift\,2 \rightarrow\rightarrow 2 = 0.006 \ \ shift \log answ = 1.014$,
es decir $\hat{b} = 1.014$

Valor de $\hat{a}$: $Shift\,2 \rightarrow\rightarrow 1 = 1.241 \quad shift \log answ = 17.42$,
es decir $\hat{a} = 17.42$

Modelo exponencial de predicción: $\boxed{\hat{y} = (17.42)\cdot(1.014)^x}$

❖ **Modelo** $y = ax^b$.

Aplicando logaritmo decimal, se obtiene $\log(y) = \log(a) + b \cdot \log(x)$

Luego, su coeficiente de correlación es de la forma

$$r_{\log(x),\log(y)} = \frac{\overline{\log(x)\log(y)} - \overline{\log(x)}\cdot\overline{\log(y)}}{s_{\log(x)} \cdot s_{\log(y)}}$$

Sobre la base de los datos originales, se crea una tabla conveniente como la de más abajo:

	log(x)	log(y)	log(x)*log(y)	$(\log(x_i)-\overline{\log(x)})^2$	$(\log(y_i)-\overline{\log(y)})^2$
	1,67209786	1,54406804	2,58183287	0,00100644	1,12801E-05
	1,68124124	1,53147892	2,57478551	0,0005099	0,000254329
	1,69897	1,51851394	2,57990964	2,3545E-05	0,000835943
	1,69897	1,54406804	2,62332529	2,3545E-05	1,12801E-05
	1,70757018	1,53147892	2,61510772	1,4047E-05	0,000254329
	1,71600334	1,5563025	2,6706203	0,00014838	7,87812E-05
	1,71600334	1,5563025	2,6706203	0,00014838	7,87812E-05
	1,71600334	1,56820172	2,6910394	0,00014838	0,000431605
	1,73239376	1,5797836	2,73680724	0,00081633	0,001046974
	1,69897	1,54406804	2,62332529	2,3545E-05	1,12801E-05
suma	17,0382231	15,4742662	26,3673736	0,00286249	0,003014583
promedio	1,70382231	1,54742662	2,63673736	0,00028625	0,000301458
D. estándar				0,01691889	0,017362555

Luego,

$$r_{\log(x),\log(y)} = \frac{2,63673736 - 1,70382231\cdot 1,54742662}{0,01691889\cdot 0,017362555} = 0,6718$$

Usando calculadora :

Limpiar la memoria: SHIFT MODE SCL

Ingresar datos: $\begin{array}{l}\log x_1, \log y_1 \rightarrow M^+ \\ \vdots \\ \vdots \\ \vdots \\ \log x_n, \log y_n \rightarrow M^+\end{array}$ $\Rightarrow$ $\begin{array}{l}\log 47, \log 35 \rightarrow M^+ \rightarrow n = 1 \\ \vdots \\ \vdots \\ \vdots \\ \log 50, \log 35 \rightarrow M^+ \rightarrow n = 10\end{array}$

Valor de $r_{\log X,\log Y}$: $Shift\,2 \rightarrow\rightarrow 3 = 0.6718$, es decir $r_{\log x,\log y} = 0.6718$

Puesto que, hasta aquí, hemos ya calculado los coeficientes de correlación lineal en los modelos linealizados pudiendo, por tanto, decidir cuál es el mejor sin necesidad de dar más detalles:

Resp.: Si se comparan los coeficientes de correlación lineal de cada modelo se observa que el mayor, **en valor absoluto**, es el del modelo lineal. Por tanto ese es el modelo que se escoge para predecir la circunferencia de la cabeza en función de la estatura de un niño.

c. Se pide, en cada modelo, determinar las estimaciones de los valores de los parámetros a y b, y escribir el modelo de predicción.

Puesto que en el desarrollo anterior ya han sido calculados estos valores:

Resp.:

- Modelo $y = a + bx$: $\hat{b} = 0,5$; $\hat{a} = 10$.
- Modelo $y = ab^x$: $\hat{b} = 1.014$; $\hat{a} = 17.42$.
- Modelo $y = ax^b$: $\hat{b} = 0.6895$; $\hat{a} = 2,3589$.

Puesto que los valores del tercer modelo mencionado enla respuesta nohabían sido calculados se detalla a continuación la determinación de $\hat{b}$ y $\hat{a}$:

$$\hat{b} = \frac{\overline{\log(x)\log(y)} - \overline{\log(x)} \cdot \overline{\log(y)}}{s^2_{\log(x)}}$$

$$= \frac{2,63673736 - 1,70382231 \cdot 1,5474266}{(0,01691889)^2} = 0.6895$$

Con calculadora: $Shift\,2 \rightarrow\rightarrow 2 =$, es decir $\boxed{\hat{b} = 0.6895}$.

$$\log(a) = \overline{\log(y)} - \hat{b} \cdot \overline{\log(x)} = 1,5474266 - 0,6895 \cdot 1,70382231 = 0,3727$$

$$\hat{a} = anti\log\left(\overline{\log(y)} - \hat{b} \cdot \overline{\log(x)}\right) = 2,3589$$

Con calculadora: $Shift\,2 \rightarrow\rightarrow 1 = 0,37271258$

$shift\log answ = 2,3589$, es decir $\boxed{\hat{a} = 2,3589}$

Modelo de predicción: $\hat{y} = (2,3589) \cdot x^{(0,6895)}$.

d. Se pide en el modelo elegido:

i. Determinar el error de estimación de la circunferencia de la cabeza de un bebé que tiene 51 cms. de estatura.

Para $x = 51 \Rightarrow \hat{y} = (10) + (0,5) \cdot (51) = 35,5\ cm$
$\Rightarrow \hat{e} = (34) - (35,5) = -1,5$.

Resp.: El error de estimación de la circunferencia de la cabeza de un bebé que tiene 51 cms. de estatura es de $-1,5$ cms. El signo negativo significa que el valor observado de y, $y_i = 34$, está por debajo del valor estimado, $\hat{y} = 35,5\ cm$.

x	y	y_{est}	$(y-y_{est})^2$
47	35	33,5	2,25
48	34	34	0
50	33	35	4
50	35	35	0
51	34	35,5	2,25
52	36	36	0
52	36	36	0
52	37	36	1
54	38	37	1
50	35	35	0
suma	506	353	353	10,5
promedio	50,6	35,3		1,05

ii. Determinar el error estándar de estimación (e.e.e.).

Creando una tabla como la adyacente se puede calcular *e.e.e.*

$$e.e.e. = \sqrt{\frac{\sum_{i=1}^{n} \hat{e}_i^2}{n}} = \sqrt{1,05} = 1,0247$$

Resp.: El error estándar de estimación de este modelo es de 1,0247 cms.

Ejemplo 2.

(**Aplicación en Ciencias de la Ingeniería**)

La tabla muestra la fuerza de tensión aplicada X a una probeta de acero, en miles de libras, y la elongación resultante, en milésimas de pulgadas, Y.

X	1	2	3	4	5	6
Y	14	33	40	63	76	85

a) Interpretar el dato (3,40)
b) Se proponen los siguientes modelos

$$i)\ \frac{1}{y-20} = a + b(x+5) \qquad ; \qquad ii)\ y = 10 + ab^x$$

Usando un indicador estadístico justifique qué modelo es el más adecuado para ajustar estos datos.

c) Del modelo seleccionado estime los parámetros mínimos cuadráticos y escriba la ecuación del modelo de ajuste.
d) ¿Cuál es el error de estimación de la elongación cuando la tensión aplicada a una probeta de acero es de 3 miles de libras?

Paso 1. Leer problema:

- ✓ Se dan $n=6$ datos experimentales.
- ✓ Se pide:
  a. **Interpretar** el punto (3,40).
  b. Dados dos modelos determinar cuál de los dos es **el mejor** para hacer pronósticos sobre valores de Y, $\hat{y}$.
  c. En el modelo seleccionado, **estimar** a y b. Determinar, además, el **modelo de predicción**.
  d. Calcular el **error de estimación** de la elongación cuando la tensión aplicada a una probeta de acero es de 3 miles de libras

Paso 2. Identificar las variables:

Variable independiente (X): Fuerza de tensión aplicada a una probeta (miles de libras).

Variable dependiente (Y): Elongación resultante (en milésimas de pulgada).

Paso 4. Desarrollar la respuesta mostrando los fundamentos y responder:

1. Interpretación del punto (3,40):

   Resp.: La fuerza de tensión de 3 mil libras aplicada a una probeta de acero produce una elongación de 40 milésima de pulgada.

	x	y	x+5	1/(y-20)	(x+5)*1/(y-20)
	1	14	6	-0,16666667	-1,00000002
	2	33	7	0,07692308	0,53846156
	3	40	8	0,05	0,4
	4	63	9	0,02325581	0,20930229
	5	76	10	0,01785714	0,1785714
	6	85	11	0,01538462	0,16923082
suma	21		51	0,01675398	0,49556605
promedio	41		8,5	0,00279233	0,082594342
D. estándar			1,70783	0,07878722	

2. El mejor modelo entre los dos propuestos:

$$i)\ \frac{1}{y-20}=a+b(x+5)$$

$$\underbrace{\frac{1}{y-20}}_{y'}=a+b\underbrace{(x+5)}_{x'}$$

$$\Rightarrow y'=a+bx'$$

$$\Rightarrow r_{\left(x+5,\frac{1}{y-20}\right)}=\frac{\overline{(x+5)\cdot 1/(y-20)}-\overline{(x+5)}\cdot\overline{1/(y-20)}}{s_{(x+5)}\cdot s_{1/(y-20)}}=0.4374$$

ii) $y = 10 + ab^x \Rightarrow y - 10 = ab^x \Rightarrow \log(y-10) = \log(a) + \log(b)\cdot x$

$$\Rightarrow r_{(x,\log(y-10))} = \frac{\overline{x\cdot\log(y-10)} - \overline{x}\cdot\overline{\log(y-10)}}{s_x s_{\log(y-10)}} = 0,9$$

	x	y	log(y-10)	xlog(y-10)
	1	14	0,60205999	0,60205999
	2	33	1,36172784	2,72345568
	3	40	1,47712125	4,43136375
	4	63	1,72427587	6,89710348
	5	76	1,81954394	9,0977197
	6	85	1,87506126	11,2503676
suma	21		8,85979015	35,0020702
promedio	3,5		1,47663169	5,83367836
D. Estándar	1,70782		0,4312272	

Resp.:

El modelo más adecuado es el modelo $y = 10 + ab^x$.

3. Estimadores mínimo cuadráticos del mejor modelo y modelo de predicción:

$$\log(b) = \frac{\overline{x\cdot\log(y-10)} - \overline{x}\cdot\overline{\log(y-10)}}{s_x^2} = 0,22816026$$

$$\Rightarrow \hat{b} = anti\log(b) = 1,6911$$

$$\log(a) = \overline{\log(y-10)} - \hat{b}\cdot\overline{\log(x)} = 0,67807076$$

$$\Rightarrow \hat{a} = anti\log(\overline{\log(y-10)} - \hat{b}\cdot\overline{\log(x)}) = 4,7651$$

Luego el modelo de predicción es $\boxed{\hat{y} = 10 + (4,7651)\cdot(1,6911)^x}$

Resp.: Los estimadores mínimo cuadráticos de a y b son $\hat{a} = 4,7651$ y $\hat{b} = 1,6911$. El modelo de predicción es $\hat{y} = 10 + (4,7651)\cdot(1,6911)^x$.

4. $x = 3 \Rightarrow \hat{y} = 10 + (4,7651)\cdot(1,6911)^3 = 33,0437$

$x = 3 \Rightarrow y = 4$ (de la tabla de valores dada).

Luego el error de estimación cuando $x = 3$ es

$$\hat{e}_{x=40} = y - \hat{y} = (40) - (33,0437) = 6,9563$$

Resp.: El error de estimación de la elongación cuando la tensión aplicada a una probeta de acero es de 3 miles de libras es de 6,96 mm aprox.

2.4.2 Ejercicios Propuestos.

1. Se tomaron las medidas de la longitud L (en centímetros) y el peso W (en gramos) de 16 unidades de una especie de pez, el *Nemipterus marginatus*. Se muestran los datos en la tabla adyacente.

L	8,1	9,1	10,2	11,9	12,2	13,8	14,8	15,7
W	6,3	9,6	11,6	18,5	26,2	36,1	40,1	47,3

L	17	18	18,7	19	20,6	21,9	22,9	23,5
W	66	69	76,4	82,5	107	120	169	173

   Se desea ajustar un modelo que explique la longitud en función del peso. Se postulan los siguientes tres modelos:

$$L = a + bW; \qquad L = a \cdot b^{W}; \qquad L = aW^{b}$$

   a. ¿Cuál es el mejor modelo para ajustar esos datos? Justifique su respuesta, con el indicador adecuado?
   b. Determine los parámetros a y b del modelo elegido.
   c. Estime según el modelo seleccionado la longitud para un pez que pese 150 gramos.
   d. Con el modelo elegido Determine el error estándar de estimación (e.e.e.).
   e. Determine e interprete el valor de R^2.

2. La tabla muestra la duración de cada erupción (X) del famoso géiser Old Faithful (Viejo Fiel) y el intervalo de tiempo transcurrido (Y) ambos en minutos, hasta la siguiente erupción (Fuente Parque Nacional de Yellowstone)

x	y	x	y	x	y
1,8	56	2,82	73	4,27	90
1,82	58	3,13	6	4,3	84
1,88	60	3,27	77	4,3	89
1,9	60	3,65	77	4,43	84
1,92	60	3,7	82	4,43	89
1,93	56	3,78	79	4,47	86
1,98	59	3,83	85	4,53	89
2,03	60	3,87	81	4,55	86
2,05	57	3,88	80	4,6	88
2,37	61	4,1	89	4,6	92

   a. ¿A mayor duración de las erupciones del Old Faithful mayor es el intervalo de tiempo transcurrido entre ellas?.

   b. Suponga que se observa una erupción que dura 2 minutos y 40 segundos. Use un modelo lineal para estimar cuándo se producirá la próxima.

3. Los valores de la tabla corresponden al efecto del gasto cardiaco X (lts/min) sobre el consumo metabólico de oxígeno (ml/min.)

x	0	1	2	3	4
y	4,3	9,89	22,75	53,32	120,3

   Determine cuál de los siguientes modelos es más adecuado para estimar el consumo metabólico de oxígeno:

$$i)\, y = a + bx \qquad ii)\, y = ab^{x} \qquad iii)\, y + ax^{-b} = 4$$

   a. De acuerdo al modelo escogido estime los parámetros a y b .
   b. Determine el error estándar de estimación con el modelo escogido y obtenga un intervalo de longitud cuatro errores estándar de estimación para estimar el consumo metabólico de oxígeno en un paciente cuyo efecto del gasto cardiaco es de 7lts/min.

4. Se pretende estimar los valores de producción Y (en miles de toneladas) de cierto material, en función del ,tiempo transcurrido X (en años) usando los datos de la tabla.

Tiempo (x)	1	2	3	4	5	6
Compras (y)	296	284	264	236	201	156

Para realizar las estimaciones se proponen los siguientes modelos

$$i)\ \ Y = A \cdot X^{2B} + 3 \qquad ii)\ \ \frac{1 - BY}{Y} = \frac{A}{X}$$

a. Fundamentando con indicadores estadísticos pertinentes indique qué modelo de los propuestos escogería usted para efectuar mejores predicciones de la producción.
b. Determine el estimador mínimo cuadrático de A y de B en el modelo de predicción de la producción.
c. ¿Cuál es el error de estimación contenido en la predicción de la producción cuando han transcurrido 5 años?.
d. ¿Cuál es el porcentaje de variación de la variación de la producción explicada por el tiempo trascurrido.

2.5. Modelo parabólico.

Resumen de conceptos y propiedades:

✓ El **modelo cuadrático** está representado por la ecuación general de segundo grado

$$Y = a + bX + cX^2,$$

donde a sigue siendo la intersección con el eje de ordenadas Y, y b y c están relacionados con la pendiente y la tasa de variación.

✓ La **ecuación de predicción** de la parábola que ajusta un conjunto de n puntos de la forma (x_i, y_i), $1 = 1, 2, ..., n$, tiene por ecuación:

$$\hat{y} = \hat{a} + \hat{b}x + \hat{c}x^2 \quad (*)$$

donde $\hat{a}$, $\hat{b}$ y $\hat{c}$ son los **estimadores mínimo cuadráticos** de a, b y c determinados al resolver el sistema de **ecuaciones normales**:

$$\begin{vmatrix} na+b\sum_{i=1}^{n}x_i+c\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}y_i \\ a\sum_{i=1}^{n}x_i+b\sum_{i=1}^{n}x_i^2+c\sum_{i=1}^{n}x_i^3=\sum_{i=1}^{n}x_iy_i \\ a\sum_{i=1}^{n}x_i^2+b\sum_{i=1}^{n}x_i^3+c\sum_{i=1}^{n}x_i^4=\sum_{i=1}^{n}x_i^2y_i \end{vmatrix}$$

Si Δ: matriz de coeficientes.

Δ_a : matriz de coeficientes en la que la columna de coeficientes de a es reemplazada por la columna de los segundos miembros de las ecuaciones.

Δ_b : matriz de coeficientes en la que la columna de coeficientes de b es reemplazada por la columna de los segundos miembros de las ecuaciones.

Δ_c : matriz de coeficientes en la que la columna de coeficientes de c es reemplazada por la columna de los segundos miembros de las ecuaciones.

Entonces $\hat{a}=\dfrac{\Delta_1}{\Delta}$ $\hat{b}=\dfrac{\Delta_2}{\Delta}$ $\hat{c}=\dfrac{\Delta_3}{\Delta}$.

✓ Para cada $x=x_0$ la estimación de y está dada por: $\hat{y}_i=\hat{a}+\hat{b}x_i+\hat{c}x_i^2$.

2.5.1 Ejercicios resueltos, paso a paso:

Ejemplo 1.

(**Aplicación en Ciencias Medioambientales**)

La siguiente tabla muestra el número medio de acres por granja en los Estados Unidos en una selección de varios años (Fuente Bureau of Labor Statistics.)

Año	1950	1960	1970	1980	1990	1994
Acre	213	297	374	426	461	478

a) Determinar el modelo cuadrático $y=a+bt+ct^2$ que ajuste los datos, donde y es número medio de acres y t el año, tomando como origen $t=0$, en 1950.
b) Estime el número medio de acres para el año 2000.
c) Calcular e interpretar R^2.

Paso 1. Leer problema:

- ✓ Se dan $n=6$ datos bidimensionales de la forma (T,Y), con valores (t,y) contenidos en la tabla.
- ✓ Se pide
  - Ajustar un modelo cuadrático a los datos. Es necesario estimar sus parámetros a, b y c.
  - Estimar y cuando $t=2000$.
  - Calcular e interpretar el coeficiente de determinación.

Paso 2. Identificar las variables:

Variable independiente: Año en que se registran los datos (T)
Variable dependiente: Número medio de acres otorgados por granja (Y).

Paso 3. Desarrollar la respuesta mostrando los fundamentos y responder:

a) Determinar las constantes del modelo $y=a+bt+ct^2$:

Ecuaciones normales y cálculos:

$$\begin{aligned} na+b\sum_{i=1}^{n}t_i+c\sum_{i=1}^{n}t_i^2&=\sum_{i=1}^{n}y_i\\ a\sum_{i=1}^{n}t_i+b\sum_{i=1}^{n}t_i^2+c\sum_{i=1}^{n}t_i^3&=\sum_{i=1}^{n}t_iy_i\\ a\sum_{i=1}^{n}t_i^2+b\sum_{i=1}^{n}t_i^3+c\sum_{i=1}^{n}t_i^4&=\sum_{i=1}^{n}t_i^2y_i \end{aligned}$$

	y	t	t^2	t^3	t^4	ty	t^2y
	213	0	0	0	0	0	0
	298	1	1	1	1	298	298
	374	2	4	8	16	748	1496
	426	3	9	27	81	1278	3834
	461	4	16	64	256	1844	7376
	478	4,4	19,36	85,18	374,8	2103	9254
Suma	2250	14,4	49,36	185,2	728,8	6271	22258

Remplazando las sumas en las ecuaciones normales:

$$\begin{aligned} &i)\quad 6a+14,4b+49,36c=2250\\ &2i)\quad 14,4a+49,36b+185,2c=6271\\ &31)\quad 49,36a+185,2b+728,8c=22258 \end{aligned}$$

Usando método de determinantes para resolver dicho sistema

$$\Delta=\begin{vmatrix} 6 & 14,4 & 49,36\\ 14,4 & 49,36 & 185,2\\ 49,36 & 185,2 & 728,8 \end{vmatrix}=1936,42$$

$$\Delta_{a_0} = \begin{vmatrix} 2250 & 14,4 & 49,36 \\ 6271 & 49,36 & 185,2 \\ 22258 & 185,2 & 728,8 \end{vmatrix} = 411279,955 \Rightarrow \frac{\Delta_a}{\Delta} = \hat{a} = 212,39$$

$$\Delta_{a_1} = \begin{vmatrix} 6 & 2250 & 49,36 \\ 14,4 & 6271 & 185,2 \\ 49,36 & 22258 & 728,8 \end{vmatrix} = 185836,8704 \Rightarrow \frac{\Delta_b}{\Delta} = \hat{b} = 95,97$$

$$\Delta_c = \begin{vmatrix} 6 & 14,4 & 2250 \\ 14,4 & 49,36 & 6271 \\ 49,36 & 185,2 & 22258 \end{vmatrix} = -15939,936 \Rightarrow \frac{\Delta_c}{\Delta} = \hat{c} = -8,23$$

Resp.: El modelo de predicción está dado por $\boxed{\hat{y} = 212,39 + (95,97)t - (8,23)t^2}$

b) $t = 5 \Rightarrow \hat{y} = 212,39 + (95,97)\cdot(5) - (8,23)\cdot(5)^2 = 486,49$

Resp.: El número medio de acres estimado para el año 2000 es de 486,49.

c) Se completa la tabla anterior:

	y	t	t^2	t^3	t^4	ty	t^2y	y_{est}	e_{est}	$(Y_{est}-Prom(Y))^2$	e^2_{est}
	213	0	0	0	0	0	0	212,39	0,61	26442	0,3721
	298	1	1	1	1	298	298	300,13	-2,13	5606	4,5369
	374	2	4	8	16	748	1496	371,41	2,59	13	6,7081
	426	3	9	27	81	1278	3834	426,23	-0,23	2625	0,0529
	461	4	16	64	256	1844	7376	464,59	-3,59	8026	12,888
	478	4,4	19,36	85,18	374,8	2103	9254	475,3252	2,6748	10065	7,1546
Sumas	2250	14,4	49,36	185,2	728,8	6271	22258	2250,08	-0,08	52776,44	0,0057
Promedio Y	375				Promedio y_{est}			375,01	R^2 = 0,999932605		e.e.e.
$Suma(Y-PromY)^2$	52780				$Suma(Y_{est}-PromY)^2$			52776,44			0,0307
Parámetros	a_{est} = 212,39		b_{est} = 96		c_{est} = 8,23		n = 6				

Y se calcula el coeficiente de determinación:

$$R^2 = \frac{\sum_{i=1}^{6}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{6}(y_i - \overline{y})^2} \cdot 100\% = \frac{52776.44}{52780} \cdot 100\% = 99.9933\%$$

Resp.: Prácticamente el 100% de la variabilidad del **Número medio de acres** otorgados, está explicada por la variable **Año** a través del modelo parabólico.

2.5.2 Ejercicios Propuestos.

1. El dueño de una distribuidora de automóviles piensa que la relación entre el número Y de autos nuevos vendidos por él en un mes dado y el número X de anuncios de su distribuidora en el diario local durante ese mes, está dada por el modelo

$$y = a + bx + cx^2$$

En la tabla aparecen datos correspondientes a los últimos seis meses. Ajustar estos datos con el modelo cuadrático planteado.

Mes	1	2	3	4	5	6
Y	10	10	15	20	30	40
X	0	1	2	2	3	4

2. Dados los datos de la tabla adyacente:

x	1,2	1,8	3,1	4,9	5,7	7,1	8,6	9,8
y	4,5	5,9	7	7,8	7,2	6,8	4,5	2,7

   a. Determine la ecuación cuadrática de predicción que se ajusta a los datos.
   b. Estime un valor de y cuando x =8,0.
   c. Determine e interprete R^2 .

3. El número y de bacterias por unidad de volumen presentes en un cultivo después de X horas está dado en la tabla.

x	0	1	2	3	4	5	6
y	30	39	58	60	90	130	190

   a. Determine la ecuación cuadrática que ajuste estos datos.
   b. Estime un valor de y cuando x =7.
   c. Determine e interprete R^2 .

2.6. Modelo de regresión múltiple.

Resumen de conceptos y propiedades:

✓ Cuando se usa más de una **variable predictora** (o independiente) para predecir los valores de una **variable de respuesta** (o dependiente), el proceso se llama **análisis de regresión múltiple**.

✓ El análisis de regresión múltiple incluye el uso de ecuaciones **lineales** y **no lineales**. En estas notas sólo nos referiremos a ecuaciones de regresión **lineal**, de la forma

$$y = a_0 + a_1 x_1 + a_2 x_2 + \cdots + a_k x_k + e$$

donde y es el valor de la **variable de respuesta** Y, y $x_1, x_2, \ldots, x_k$ son k variables con las que se sospecha se puede predecir el valor de Y (variables predictoras). Los coeficientes $a_1, a_2, \ldots, a_k$ representan ponderaciones de los aportes de las variables predictoras y a_0 es un aporte básico general.

- ✓ Para realizar estimaciones de y, se debe estimar, con los puntos experimentales, las estimaciones de los **parámetros** $a_0, a_1, a_2, \ldots, a_k$ del modelo.

- ✓ El método de mínimos cuadrados permite encontrar los estimadores de los parámetros $a_0, a_1, a_2, \ldots, a_k$ las ecuaciones normales.

- ✓ En el caso de dos variables predictoras $(k = 2)$ el sistema de ecuaciones normales está dado por el sistema de ecuaciones normales:

$$\begin{aligned} na_0 + a_1\sum_{i=1}^{n} x_{i1} + a_2\sum_{i=1}^{n} x_{i2} &= \sum_{i=1}^{n} y_i \\ a_0\sum_{i=1}^{n} x_{i1} + a_1\sum_{i=1}^{n} x_{i1}^2 + a_2\sum_{i=1}^{n} x_{i1}x_{i2} &= \sum_{i=1}^{n} x_{i1}y_i \\ a_0\sum_{i=1}^{n} x_{i2} + a_1\sum_{i=1}^{n} x_{i1}x_{i2} + a_2\sum_{i=1}^{n} x_{i2}^2 &= \sum_{i=1}^{n} x_{i2}y_i \end{aligned}$$

Resolviendo este sistema se encuentran las estimaciones $\hat{a}_0$, $\hat{a}_1$ y $\hat{a}_2$. Con estas estimaciones se logra el modelo de predicción

$$\hat{y} = \hat{a}_0 + \hat{a}_1 x_1 + \hat{a}_2 x_2$$

2.6.1 Ejercicios resueltos, paso a paso:

Ejemplo 1.

(**Aplicación en Ciencias Sociales**)

1. El objetivo de los siguientes datos es el relacionar Y, número de horas diarias frente al televisor, con el estado civil X y la edad Z, en una muestra de 20 personas escogidas al azar para observar del vector (Y, X, Z) con valores (y, x, z) mostrados en la tabla. El estado civil se encuentra codificado: casada(o)=1, soltera(o) = 0).

   a) Escriba las ecuaciones normales.

b) Encuentre la ecuación de predicción de mínimos cuadrados para los datos obtenidos.
c) Estime las horas diarias frente al televisor de una persona casada de 60 años.

Horas (Y)	Estado civil (x)	Edad (z)
1,1	1	73
1,6	1	66
2	0	65
2,5	0	65
2,8	1	68
2,8	1	69
3	1	82
3	1	83
3	1	82
3,2	0	72
3,2	1	69
3,3	0	71
3,3	0	71
3,4	0	80
3,5	0	73
3,6	0	75
3,6	0	78
3,7	0	66
3,9	1	80
4	0	55

Esquema de desarrollo:

Paso 1. Leer el problema:

- ✓ Se toma una muestra de $n=20$ datos de una variable tridimensional (Y,X,Z) con valores (y_i,x_i,z_i), $i=1,2,...,20$, se muestran en una tabla.
- ✓ Se quiere pronosticar el valor de Y sobre la base de valores de X y de Z.
- ✓ Se pide:
  - Escribir las ecuaciones normales.
  - Determinar la ecuación de predicción.
  - Hacer una estimación de Y para valores dados de X y de Z.

Paso 2. Identificar las variables:

Variables predictoras (independientes):

X : Estado civil de la persona (0: soltera, 1: casada).
Z : Edad de la persona.

Variable de respuesta (dependiente):

Y : Número de horas diarias que la persona pasa frente al televisor.

Paso 4. Desarrollar la respuesta mostrando los fundamentos y responder: Usando la tabla para determinar las sumas ,

a. Las ecuaciones normales son:

$$\begin{aligned} na_0 + a_1\sum_{i=1}^{n} x_{i1} + a_2\sum_{i=1}^{n} x_{i2} &= \sum_{i=1}^{n} y_i \\ a_0\sum_{i=1}^{n} x_{i1} + a_1\sum_{i=1}^{n} x_{i1}^2 + a_2\sum_{i=1}^{n} x_{i1}x_{i2} &= \sum_{i=1}^{n} x_{i1}y_i \\ a_0\sum_{i=1}^{n} x_{i2} + a_1\sum_{i=1}^{n} x_{i1}x_{i2} + a_2\sum_{i=1}^{n} x_{i2}^2 &= \sum_{i=1}^{n} x_{i2}y_i \end{aligned}$$

Reemplazando los valores de de la tabla de cálculos anterior:

Resp.: Las ecuaciones normales de este ajuste son

$$\begin{array}{lll} \text{i)} & 20a_0 + 9\ a_1 + 1443\ a_2 = 605 \\ \text{ii)} & 9\ a_0 + 9\ a_1 + 672\ a_2 = 24,41443 \\ \text{iii)} & a_0 + 672a_1 + 105103a_2 = 4377,3 \end{array}$$

b. El sistema lo resolvemos a través de determinantes

$$\Delta = \begin{vmatrix} 20 & 9 & 1443 \\ 9 & 9 & 672 \\ 1443 & 672 & 105103 \end{vmatrix} = 87804$$

$$\Delta_{a_0} = \begin{vmatrix} 60,5 & 9 & 1443 \\ 24,4 & 9 & 672 \\ 4377,3 & 672 & 105103 \end{vmatrix} = 113630,4 \Rightarrow b_0 = \frac{\Delta_{b_0}}{\Delta} = 1,29414$$

$$\Delta_{a_1} = \begin{vmatrix} 20 & 60,5 & 1443 \\ 9 & 24,4 & 672 \\ 1443 & 4377,3 & 105103 \end{vmatrix} = -61504 \Rightarrow b_0 = \frac{\Delta_{b_1}}{\Delta} = 0,700469$$

$$\Delta_{a_2} = \begin{vmatrix} 20 & 9 & 60,5 \\ 9 & 9 & 24,4 \\ 1443 & 672 & 4377,3 \end{vmatrix} = 2490 \Rightarrow b_2 = \frac{\Delta_{b_2}}{\Delta} = 0,0283586$$

Resp.:

b. La ecuación de regresión es: $\hat{y} = 1,29414 - 0,700469x + 0,0283586z$

c. Una persona casada de 60 años pasa frente al televisor 0.622 horas diarias.
En efecto, reemplazando en la ecuación de predicción $x = 1$ y $z = 60$ se obtiene:

$$\hat{y} = 1,29414 - 0,700469 \cdot 1 + 0,0283586 \cdot 60 = 0,622 \text{ horas.}$$

2.6.2 Ejercicios Propuestos.

1. Los individuos por Malec fueron 16 graduados de un programa de rehabilitación de lesiones cerebrales pos agudas. Los investigadores examinaron las relaciones entre diversas variables que incluían el resultado de trabajo (Y) (la escala iba de 1 para desempleado, hasta 5 que representa empleo competitivo sin prestaciones), la calificación al momento de la evaluación inicial (X) sobre el inventario de adaptabilidad Portland (PAI, por sus siglas en inglés) y el tiempo de permanencia en días (Z).

   Las mediciones registradas sobre estas tres variables son las siguientes:

y	x	z
5	67	19
4	157	17
2	242	23
4	255	14
1	227	27
4	140	22
1	179	23
4	258	18

y	x	z
4	85	16
5	52	22
3	296	15
1	256	30
4	198	21
1	224	22
4	126	19
4	156	8

a) Escriba las ecuaciones normales

b) Determinar la ecuación de regresión múltiple $\hat{y} = \hat{a} + \hat{b}x + \hat{c}z$

2. Los ornitólogos han notado que hay más tipos de pájaros anidando en terrenos arbolados que en campos de cultivo de tamaños similares. Los datos de las tablas adjunta fueron obtenidos al investigar la relación de la diversidad de los árboles (x) y la de las especies de cultivo(z) tienen en la variabilidad de las especies de pájaros

a) Escriba las ecuaciones normales

b) Determinar la ecuación de regresión múltiple

$$\hat{y} = \hat{a} + \hat{b}x + \hat{c}z$$

	Lugar	y	x	z
	A	0,639	0,043	0,972
	B	1,266	0,448	1,911
	C	2,265	0,745	2,344
	D	2,403	0,943	1,768
	E	1,721	0,731	1,372
	F	2,739	1,009	2,503
	G	1,332	0,577	1,367
	H	2,285	0,859	1,776
	I	2.277	1,021	2,464
	J	2,127	0,825	2,176
	K	2,567	1,093	2,816
Suma		2296,34	8,294	21,469

3. Los siguientes datos representan el índice de P-450IA2 (Y), número de cigarrillos diarios fumados por día (X) y el nivel de cotidiana urinaria (Z).

a. Obtener la ecuación de regresión múltiple

$$\hat{y} = \hat{a} + \hat{b}x + \hat{c}z$$

b. Obtenga y, cuando $x = 5$ y $z = 0$.

y	x	z
4,1648	1	0
3,7314	1	0
5,7481	1	0
4,437	1	0
6,4687	1	0
3,8923	3	0
5,2952	8	10,595
4,6031	8	4,6154
5,8112	8	27,1902
3,689	8	5,5319

y	x	z
3,3722	8	2,7778
8,0213	10	19,7856
10,8367	10	22,8045
4,1148	15	0
5,5429	15	14,5139
11,3531	15	36,7113
7,5637	20	21,2267
7,2158	20	21,1273
13,5	24	63,2125

4. Los siguientes datos representan el tiempo de posoperatoria en días (Y), número de problemas médicos actualmente (X) y el tiempo de permanencia preoperatoria en días (Z) de 20 pacientes intervenidos quirúrgicamente en un hospital.

a. Determinar la ecuación de regresión múltiple $\hat{y} = \hat{a} + \hat{b}x + \hat{c}z$

b. Calcule el tiempo de permanencia posoperatoria cuando el número de problemas médicos es 5 y el tiempo de permanencia preoperatoria es de, 5 días.

y	x	z
6	1	1
6	2	1
11	2	2
9	1	3
16	3	3
16	1	5
4	1	1
8	3	1
11	2	2
13	3	2

y	x	z
13	1	4
9	1	2
17	3	3
17	2	4
12	4	1
6	1	1
5	1	1
12	3	2
8	1	2
9	2	2