

## CAPITULO V

### ANALISIS DE REGRESIÓN

#### 5.1. Introducción

El objetivo de muchas investigaciones científicas es el comprender y explicar las relaciones entre variables. Frecuentemente, se requiere conocer como y en que medida una variable de respuesta se relaciona con un grupo de variables. El análisis de regresión es una técnica estadística para el modelamiento y la investigación de la relación entre dos o más variables. Por ejemplo, en un proceso químico, supóngase que el rendimiento del proceso esta relacionado con la temperatura de operación; o las siguientes cuestiones: ¿cuál es la cantidad de fertilizante aplicado, relacionado con la producción del cultivo?; ¿qué relación existe entre la cantidad de alimento consumido y el aumento de peso en los animales?; ¿cuál es el precio de una mercancía afectada por la oferta?; etc. El análisis de regresión puede usarse para construir un modelo matemático que permita predecir la relación entre las variables de interés. El modelo que se obtenga también puede usarse para la optimización del proceso, tal como hallar la temperatura que maximiza el rendimiento, o para fines de control.

En muchos problemas de esta clase, las observaciones de la “variable independiente” se hacen sin error que es insignificante al compararlo con el error (variación aleatoria) de la “variable dependiente”. Por ejemplo, al medir la cantidad de óxido en la superficie de un metal, la temperatura de calentamiento se puede controlar (variable independiente) con buena precisión; pero, la medida del espesor de óxido (variable dependiente) se encontrará sujeta a considerables variaciones aleatorias. Así, aunque la variable independiente se pueda fijar en un valor ‘ $x$ ’, las medidas repetidas de la variable dependiente nos darán valores que difieren considerablemente. Las diferencias entre los valores de ‘ $y$ ’ se pueden atribuir a diversas causas, principalmente a errores de medida y a la existencia de otras variables “incontrolables” que pueden influir en el valor de ‘ $y$ ’ cuando ‘ $x$ ’ permanece fija. Luego, las medidas del espesor de óxido variarán para diferentes piezas calentadas durante el mismo tiempo a la misma temperatura, debido a la dificultad de medir los espesores, como a las posibles diferencias en la composición de la atmósfera del horno, a las condiciones de la superficie de la pieza, etc.

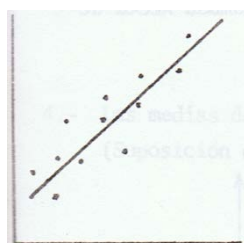
De esta discusión, debe entenderse que ‘ $y$ ’ es el valor de una variable aleatoria cuya distribución depende de ‘ $x$ ’. En la mayoría de situaciones de este tipo, nos interesa principalmente la relación entre ‘ $x$ ’ y la media de la distribución correspondiente de ‘ $y$ ’. Nos referimos a esta relación llamándola “curva de regresión”.

#### 5.2. Diagramas de dispersión.

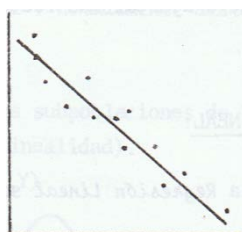
El primer paso a realizar en el estudio de la relación entre dos variables es el diagrama de dispersión que consiste en representar los pares de valores  $(X_i, Y_i)$  como puntos en un sistema de ejes cartesianos. Debido a la variación del muestreo, los puntos estarán dispersos.

Si los puntos muestran una tendencia lineal positiva o negativa, se puede ajustar una línea recta, que servirá, entre otras cosas, para predecir valores de ‘ $Y$ ’ correspondientes a valores de ‘ $X$ ’.

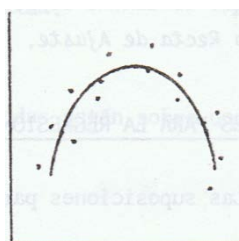
En las figuras 5.1 se representan algunos diagramas de dispersión típicos:



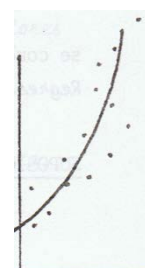
(a)



(b)



(c)



(d)

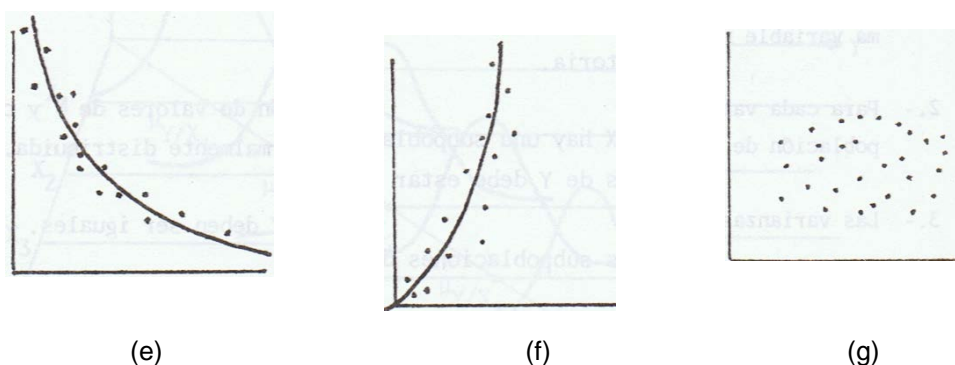


Fig. 5.1: Diagramas de dispersión que muestran (a) relación lineal positiva; (b) relación lineal negativa; (c) relación parabólica; (d) relación exponencial; (e) y (f) relaciones potenciales; y (g) ninguna relación.

### 5.3. Regresión lineal simple.

Después de que se ha determinado el modelo matemático a utilizar y se conoce que es lineal, se procede a ajustar una recta llamada *Recta de Regresión* o *Recta de Ajuste*.

La elaboración de la Recta de regresión toma en consideración los siguientes supuestos:

1. Los valores de la variable independiente 'X' son fijos; a 'X' se le llama variable no aleatoria
2. Para cada valor de 'X' hay una sub-población de valores de 'Y'; cada sub-población de valores de 'Y' debe estar normalmente distribuida.
3. Las varianzas de las sub-poblaciones de 'Y' deben ser iguales.
4. Las medias de las sub-poblaciones de 'Y' todas están sobre una recta, (suposición de linealidad).
5. Los valores de 'Y' son estadísticamente independientes; es decir, los valores de 'Y' correspondientes a un valor de 'X' no dependen de los valores de 'Y' para otro valor de 'X'

Como ilustración considérese los datos de rendimiento en carretera en kilómetros por galón de gasolina (KPG) para una muestra de 10 automóviles, según se observa en la Tabla 5.1

Tabla 5.1: Rendimiento en kilómetros por galón de diversos tipos de automóviles.

Nro.	Automóvil	Peso /(1000) Kg.	Rendimiento, KPG
1	Toyota Corona	1.179	44.7
2	Ford mustang Ghia	1.315	37.1
3	Mazda GLC	0.907	55.5
4	VW Rabbit	0.862	51.9
5	Buick Century	1.542	32.8
6	AMC Concord	1.542	29.3
7	Chevy Caprice	1.724	27.3
8	Ford Country	1.600	24.8
9	Chevette	0.998	48.8
10	AMC Sprint	1.225	44.7

La Figura 5.2 presenta el diagrama de dispersión de los datos contenidos en la Tabla 5.1. El diagrama es solo una gráfica en la que cada par  $(x, y)$  está representado por un punto en el sistema de coordenadas  $X, Y$ . El análisis de este diagrama indica que, si bien una curva no pasa exactamente por todos los puntos, existe una evidencia fuerte de que los puntos están dispersos de manera aleatoria alrededor de una línea recta.

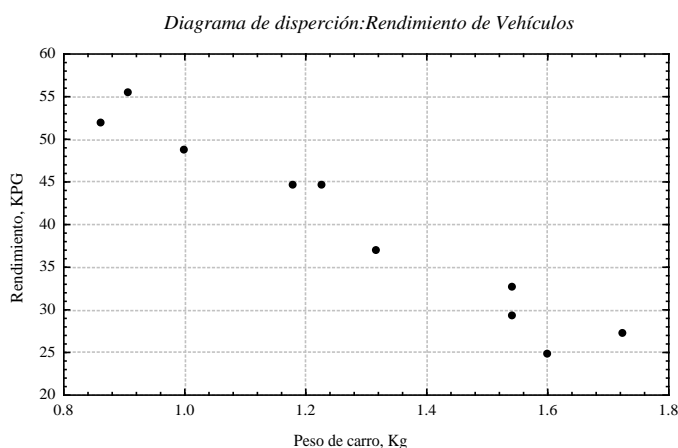


Figura 5.2: Diagrama de dispersión para el rendimiento de vehículos.

El consumo de combustible mantiene cierta relación con el esfuerzo (fuerza por distancia) requerido para desplazar el vehículo. Ya que la fuerza es proporcional al peso, es de esperarse que el consumo de combustible también sea proporcional al peso. Por consiguiente, es razonable suponer que la media de la variable aleatoria  $Y$  esta relacionada con  $X$  por la siguiente relación lineal:

$$E(Y) = \mu = \beta_0 + \beta_1 X$$

donde la pendiente,  $\beta_1$ , y la ordenada al origen,  $\beta_0$ , de la recta reciben el nombre de **coeficientes de regresión**. Si bien la media de  $Y$  es una función lineal de  $X$ , el valor real observado de  $Y$  no cae de manera exacta sobre la recta. La manera apropiada para generalizar este hecho con un **modelo probabilístico lineal** es suponer que el valor esperado de  $Y$  es una función **lineal** de  $x$ , pero que para un valor fijo de  $X$  el valor real de  $Y$  esta determinado por el valor medio de la función (el modelo lineal) más un término que representa un error aleatorio, así:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

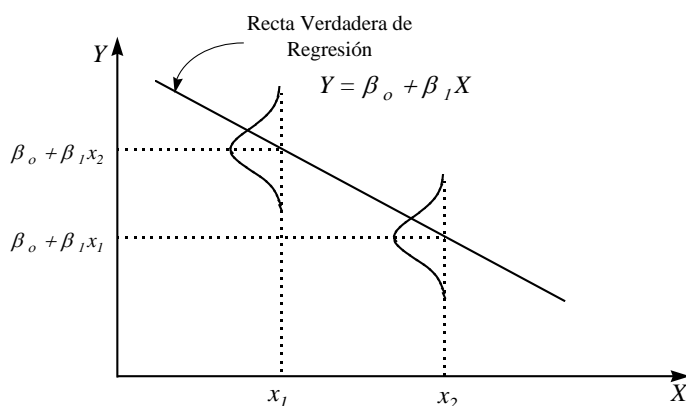
donde  $\varepsilon$  es el error aleatorio. Este modelo recibe el nombre de modelo de regresión lineal, ya que sólo tiene una variable independiente o regresor ( $X$ ).

Naturalmente, una relación precisa entre peso y consumo de combustible debe tener en consideración otros factores tales como eficiencia de uso de combustible, diseño de motor, forma del chasis, etc., factores que al no ser considerados, contribuyen a hacer más dispersos los datos. A pesar de ello supóngase que el modelo de regresión verdadero para el caso del rendimiento de los automóviles se ajusta a la línea recta mostrada en la Figura 6.2. El modelo de regresión verdadero:

$$E(Y) = \mu = \beta_0 + \beta_1 X$$

es una recta de valores promedio; esto es, cualquier punto de la recta corresponde al *valor esperado* de  $Y$  para su correspondiente  $X$ . La pendiente  $\beta_1$  puede interpretarse como el cambio de la media de  $Y$  por unidad de cambio de  $X$ . Además, la variabilidad de  $Y$  en un valor particular de  $X$  está determinada por la varianza del error. Esto implica que existe una distribución de valores de  $Y$  para cada  $X$ , y que la varianza de esta distribución es la misma para cada  $X$ . Nótese que se ha utilizado una distribución normal para describir la variación aleatoria de  $\varepsilon$ . La variación de  $\varepsilon$  mide que tan dispersos se encuentran los valores medidos con respecto a los calculados, de modo que, según el valor de  $\sigma^2$  de  $\varepsilon$  determina el grado de dispersión de las observaciones. Por lo tanto, cuando  $\sigma^2$  tiene un valor pequeño, los valores observados de  $Y$  caen cerca de la línea, y cuando  $\sigma^2$  es grande, los valores observados de  $Y$  pueden desviarse considerablemente de la línea. Dado que  $\sigma^2$  es constante, la variabilidad en  $Y$  para cualquier valor de  $X$  es la misma.

La ecuación de regresión proporciona una “estimación” de la ecuación de la línea de regresión cuya expresión verdadera es ‘desconocida’ (analogía con el promedio ‘ $\mu$ ’). El error real al predecir  $Y$  es  $\varepsilon$ , y este error se estima por la cantidad:  $Y_i - (\beta_0 + \beta_1 X_i)$ . Se trata de determinar  $\beta_0$  y  $\beta_1$  de tal forma que los errores estimados sean tan pequeños como sea posible. Se trata entonces de que determinar  $\beta_0$  y  $\beta_1$  tal que  $\sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_i)]^2$  sea mínima. Esto equivale a hacer mínima la suma de los cuadrados de las distancias entre cada punto observado experimentalmente y el valor correspondiente a la ecuación de regresión, (criterio de mínimos cuadrados).

Figura 5.3: Distribución de Y para valores de  $X_1$  y  $X_2$ 

El caso de regresión lineal simple considera solo un *regresor* o *predictor*  $X$ , y una variable dependiente o respuesta  $Y$ . Existen tres parámetros en este modelo: los coeficientes  $\beta_0$ ,  $\beta_1$  y la varianza  $\sigma^2$  la cual estima la dispersión de los datos alrededor de la línea de regresión. Normalmente, estos parámetros son desconocidos y deben ser calculados de los datos de muestra. Las estimaciones de  $\beta_1$  y  $\beta_2$  deben dar como resultado una línea que, en algún sentido, se 'ajuste mejor' a los datos. El científico alemán Karl Gauss (1777-1855) propuso estimar esos parámetros de modo que se minimice la suma de los cuadrados de las desviaciones en los valores de  $Y$ . Este criterio para estimar los coeficientes de regresión se conoce como **método de los mínimos cuadrados**.

El método de mínimos cuadrados permite calcular los estimadores  $\hat{\beta}_0$  y  $\hat{\beta}_1$  de  $\beta_0$  y  $\beta_1$  mediante las siguientes expresiones:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i x_i - \frac{\left(\sum_{i=1}^n y_i\right)\left(\sum_{i=1}^n x_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

donde

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Una expresión equivalente para  $\hat{\beta}_1$  está dada por :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i \bar{y}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Por tanto, la línea de regresión estimada o ajustada es:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Observese que cada par de observaciones satisface la relación :

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i \quad i=1,2,\dots,n$$

donde  $e_i = y_i - \hat{y}$  recibe el nombre de residuo.

El residuo describe el error en el ajuste del modelo en la  $i$ -ésima observación  $y_i$ . Más adelante se utilizarán los residuos para proporcionar información sobre la adecuación del modelo ajustado.

**Ejemplo**

Para los datos de la Tabla 5.1 de rendimiento de combustible de vehículos, tenemos:

$$\sum_{i=1}^{10} x_i = 12,894 \quad \sum_{i=1}^{10} x_i^2 = 17.469 * 10^6 \quad \sum_{i=1}^{10} y_i = 396.9$$

$$\sum_{i=1}^{10} y_i^2 = 16,822.55 \quad \sum_{i=1}^{10} x_i y_i = 482,527.4$$

$$\hat{\beta}_1 = \frac{482,527.4 - \frac{396.9 * 12,894}{10}}{17.469 * 10^6 - \frac{(12,894)^2}{10}} = -0.03465$$

$$\hat{\beta}_0 = 39.69 + 34.66 * 1.289 = 84.366$$

Por lo tanto el modelo queda definido como:

$$Y = 84.366 - 0.03465 X$$

Un programa estadístico de computadora proporciona el modelo y gráfico de regresión, según se muestra en la Figura 5.4.

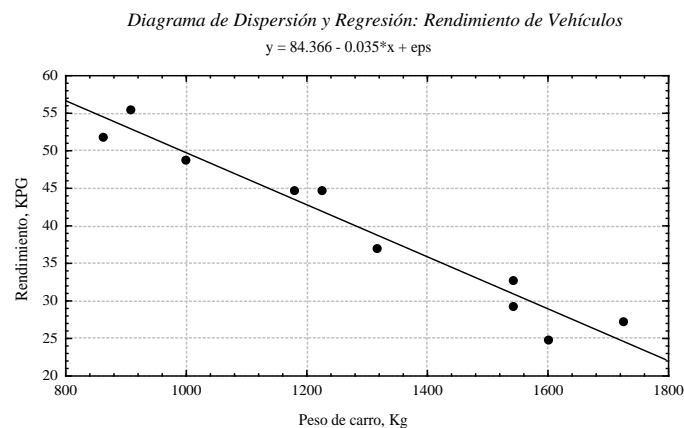


Figura 5.4. Modelo de regresión y gráfico para rendimiento de vehículos.

La pendiente estimada por  $\hat{\beta}_1 = -0.035$  KPG/Kg. significa que el aumento de 1 Kg. en el peso del vehículo este avanza 0.035 Km. (35 m.) menos por cada galón consumido; ó también, que una persona dentro del vehículo que pese 70 Kg. ocasionará una reducción en el rendimiento del vehículo de 2.45 Km. Esto último dicho de otra forma implica que, para un vehículo que rinde 40 KPG, el transporte de una persona que pesa 70 Kg. implicará un consumo de  $2.45/40 = 0.06$  galones por cada 40 Km. de recorrido.

Nótese que los datos incluyen valores de peso entre 862 y 1,724 Kg. por lo que para valores fuera de este rango, la eficiencia de vehículo puede ser diferente. La intersección  $\hat{\beta}_0 = 84.37$  no se le puede asociar ningún significado, pues implicaría el rendimiento de un vehículo de peso 0 Kg.

**5.3.1 Valores residuales y ajustados**

Los estimados por mínimos cuadrados de  $\hat{\beta}_0$  y  $\hat{\beta}_1$  permiten estimar (ajustar) la línea de regresión dada por:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

El cálculo de ésta expresión para los niveles  $x_1, \dots, x_n$  proporciona los **valores ajustados**:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \quad \text{para } i = 1, 2, \dots, n$$

Las respectivas diferencias entre las observaciones  $Y_1, Y_2, \dots, Y_n$  y los valores ajustados  $\hat{Y}_1, \hat{Y}_2, \dots, \hat{Y}_n$  son denominados *residuales*, y están dados por:

$$\begin{aligned} e_i &= Y_i - \hat{Y} \\ &= Y_i - \left( \hat{\beta}_0 - \hat{\beta}_1 X_i \right) \quad i=1, 2, \dots, n. \end{aligned}$$

La estimación de residuos para el ejemplo de consumo de combustible por automóviles se muestra en el cuadro 5.1, tal como se obtuvo de un programa estadístico de computadora.

Obsérvese que la suma de residuales es cero, lo cual es general para toda evaluación de residuos, es decir, para un análisis de regresión:

$$\sum_{i=1}^n e_i = 0$$

Cuadro 5.1: Análisis de residuos para el modelo de regresión de consumo de combustible.

STATISTICA: Multiple Regression			
CASE	OBSERVED VALUE	PREDICTD VALUE	RESIDUAL
1	44.70	43.51521	1.18479
2	37.10	38.80299	-1.70300
3	55.50	52.93966	2.56035
4	51.90	54.49884	-2.59884
5	32.80	30.93774	1.86226
6	29.30	30.93774	-1.63775
7	27.30	24.63169	2.66831
8	24.80	28.92812	-4.12812
9	48.80	49.78662	-.98663
10	44.70	41.92138	2.77863
Minimum	24.80	24.63169	-4.12812
Maximum	55.50	54.49884	2.77863
Mean	39.69	39.69000	-.00000
Median	40.90	40.36218	.09908

Un modelo lineal en términos generales esta definido por:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \varepsilon$$

en donde el término  $\varepsilon$  es una variable aleatoria con media cero y varianza  $\sigma^2$ . Puesto que los valores de  $X$  son fijos,  $Y$  es una variable aleatoria de promedio  $\mu = \beta_0 + \beta_1 X$  y varianza  $\sigma^2$ . Por consiguiente, los valores  $\hat{\beta}_0$  y  $\hat{\beta}_1$  dependen de los valores de  $Y$  observados. Por consiguiente, los estimadores de mínimos cuadrados de los coeficientes de regresión pueden verse como variables aleatorias.

Para obtener inferencias con respecto a los coeficientes de regresión  $\beta_0$  y  $\beta_1$  es necesario estimar la varianza, la que corresponde a la varianza del término de error  $\varepsilon$  en el modelo de regresión. Esta varianza refleja la variación alrededor de la verdadera recta de regresión.

Los residuos  $e$  se emplean en el cálculo de  $\sigma^2$ . La suma de los cuadrados de los residuos, o **Suma de los Cuadrados del Error,  $SS_E$  (Sum Square Error)** es:

$$SS_E = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n \left( Y_i - \hat{Y} \right)^2$$

por lo tanto:

$$\hat{\sigma}^2 = \frac{SS_E}{n-2}$$

### 5.3.2 Inferencias en el modelo de regresión

El modelo de regresión pretende usar información de una variable independiente para explicar la variabilidad en la respuesta  $Y$ . Ignorando la información contenida en  $x_1, x_2, \dots, x_n$ , se puede medir la variabilidad entre las respuestas  $Y_1, Y_2, \dots, Y_n$  utilizando:

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - \frac{\left(\sum_{i=1}^n Y_i\right)^2}{n} = SS_{TO}$$

que se le denomina suma total de cuadrados,  $SS_{TO}$ , (*total sum of squares*).

Las variaciones de los valores ajustados según la ecuación de regresión con respecto al promedio de los valores de  $Y$  se mide según:

$$\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = SS_R$$

y se denomina suma de cuadrados debido a la regresión,  $SS_R$ , (*regression sum of squares*).

Generalmente, no toda la variación es explicada por el modelo de regresión. Los residuos,  $e_i$ , expresan un componente adicional, que es medida por:

$$SS_E = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

que se denomina suma de cuadrados debida al error,  $SS_E$ , (*error sum squares*). De todo esto se puede demostrar que:

$$SS_{TO} = SS_R + SS_E$$

Todos estos conceptos conducen a la definición del **Coefficiente de Determinación**,  $R^2$  que es una medida estadística de el ajuste de los valores medidos a la ecuación de regresión. Se representa por:

$$R^2 = \frac{SS_R}{SS_{TO}} = 1 - \frac{SS_E}{SS_{TO}} = \frac{\left[ \sum_{i=1}^n X_i Y_i - \left( \sum_{i=1}^n X_i \right) \left( \frac{\sum_{i=1}^n Y_i}{n} \right) \right]^2}{\left[ \sum_{i=1}^n X_i^2 - \frac{\left( \sum_{i=1}^n X_i \right)^2}{n} \right] \left[ \sum_{i=1}^n Y_i^2 - \frac{\left( \sum_{i=1}^n Y_i \right)^2}{n} \right]}$$

Para ilustrar el significado de esas expresiones, supóngase que  $\hat{\beta}_1 = 0$  lo cual indica que no existe relación entre  $X$  y  $Y$ ; es decir  $Y = \beta_0$ ; ó  $Y_1 = Y_2 = \dots = Y_n$ . Por lo tanto:  $\hat{Y} = \bar{Y}$  y  $SS_R = 0$  y  $SS_{TO} = SS_E$  y  $R^2 = 0$ . En una segunda situación, supóngase que todos los puntos de la ecuación ajustada pasan por los punto medidos; o sea  $\hat{Y}_i = Y_i$  entonces  $SS_E = 0$  y  $SS_{TO} = SS_R$ , y  $R^2 = 1$  lo cual indica que existe un excelente ajuste entre lo real y calculado. De esto último y en la ecuación:

$$R^2 = \frac{SS_R}{SS_{TO}} = 1 - \frac{SS_E}{SS_{TO}}$$

se puede observar que  $R^2 = 1$  cuando hay un excelente ajuste entre lo medido y calculado; por el contrario,  $R^2 = 0$  cuando no existe ninguna relación entre  $X$  y  $Y$ . De esto sigue que:

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

Esto explica que mientras el valor de  $R^2$  sea más cercano a 1 mejor será el ajuste predicho por la ecuación de regresión.

En el ejemplo de el rendimiento de vehículos, se calcula fácilmente  $R^2 = 0.94706$ , lo cual indica que existe un buen ajuste predicho por la ecuación de regresión.

La salida que proporciona un programa de computadora se muestra en Cuadro 5.2:

Cuadro 5.2: Resultados de Regresión de un programa de computadora

```

STATISTICA: Multiple Regression
data file: CARROS1.STA [ 10 cases with 3 variables ]
VARIABLES:
  2: PESO_KG   -
  3: RENDIMIE
Missing data casewise deleted
MULTIPLE REGRESSION RESULTS:
Variables were entered in one block

Dependent Variable: RENDIMIE
Multiple R:          .973172525
Multiple R-Square:   .947064764
Adjusted R-Square:   .940447859
Number of cases:     10
F ( 1,      8) = 143.1281      p < .000002
Standard Error of Estimate: 2.660332365
Intercept:  84.366006027 Std.Error: 3.827912 t( 8) = 22.040 p < .000000

```

### 5.3.3 Prueba de hipótesis en la regresión lineal simple

Una parte importante al evaluar la adecuación de un modelo de regresión lineal es la prueba de hipótesis estadísticas sobre los parámetros del modelo y la construcción de ciertos intervalos de confianza. Para probar hipótesis sobre la pendiente y la ordenada al origen del modelo de regresión, debe hacerse la hipótesis adicional de que el componente del error del modelo  $\varepsilon$ , tiene una distribución normal, es decir  $\varepsilon$  es  $N(0, \sigma)$ .

#### 5.3.3.1. Uso de pruebas $t$

Si se desea probar la hipótesis de que la pendiente es igual a una constante,  $\beta_{10}$  las hipótesis apropiadas son:

$$H_0 : \beta_1 = \beta_{10}$$

$$H_1 : \beta_1 \neq \beta_{10}$$

donde se ha considerado una hipótesis alternativa a dos colas (bilateral). Para evaluación de estas hipótesis se calcula el estadístico de prueba siguiente:

$$T_0 = \frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} = \frac{\hat{\beta}_1 - \beta_{10}}{se(\hat{\beta}_1)}$$

donde

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} = se(\hat{\beta}_1) = \text{error estándar estimado de la pendiente}$$

$T_0$  tiene una distribución  $t$  con  $n-2$  grados de libertad. Puede rechazarse  $H_0$  si:

$$|T_0| > t(\alpha/2, n-2)$$

De modo similar para  $\beta_0$  se pueden probar las hipótesis:

$$H_0 : \beta_0 = \beta_{00}$$

$$H_1 : \beta_0 \neq \beta_{00}$$

donde se ha considerado una hipótesis alternativa a dos colas (bilateral). Para evaluación de estas hipótesis



se calcula el estadístico de prueba siguiente:

$$T_0 = \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{\hat{\sigma}^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]}} = \frac{\hat{\beta}_0 - \beta_{00}}{se(\hat{\beta}_0)}$$

$$\text{donde: } \sqrt{\hat{\sigma}^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]} = se(\hat{\beta}_0) = \text{error estándar de la ordenada al origen}$$

$T_0$  tiene una distribución  $t$  con  $n-2$  grados de libertad. Puede rechazarse  $H_0$  si:

$$|T_0| > t(\alpha/2, n-2)$$

Un caso especial muy importante de las hipótesis anteriores es:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

las cuales están relacionadas con la significancia de la regresión. Aquí, si se rechaza  $H_1$  y acepta  $H_0$  es equivalente a concluir que no hay ninguna relación lineal entre  $X$  e  $Y$ .

Ejemplo: Luego de una evaluación y cálculo de un grupo de 20 datos se obtuvo el siguiente modelo lineal;

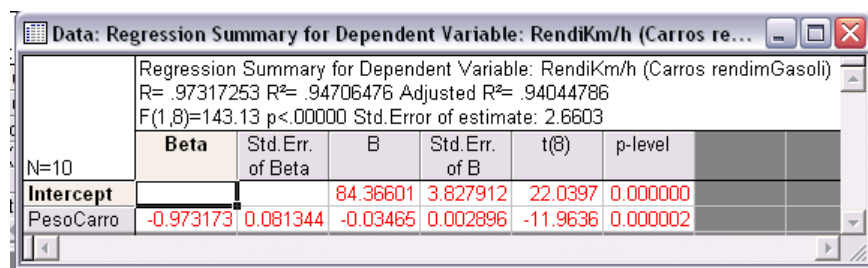
$$Y = 74.20 + 14.97 X$$

Se calculo  $S_{xx} = 0.68$  y  $\sigma^2 = 1.17$ . De esto:

$$T_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2 / S_{xx}}} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{14.97}{\sqrt{1.17/0.68}} = 11.41$$

de Tablas se obtiene  $t(0.05, 18) = 2.88$ . Con estos resultados se decide rechazar  $H_0$ . El valor  $p$  de esta prueba es  $1.13 \times 10^{-9}$ .

Para el caso de los datos de la Tabla 5.1 (rendimiento de vehículos como función del peso), una salida típica de un programa de computadora es la siguiente:



Data: Regression Summary for Dependent Variable: Rendikm/h (Carros re...)						
Regression Summary for Dependent Variable: Rendikm/h (Carros rendimGasoli)						
R= .97317253 R²= .94706476 Adjusted R²= .94044786						
F(1,8)=143.13 p<.00000 Std.Error of estimate: 2.6603						
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(8)	p-level
N=10						
Intercept			84.36601	3.827912	22.0397	0.000000
PesoCarro	-0.973173	0.081344	-0.03465	0.002896	-11.9636	0.000002

Como se observa, los valores ' $t$ ' para  $\beta_0$  y  $\beta_1$  son 22.0397 y -11.9636 respectivamente. El correspondiente valor ' $p$ ' indica la significancia de esos regresores en el modelo matemático establecido.

### 5.3.3.2. Análisis de la varianza y prueba $F$

Para probar la significancia de una regresión se utiliza el método de análisis de la varianza, ANAVA. Como base para la prueba, el procedimiento particiona la variabilidad total en componentes más manejables. La identidad del análisis de la varianza es el siguiente:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Los dos componentes del lado derecho de esta ecuación miden respectivamente: a) la cantidad de variabilidad en  $y_i$  tomada en cuenta por la recta de regresión; y, b) la variación residual que queda sin explicar por la recta.

Lo usual es utilizar las siguientes definiciones:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = S_{yy} \text{ (Suma Total de Cuadrados Corregida). También simbolizado por } SS_{TO}$$

$$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = SS_R \text{ (Suma de Cuadrados de la Regresión)}$$

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = SS_E \text{ (Suma de Cuadrados de los Errores)}$$

De modo que la última ecuación puede escribirse:

$$S_{yy} = SS_R + SS_E$$

La suma total de cuadrados  $S_{yy}$  tiene  $(n-1)$  grados de libertad, y  $SS_R$  y  $SS_E$  tiene  $(1)$  y  $(n-2)$  grados de libertad respectivamente..

Es posible demostrar que  $SS_E/\sigma^2$  y  $SS_R/\sigma^2$  son variables aleatorias independientes con distribución *Chi cuadrado* con  $(n-2)$  grados de libertad. Para el contraste de las hipótesis:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

se utiliza el estadístico  $F$  expresado de la siguiente forma:

$$F = \frac{(SS_R / \sigma^2) / 1}{(SS_E / \sigma^2) / (n-2)} = \frac{SS_R / 1}{SS_E / (n-2)} = \frac{MS_R}{MS_E}$$

$$\text{observese que } MS_E = \hat{\sigma}^2$$

Este estadístico tiene distribución  $F(\alpha; 1, n-2)$  con lo que  $H_0$  debe rechazarse si  $F_0 > F(\alpha; 1, n-2)$ .

Las cantidades  $MS_R$  y  $MS_E$  reciben el nombre de **medias de cuadrados**. En general, una media de cuadrados siempre se calcula dividiendo una suma de cuadrados entre su número de grados de libertad. Lo usual es acomodar el procedimiento de prueba en una Tabla de Análisis de la Varianza, según se observa en la Tabla 5.2.

Tabla 5.2: Tabla de ANAVA para un modelo de regresión linear simple.

Fuente de variación	Suma de cuadrados, SS	Grados de libertad, gl	Media de cuadrados, MS	F	Valor p
Regresión	$SS_R$	1	$MS_R = SS_R/1$	$MS_R / MS_E$	Area de probabilidad
Error	$SS_E$	$n - 2$	$MS_E = SS_E/(n-2)$		
Total	$SS_{TO}$	$n - 1$			

Los grandes valores de  $F$  indican que  $\beta_1$  es diferente de cero. Esto implica de que si la relación  $F$  excede el  $100\alpha$  punto de porcentaje de la distribución  $F(1, n-2)$ , o de otra forma, si  $F > F(\alpha; 1, n-2)$  se rechaza  $H_0$  y acepta  $H_1$  a un nivel de significación  $\alpha$ , lo que significa que  $\beta_1$  es diferente de cero. El valor  $p$  es un valor de área o probabilidad asociado al valor de  $F_0$ . Si el valor  $p$  es menor que el nivel de significación fijado,  $\alpha$ , se acepta  $H_1$ .

Los resultados de análisis de la varianza obtenidos de un programa estadístico computadora para los datos de la Tabla 5.1 son los siguientes:

Effect	Sums of Squares	df	Mean Squares	F	p-level
Regress.	1012.970	1	1012.970	143.1281	0.000002
Residual	56.619	8	7.077		
Total	1069.589				

El valor pequeño de ' $p\text{-level}$ ' indica una alta dependencia de ' $Y$ ' con respecto a ' $X$ '.

#### 5.3.4 Intervalos de confianza

Además de las estimaciones puntuales de la pendiente y la ordenada al origen, es posible obtener

estimaciones de los intervalos de confianza para estos parámetros. El ancho de estos intervalos es una medida de la calidad total de la recta de regresión. Si los términos de error,  $\varepsilon_i$  en el modelo de regresión están distribuidos de manera normal e independiente, entonces:

$$\frac{\hat{\beta}_1 - \beta_{10}}{\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}} \quad y \quad \frac{\hat{\beta}_0 - \beta_{00}}{\sqrt{\hat{\sigma}^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]}}$$

son variables aleatorias con distribución  $t$  y  $n-2$  grados de libertad. El intervalo de confianza para la pendiente  $\beta_1$  del  $(100-\alpha)$  por ciento en regresión lineal simple es:

$$\hat{\beta}_1 - t_{\alpha/2, n-2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2, n-2} \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

De manera similar, el intervalo de confianza para la ordenada al origen  $\beta_0$  es:

$$\hat{\beta}_0 - t_{\alpha/2, n-2} \sqrt{\hat{\sigma}^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]} \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2, n-2} \sqrt{\hat{\sigma}^2 \left[ \frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]}$$

Los intervalos de confianza graficados para los valores de  $\beta_1$  para los datos de la Tabla 5.1, se muestran en la Figura 5.5, representando el intervalo de confianza al 95% para el modelo de ajuste.

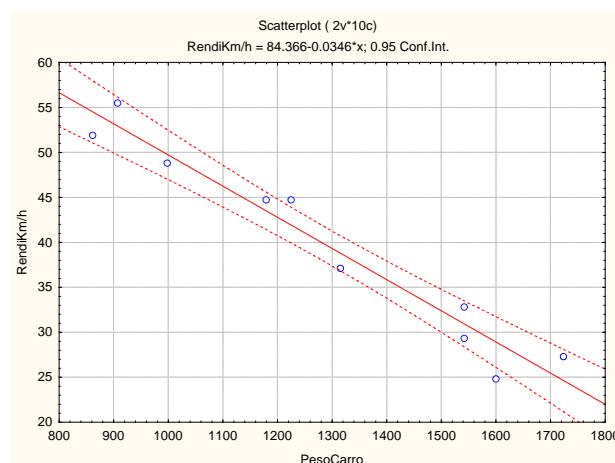


Figura 5.5: Curva de regresión e intervalos de confianza para el rendimiento de vehículos

### 5.3.5 Evaluación de la adecuación del modelo

Un modelo de regresión lineal simple puede que ajuste datos que presentan una escasa relación linear. Un ajuste lineal no es apropiado para datos que siguen un modelo cuadrático u otro más complicado. Tales modelos puede no ser consistentes con los supuestos asociados a los modelos de regresión lineal simple. Además, este modelo de regresión debe ser ajustarse a datos para los cuales la variabilidad en los valores respuesta de  $Y$  son aproximadamente constantes para todos los valores de  $X$ . Por ejemplo, el modelo no sería adecuado para datos en los que la variabilidad en  $Y$  se incrementa al incrementarse los valores de  $X$ . Esto estaría en contradicción con el principio de constancia de la varianza.

#### 5.3.5.1. Análisis residual.

Es muy importante revisar los supuestos para un modelo de regresión. Las principales herramientas útiles para verificar el modelo son la evaluación de residuos, tales como:

- Gráficos de residuos versus los valores ajustados de  $Y$ .
- Gráficos de residuos versus los valores de  $X$ .
- Gráficos de residuos versus otras variables que no fueron incluidas en el modelo original, e. g. orden cronológico, si los datos son obtenidos secuencialmente.

Si los supuestos de regresión satisfacen el modelo, no debe observarse ningún 'patrón' en esos gráficos. Los residuos deben aparecer variando aleatoriamente en un área horizontal  $2\sigma$  alrededor de la línea de cero. Por ejemplo, los gráficos de la Figura 5.6 muestran patrones que no satisfacen el requisito de

constancia de la varianza. En el primero de ellos por ejemplo, se observan desviaciones negativas en los valores extremos y desviaciones positivas en los valores medios; en el segundo gráfico, las desviaciones se ejincrementan para valores altos. De ellos se deduce que el modelo obtenido en cada caso no se ajusta plenamente a los datos obtenidos.

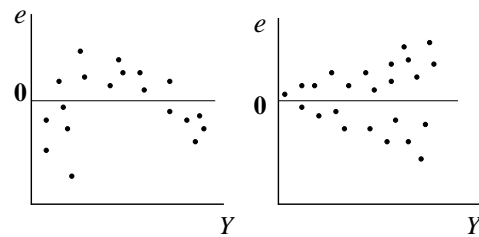


Figura 5.6: Gráficos de residuos que muestran determinados patrones.

Con los datos de la Tabla 5.1 y con la aplicación de un programa estadístico de computadora se obtienen los gráficos de las Figuras 5.7 y 5.8.

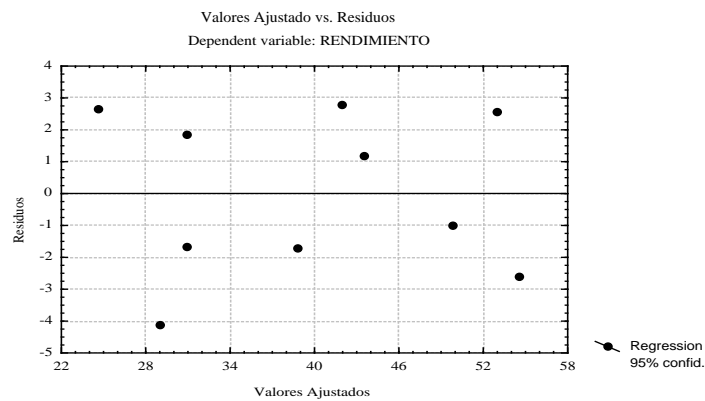


Figura 5.7: Relación de valores calculados y residuos de ajuste.

La Figura 5.6 muestra el gráfico de los valores calculados vs residuos, esto es  $\hat{Y}_i$  vs  $(Y_i - \hat{Y}_i)$ . Se observa en el que los 10 datos presentan residuos distribuidos sin patrones evidentes. Semejante conclusión se obtiene de la Figura 5.7. Obviamente, los datos son muy pocos para obtener una conclusión definitiva acerca de la validez del modelo en función de la evaluación de los residuos, pero queda con ello definido el procedimiento de análisis para cuando se disponga de mayor información.

De encontrarse patrones en estos gráficos de residuos estaría indicando la presencia de un componente no explicado por el modelo. Por ejemplo, si al obtener mayor información de rendimiento de vehículos se observasen algunos patrones, se evaluaría la posibilidad de considerar otras variables, tales como numero de cilindros, desplazamiento de los pistones, potencia, etc.

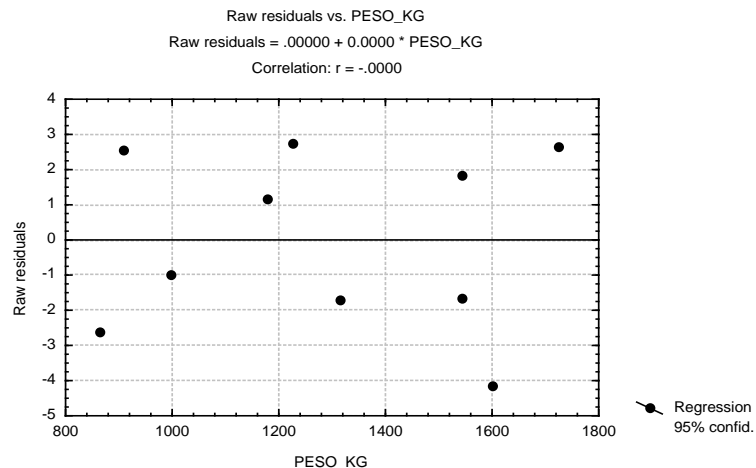


Figura 5.8: Relación de variable independiente y residuos de ajuste.

### 5.3.5.2. Coeficiente de determinación ( $R^2$ )

Como se vio anteriormente, la cantidad:

$$R^2 = \frac{SS_R}{SS_{TO}} = 1 - \frac{SS_E}{SS_{TO}}$$

recibe el nombre de coeficiente de determinación y se utiliza con mucha frecuencia para juzgar la adecuación de un modelo de regresión. A menudo se hace frecuencia de manera vaga a  $R^2$  como la cantidad de variabilidad en los datos que es explicada o tomada en cuenta por el modelo de regresión. Para el ejemplo de rendimiento de combustible de vehículos, se determinó el valor de  $R^2 = 0.9471$ ; lo que significa, que el modelo determinado toma en cuenta el 94.71% de la variabilidad presente en los datos.

El estadístico  $R^2$  debe emplearse con precaución, ya que siempre es posible hacer  $R^2$  igual a uno mediante la adición al modelo de un número suficiente de términos. Por ejemplo, puede obtenerse un ajuste “perfecto” a  $n$  puntos con un polinomio de grado  $(n-1)$ . Además  $R^2$  siempre aumenta si se añade una variable al modelo, lo que no implica necesariamente que el nuevo modelo sea mejor que el anterior. Aménos que la suma de los cuadrados de los errores del nuevo modelo se vea disminuida por una cantidad igual que al error cuadrático medio original, el nuevo modelo tendrá un error cuadrático medio mayor que el anterior debido a la pérdida de un grado de libertad en el error; por lo tanto, en realidad el nuevo modelo es peor que el anterior.

Se debe tener en cuenta que  $R^2$  no mide cuan apropiado resulta ser el modelo, ya que esto puede inflarse de manera artificial con la adición al modelo de términos polinomiales en  $X$  de grado superior. Incluso,  $R^2$  puede ser grande si  $X$  y  $Y$  están relacionadas de manera NO lineal. Finalmente, a pesar de  $R^2$  sea grande, esto no necesariamente implica que el modelo de regresión proporcionará predicciones precisas de observaciones futuras.

### 5.3.6. Transformaciones que llevan a una línea recta

Al reconocer la existencia de falta de ajuste en un modelo de regresión se deben tomar acciones para corregir el defecto. Naturalmente, un diagrama de dispersión es el primer análisis a realizar, el cual puede en muchos casos revelar la falta de ajuste lineal. El coeficiente de determinación, el análisis de residuos son pruebas definitivas de la falta de ajuste. En el caso de que la varianza no es constante, se debe intentar alguna transformación. Algunos ejemplos gráficos de estas transformaciones se dan en la Figura 5.9.

En algunas circunstancias, la función no lineal puede expresarse como una línea recta mediante el empleo de una transformación adecuada. Estos modelos no lineales son denominados *intrínsecamente lineales*. Como ejemplo de un modelo no lineal que es intrínsecamente lineal, considérese la siguiente función exponencial:

$$Y = \beta_0 e^{\beta_1 X} \varepsilon$$

Esta función es intrínsecamente lineal, puesto que puede transformarse en una línea recta mediante una transformación logarítmica:

$$\ln Y = \ln \beta_0 + \beta_1 X + \ln \varepsilon$$

Esta transformación requiere que los términos de error transformados  $\ln \varepsilon$  sean normales, con media cero y varianza  $\sigma^2$  y que estén distribuidos de manera independiente.

Otra función intrínsecamente lineal es:

$$Y = \beta_0 + \beta_1 \left( \frac{1}{X} \right) + \varepsilon$$

Mediante el empleo de la transformación recíproca  $Z = 1/X$ , el modelo queda linealizado como:

$$Y = \beta_0 + \beta_1 Z + \varepsilon$$

Si la varianza de  $Y$  se incrementa proporcionalmente al nivel de  $Y$ , se puede tratar transformaciones como:  $\sqrt{Y}$ ,  $1/Y$ ,  $\log Y$ , etc u otras potencias de  $Y$ . Los gráficos de estas transformaciones versus los valores de  $X$  indicarán que transformación en particular estabiliza mejor la varianza.

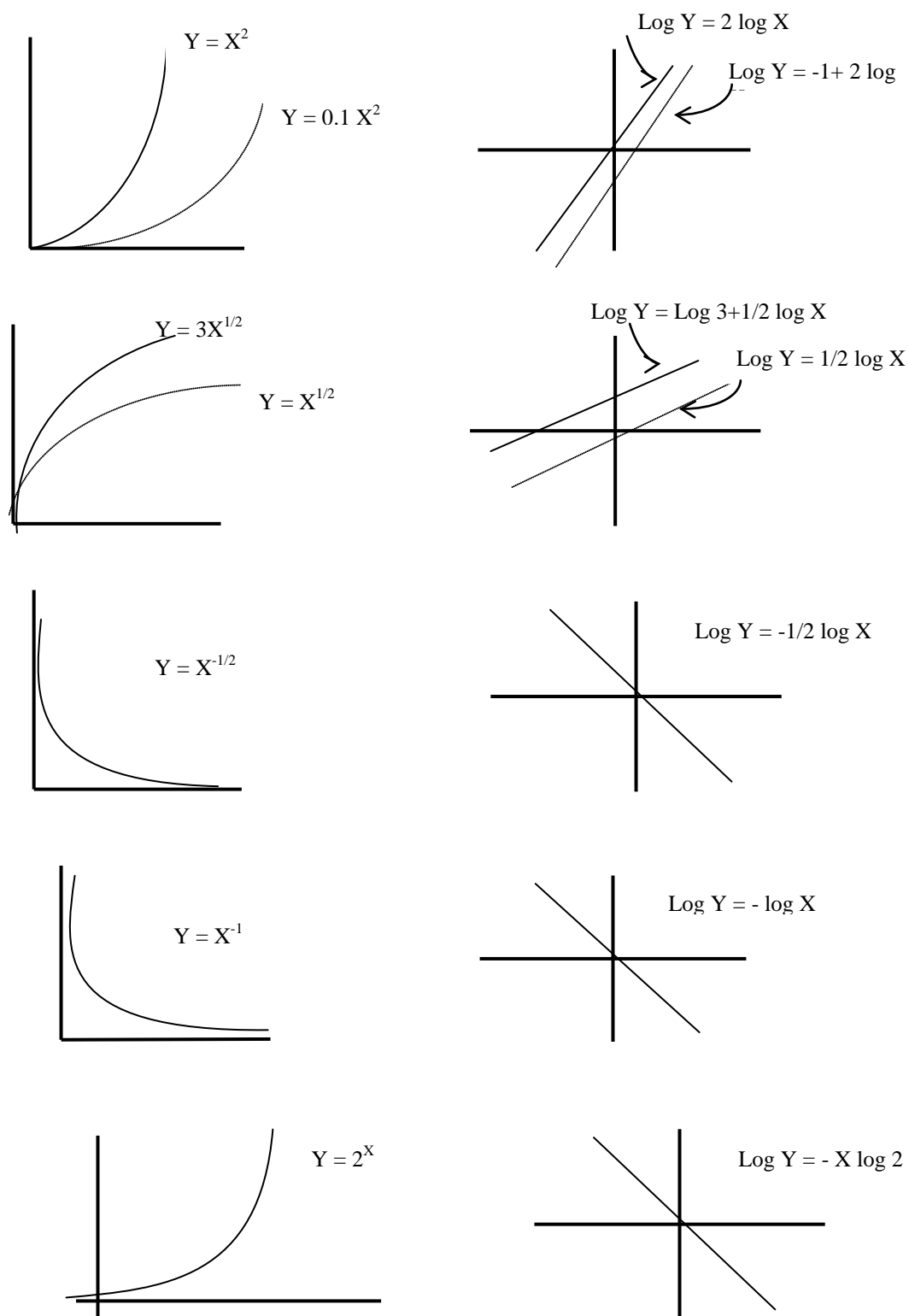


Figura 5.9: Diversas curvas formadas con sus transformaciones logarítmicas.

#### 5.4. Regresión no lineal (exponencial y potencial)

Cuando se sospecha que la relación es de tipo exponencial, se propone una ecuación de regresión de la forma:

$$Y = c d^X \quad (\text{Como sugiere el nombre exponencial, la variable independiente 'X' aparece en el exponente})$$

Por necesidad teórica y conveniencia práctica, se "*transforma*" la ecuación a otra, tomando logaritmos en ambos lados

$$\text{Log}(Y) = \text{Log}(c) + X \text{Log}(d)$$

Haciendo:  $\text{Log}(c) = a$  y  $\text{Log}(d) = b$

La ecuación exponencial es transformada en:

$$\text{Log}(Y) = a + bX$$

que es lineal en  $\text{Log}(Y)$  y  $X$ ; la cual, es una función semilogarítmica, de manera que si se lleva los puntos a papel semilogarítmico, se obtiene una recta.

Para transformar de nuevo la ecuación semilogarítmica a la forma original, solo se necesita tomar la función inversa del logaritmo; esto es, la función exponencial de la base adecuada según haya sido la base de los logaritmos con que se este trabajando. En el caso presentado:

$$c = 10^a$$

$$d = 10^b$$

$$Y = 10^a * 10^{bX}$$

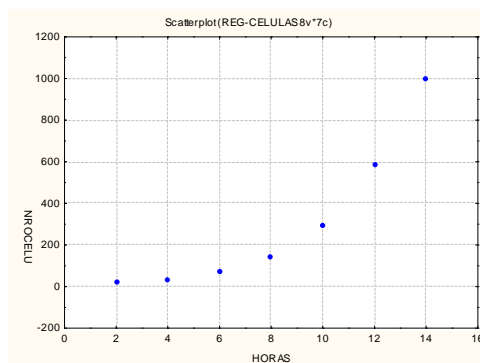
### **Ejemplo:**

Los datos de la siguiente tabla se obtuvieron de observaciones periódicas hechas durante el crecimiento de una población de células. Se efectuaron recuentos cada dos horas.

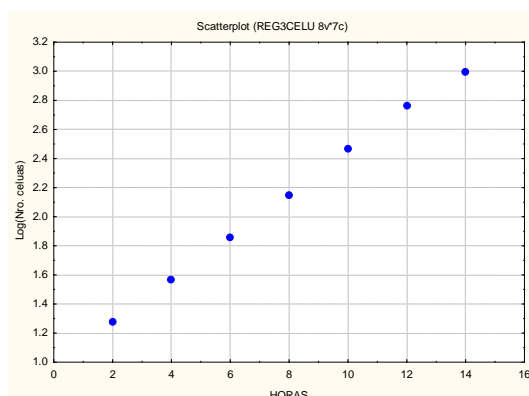
Horas	Nro. de células	Log(Nro. de células)
2	19	1.279
4	37	1.568
6	72	1.857
8	142	2.152
10	295	2.470
12	584	2.766
14	995	2.998

### **Solución:**

El diagrama de dispersión para los datos del número de células se da en la siguiente Figura. Obsérvese la relación exponencial que se establece entre el tiempo y el número de células.



Para los mismos dato, el gráfico de Log(Nro. de células) versus tiempo se observa en la siguiente Figura



Debido a que la relación entre  $X$  y  $\text{Log}(Y)$  es lineal, se pueden aplicar los principios de regresión lineal. Los valores del modelo matemático correspondiente, se observan en el siguiente cuadro:

Regression Summary for Dependent Variable: LOGNROCE (REG3CE)						
R= .99949393 R²= .99898812 Adjusted R²= .99878574						
F(1,5)=4936.3 p<.00000 Std.Error of estimate: .02197						
N=7	Beta	Std.Err. of Beta	B	Std.Err. of B	t(5)	p-level
Intercept			0.989216	0.018564	53.28674	0.000000
HORAS	0.999494	0.014226	0.145824	0.002076	70.25879	0.000000

Así, el modelo matemático se escribe:

$$\text{Log}(Y) = 0.9892 + 0.1458 X$$

La ecuación exponencial se escribe:

$$Y = (9.7545) (1.399)^X$$

La validación del modelo se efectúa de modo análogo a lo hecho con el modelo lineal, descrito en la sección anterior.

### 5.5. Regresión potencial o doble logarítmica

En ciertas ocasiones se puede tener que una función potencial tal como:

$$Y = c X^b$$

puede representar la relación entre  $X$  y  $Y$  en la muestra. En esta, se desea hallar  $c$  y  $b$ . Para esto, se toma el logaritmo en ambos lados de la ecuación, obteniendo:

$$\text{Log}(Y) = \text{Log}(c) + b \text{Log}(X)$$

El resultado es una transformación doble logarítmica, porque ambas variables se expresan en logaritmos. Si se hace:

$$\text{Log}(c) = a, \quad \text{Log}(X) = U \quad \text{Log}(Y) = W$$

Y la ecuación potencial se puede escribir:

$$W = a + b U$$

Para transformar nuevamente la ecuación a la forma original, se toma la función inversa del logaritmo, entonces:

$$c = 10^a$$

$$Y = (10^a) X^b$$

### Ejemplo:

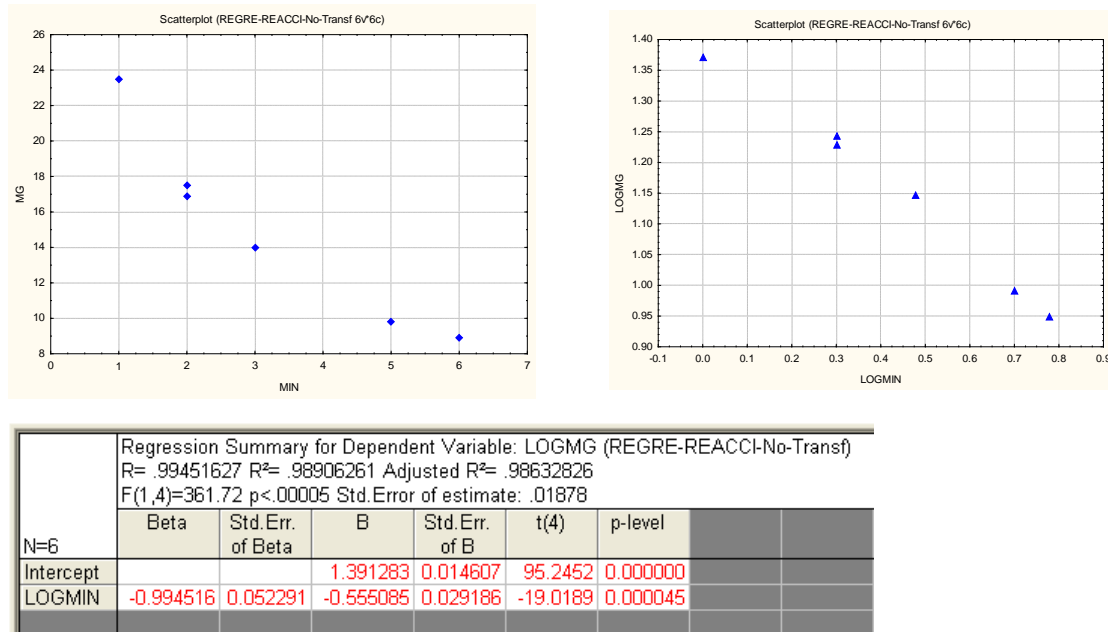
Las cantidades  $Y$  de una sustancia no transformada en seis reacciones similares después de  $X$  minutos están dadas en la siguiente Tabla.

<u>X, min</u>	<u>Y, mg</u>	<u>Log(X)</u>	<u>Log(Y)</u>
1	23.5	0.000	1.371
2	16.9	0.301	1.228
3	17.5	0.301	1.243
4	14.0	0.477	1.146
5	9.8	0.699	0.991
6	8.9	0.778	0.949



**Solución:**

Siguiendo la secuencia explicada previamente, tenemos:

**5.6. Modelo de regresión lineal múltiple**

Muchas aplicaciones del análisis de regresión involucran situaciones donde se tiene más de una variable de regresión. Un modelo de regresión que contiene más de un regresor recibe el nombre de modelo de regresión múltiple.

Como ejemplo, supóngase que la vida eficaz de una herramienta de corte depende de la velocidad de corte y el ángulo de la herramienta. El rendimiento de un proceso químico no solo puede estar afectado por la concentración de un elemento, sino por la concentración de varias sustancias, temperatura, etc. Un modelo de regresión múltiple que puede describir esta relación es el siguiente:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

donde  $Y$  representa la vida de la herramienta;  $X_1$  la velocidad de corte;  $X_2$  el ángulo de la herramienta; y  $\varepsilon$  un término de error aleatorio. Se utiliza el término *lineal* porque la ecuación del modelo de regresión múltiple es una función lineal de los parámetros conocidos  $\beta_0$ ,  $\beta_1$  y  $\beta_2$ . La ecuación del modelo de regresión describe un plano en el espacio tridimensional, Figura 5.10. El parámetro  $\beta_0$  es la intersección del plano. Los parámetros  $\beta_1$  y  $\beta_2$  se conocen como coeficientes de regresión parciales, ya que  $\beta_1$  mide el cambio esperado en  $Y$  por unidad de cambio de  $X_1$  cuando  $X_2$  se mantiene constante, y  $\beta_2$  mide el cambio esperado en  $Y$  por unidad de cambio en  $X_2$  cuando  $X_1$  se mantiene constante.

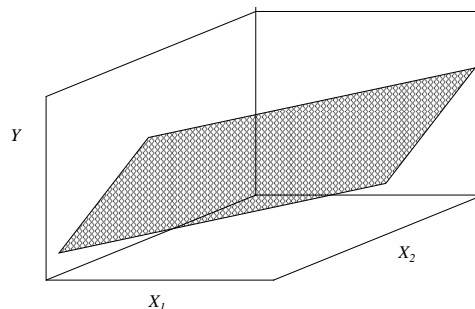


Figura 5.10: Representación gráfica de un modelo de regresión lineal múltiple con dos regresores.

En general, la variable dependiente o respuesta  $Y$ , puede estar relacionada con  $k$  variables independientes o regresores. El modelo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$$

recibe el nombre de modelo de regresión lineal múltiple con  $k$  variables de regresión. Este modelo describe un hiperplano en el espacio de dimensión  $k$  formado por las variables de regresión. Los modelos que tienen una estructura más compleja que la dada por esta ecuación se pueden analizar con frecuencia por técnicas de regresión lineal múltiple.

### 5.6.1 Estimación de los parámetros

Generalmente los coeficientes de regresión  $\beta_0, \beta_1, \beta_2, \dots, \beta_k$  así como la varianza de los errores son desconocidos y deben ser estimados a partir de datos tomados de la práctica real. Al igual que en el modelo de regresión lineal simple, se utiliza el criterio de los mínimos cuadrados para calcular esos coeficientes. En general, las ecuaciones que permiten el cálculo respectivo son:

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n X_{i1} + \hat{\beta}_2 \sum_{i=1}^n X_{i2} + \dots + \hat{\beta}_k \sum_{i=1}^n X_{ik} &= \sum_{i=1}^n Y_i \\ \hat{\beta}_0 \sum_{i=1}^n X_{i1} + \hat{\beta}_1 \sum_{i=1}^n X_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^n X_{i1} X_{i2} + \dots + \hat{\beta}_k \sum_{i=1}^n X_{i1} X_{ik} &= \sum_{i=1}^n X_{i1} Y_i \\ &\vdots \\ \hat{\beta}_0 \sum_{i=1}^n X_{ik} + \hat{\beta}_1 \sum_{i=1}^n X_{ik} X_{i1} + \hat{\beta}_2 \sum_{i=1}^n X_{ik} X_{i2} + \dots + \hat{\beta}_k \sum_{i=1}^n X_{ik}^2 &= \sum_{i=1}^n X_{ik} Y_i \end{aligned}$$

Estas son llamadas las ecuaciones normales. La solución de estas ecuaciones es relativamente sencilla aunque algo laborioso el cálculo manual dependiendo de la cantidad de regresores a considerar. En vista de que los programas estadísticos de computadora presentan la opción de regresión múltiple, la obtención del modelo de regresión se torna muy práctico con el uso de un programa de computadora.

### 5.6.2 Medidas de adecuación del modelo

Pueden emplearse varias técnicas para medir la adecuación de un modelo de regresión, entre las que se tiene las siguientes:

#### a) Coeficiente de determinación múltiple

Definido como:

$$R^2 = \frac{SS_R}{SS_{TO}} = 1 - \frac{SS_E}{SS_{TO}}$$

es una medida de la magnitud de la reducción de la variabilidad de  $Y$  obtenida mediante el empleo de las variables de regresión. Al igual que en el caso de la regresión simple  $0 \leq R^2 \leq 1$ . Sin embargo, un valor grande de  $R^2$  no necesariamente implica que el modelo de regresión sea bueno. La adición de una variable al modelo siempre aumenta  $R^2$ , sin importar si la variable es o no estadísticamente significativa. Es así como modelos que tienen valores grandes de  $R^2$  pueden proporcionar predicciones pobres de nuevas observaciones o estimaciones de la respuesta.

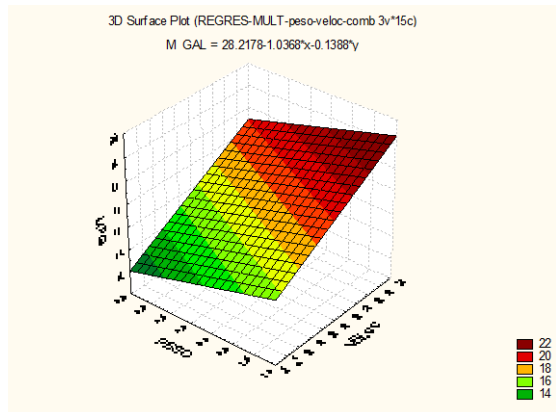
#### b) Análisis residual

Los residuos del modelo de regresión múltiple, definidos como  $e_i = Y_i - \hat{Y}_i$  desempeñan un papel importante al juzgar la adecuación del modelo, al igual que lo tienen con la regresión lineal simple. Para tal propósito son muy útiles las gráficas de los residuos. Una gráfica de probabilidad normal de los residuos es un elemento apropiado para juzgar sobre la validez del modelo; desviaciones de los residuos con respecto a la normalidad son indicio de que el modelo no es el más adecuado.

#### **Ejemplo:**

Para comprobar el efecto de velocidad en el consumo de gasolina, se probó automóviles de varios pesos a varias velocidades. Los datos se muestran en la siguiente Tabla:

Peso del vehículo, miles de libra	Millas por galón a las velocidades seleccionadas				
	30 MPH	40 MPH	50 MPH	60 MPH	70 MPH
2.29	21.55	20.07	19.11	17.83	16.72
3.98	18.25	20.00	16.32	15.77	13.61
5.25	18.33	19.28	15.62	14.22	12.74



Regression Summary for Dependent Variable: M_GAL (REGRES-MULT)						
R= .93719760 R²= .87833935 Adjusted R²= .85806257 F(2,12)=43.318 p<.000000 Std. Error of estimate: .97009						
N=15	Beta	Std. Err. of Beta	B	Std. Err. of B	t(12)	p-level
Intercept			28.21775	1.215012	23.22427	0.000000
PESO	-0.505354	0.100690	-1.03683	0.206584	-5.01893	0.000300
VELOC	-0.789276	0.100690	-0.13883	0.017711	-7.83871	0.000005

Analysis of Variance; DV: M_GAL (REGRES-MULT)					
Effect	Sums of Squares	df	Mean Squares	F	p-level
Regress.	81.52919	2	40.76460	43.31750	0.000003
Residual	11.29278	12	0.94107		
Total	92.82197				

El modelo matemático para los datos es:

$$Y = 28.22 - 1.04 (\text{Peso}) - 0.14 (\text{Veloc})$$

Los regresores,  $\beta_0$ ,  $\beta_1$ ,  $\beta_2$  son significativos en el modelo, (ver “p-level” para cada regresor).

De la Tabla de ANOVA se deduce la validez del modelo íntegramente, (ver “p-level”).

Los valores observados, calculados y residuales se observa en el cuadro inferior; también el valor estimado para valores específicos de las variables independientes.

Predicted & Residual Values (REGRES-MULT)			
Case No.	Observed Value	Predicted Value	Residual
1	21.55000	21.67842	-0.12842
2	20.07000	20.29008	-0.22008
3	19.11000	18.90175	0.20825
4	17.83000	17.51342	0.31658
5	16.72000	16.12508	0.59492
6	18.25000	19.92618	-1.67618
7	20.00000	18.53784	1.46216
8	16.32000	17.14951	-0.82951
9	15.77000	15.76118	0.00882
10	13.61000	14.37284	-0.76284
11	18.33000	18.60941	-0.27941
12	19.28000	17.22107	2.05893
13	15.62000	15.83274	-0.21274
14	14.22000	14.44441	-0.22441
15	12.74000	13.05607	-0.31607
Minimum	12.74000	13.05607	-1.67618
Maximum	21.55000	21.67842	2.05893
Mean	17.29467	17.29467	0.00000
Median	17.83000	17.22107	-0.21274

Predicting Values for (REGRES-MULT)			
Variable	B-Weight	Value	B-Weight * Value
PESO	-1.03683	2.50000	-2.59207
VELOC	-0.13883	55.00000	-7.63583
Intercept			28.21775
Predicted			17.98985
-95.0%CL			17.15388
+95.0%CL			18.82582

## 5.7. Modelos de regresión polinomiales

El modelo lineal:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \varepsilon$$

es un modelo general que puede emplearse para ajustar cualquier relación lineal en los parámetros desconocidos  $\beta_i$ . Esto incluye la clase importante de modelos de regresión polinomiales. Por ejemplo, el polinomio de segundo grado en una variable:

$$Y = \beta_0 + \beta_1 X_1 + \beta_{11} X_1^2 + \varepsilon$$

y el polinomio de segundo grado de dos variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \varepsilon$$

son modelos de regresión no lineales.

Los modelos de regresión polinomiales se utilizan mucho cuando la respuesta es curvilínea y el posible aplicar los principios generales de la regresión múltiple.

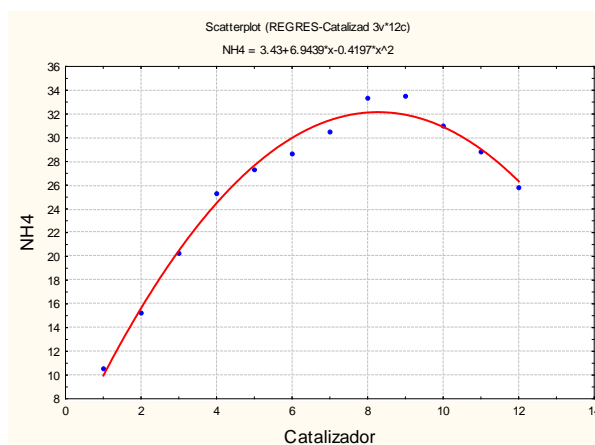
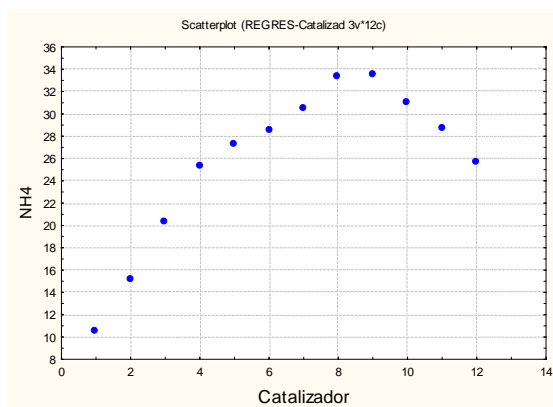
### Ejemplo:

La cantidad de un catalizador en disolución se puede determinar cuantitativamente en relación al efecto catalítico que produce. Así, un investigador hace un experimento para determinar si el catalizador actúa catalíticamente sobre el sustrato elegido y observar si hay aparición de  $\text{NH}_4$ . Para medir la cantidad de producto formado,  $\text{NH}_4$ , se utiliza HCl. Los datos se muestran en la siguiente Tabla:

Catalizador; mL	$\text{NH}_4$	(Catalizador) <sup>2</sup>
1	10.53	1.0
2	15.18	4.0
3	20.25	9.0
4	25.31	16.0
5	27.33	25.0
6	28.60	36.0
7	30.52	49.0
8	33.24	64.0
9	33.50	81.0
10	31.00	100.0
11	28.75	121.0
12	25.75	144.0

Analizar los datos del experimento por métodos de regresión.

### Solución:



Regression Summary for Dependent Variable: NH4 (REGRES-Cat)						
R= .99253115 R <sup>2</sup> = .98511809 Adjusted R <sup>2</sup> = .98181100 F(2,9)=297.88 p<.00000 Std.Error of estimate: .96214						
	Beta	Std. Err. of Beta	B	Std. Err. of B	t(9)	p-level
N=12						
Intercept			3.430000	0.994401	3.4493	0.007283
Catalizador	3.50947	0.177749	6.943906	0.351697	19.7440	0.000000
X2	-2.83285	0.177749	-0.419730	0.026336	-15.9374	0.000000

Analysis of Variance; DV: NH4 (REGRES-Catalizad)					
Effect	Sums of Squares	df	Mean Squares	F	p-level
Regress.	551.5056	2	275.7528	297.8805	0.000000
Residual	8.3314	9	0.9257		
Total	559.8370				

Las interpretaciones se dejan para ejercicio del lector.

### 5.8. Ejemplos de aplicación

Con los siguientes ejemplos se desarrollaran aplicaciones prácticas de los conceptos de regresión lineal

#### 5.8.1. Evaluación en el cambio de rendimiento de proceso

En uno de los pasos para eliminación de agua de concentrados finos en una Planta Concentradora se usa un proceso de filtración en batch usando una serie de filtros de placa-y-lona. Un día, y sin aparente razón, la capacidad del circuito de filtración se redujo en un 20% aproximadamente y ha permanecido por varios días en tal situación. El Superintendente de Planta quiere saber por qué y si será necesario comprar filtros adicionales para recuperar los anteriores niveles.

El tiempo requerido para coleccionar un volumen  $V$  de filtrado de una pulpa con una concentración de sólidos de concentración  $C$  (volumen de sólido por unidad de volumen de filtrado) a una presión constante  $\Delta P$  se representa por el siguiente modelo semiteórico:

$$t = \frac{V^2 r \mu C}{2 A^2 \Delta P} \quad (5.1)$$

donde  $\mu$  es la viscosidad del fluido,  $A$  es el área de filtración y  $r$  es la resistencia específica de la torta. Cuando la torta filtrada es compresible, la resistencia de torta  $r$  se incrementa con el incremento de la presión,  $\Delta P$  según la siguiente relación empírica:

$$r = r_o + \beta (\Delta P)^n \quad (5.2)$$

Los valores previamente determinados de las constantes para el concentrado normal de la planta son

$$r_o = 3.1 \times 10^6 \text{ (ft)}^{-2} \quad \beta = 1.1 \times 10^5 \text{ (psi}^{-0.9} \text{ ft}^{-2}) \quad n = 0.90$$

Un análisis de los registros de operación de planta determina que no hubo cambios significativos en las condiciones de operación en las últimas semanas a excepción del tamaño del concentrado, que en la actualidad es algo más fino de lo que era antes al problema surgido. Al revisarse la literatura técnica se encuentra que el parámetro  $r_o$  de la ecuación (5.2) es sensible a la distribución de tamaños; por lo tanto, se decide reevaluar la relación para un material más fino.

Se efectúan unas pruebas experimentales cuyos resultados se anotan en la siguiente Tabla:

**DATOS EXPERIMENTALES**

$\Delta P_i \text{ (psi)}$	$r_i \text{ (ft)}^{-2}$
40	$10.34 \times 10^6$
40	$10.92 \times 10^6$
70	$13.13 \times 10^6$
70	$12.23 \times 10^6$
60	$12.31 \times 10^6$
60	$11.76 \times 10^6$
30	$9.85 \times 10^6$
30	$9.71 \times 10^6$
50	$11.57 \times 10^6$
50	$11.16 \times 10^6$
80	$13.21 \times 10^6$
80	$13.72 \times 10^6$

El procedimiento a seguir para analizar y proponer una solución al problema surgido, puede efectuarse según el siguiente esquema:

A) Revisar la validez de la ecuación 5.2:

1. Inténtese ajustar los datos experimentales con los siguientes modelos:

$$a) r = r_o + \beta X$$

$$b) r = r_o e^{\beta X}$$

$$c) r = r_o X^\beta$$

considerando  $X = \Delta P^g$ , intente la linearización de las ecuaciones  $b$  y  $c$ .

2. Identificar el mejor modelo en base a un análisis de residuos.
3. Haga un gráfico de la curva ajustada y el intervalo al 95% de confianza.

B) Verificar los cambios en los parámetros  $r_o$  y  $\beta$  de la ecuación 4.2

1. Efectuar la prueba de comparación de  $r_o$  estimado de los datos experimentales con el valor previo al problema.
2. Efectuar la prueba de comparación de  $\beta$  estimado de los datos experimentales con el valor previo al problema.

### 5.8.2. Análisis de modelos de regresión

#### 1. Velocidad de una máquina:

El número de piezas producidas por una maquina, ( $Y$ ) se sabe que esta relacionada linealmente a la velocidad ajustada en la maquina ( $X$ ). Los datos de la Tabla siguiente fueron obtenidos recientemente:

$i$	1	2	3	4	5	6	7	8	9	10	11	12
$X_i$	200	400	300	400	200	300	300	400	200	400	200	300
$Y_i$	28	75	37	53	22	58	43	96	46	52	30	69

- a) Asumiendo que el modelo de regresión es apropiado, obténgase una función de regresión y un gráfico de residuos vs.  $X$ . Qué se observa del gráfico de residuos?
- b) Calcular la varianza de muestra  $S^2$  de la  $Y$  observaciones para cada una de las tres velocidades de máquina:  $X = 200, 300$  y  $400$ . Que se sugiere de los valores des la varianza de las tres muestras con respecto a la igualdad de esos valores?
- c) Calcular:  $\frac{\bar{Y}}{\sqrt{S}}$ ,  $\frac{\bar{Y}}{S}$  y  $\frac{\bar{Y}}{S^2}$  para cada uno de los niveles de  $X$ . Sugiera una transformación apropiada para estabilizar la varianza.
- d) Haga la transformación sugerida en la parte  $c$  y obtenga la línea de regresión con los datos transformados. Haga el gráfico de residuos vs.  $X$  ¿Qué concluye de este gráfico?

#### 2. Modelo de hidrociclones

Una compañía minera esta tratando de desarrollar un modelo para hidrociclones que utilizan en su planta de molienda. Se decide utilizar el modelo empírico de Rao y Lynch, el cual consiste de cuatro ecuaciones. Para una de ellas, la ecuación de agua en el overflow, se han desarrollado dos modelos:

Modelo expandido:

$$WOF = K_1 + B_1(WF) + B_2(SPIG) + B_3(WF \cdot SPIG) + B_4(WF^2) + B_5(SPIG)^2$$

Modelo reducido:

$$WOF = K_2 + C_1(WF) + c_2(SPIG)$$

donde:

WOF = Flujo másico de agua en el overflow (tph)

WF = Flujo másico de agua en el pulpa de alimentación (tph)

SPIG = diámetro del apex (pulgadas)

- a) Ajustar los dos modelos a los datos experimentales obtenidos.
- b) Simplificar el modelo expandido, en lo posible, eliminando una variable a la vez, hasta encontrar la forma más simple.
- c) Compare el modelo reducido obtenido con el modelo reducido de Rao y Lynch.

DATOS EXPERIMENTALES		
<u>WOF (tph)</u>	<u>WF (tph)</u>	<u>SPIG (pulq.)</u>
38.4	50.0	2.5
34.8	50.0	3.0
28.2	50.0	3.5
66.1	75.0	2.5
63.3	75.0	3.0
56.4	75.0	3.5
93.1	100.0	2.5
86.0	100.0	3.0
81.8	100.0	3.5
119.5	125.0	2.5
114.7	125.0	3.0
110.9	125.0	3.5

**EJERCICIOS GRUPO 5**

1. Se usa un reactivo químico para obtener un precipitado de una sustancia en una solución dada. Los datos son los siguientes:

Reactivo	7.2	4.8	5.2	4.9	5.4	6.4	6.8	8.0	6.0	6.7	7.0	8.0	7.3	4.6	4.2
Precipitado	8.4	5.4	6.3	6.8	8.0	11.1	12.3	13.3	8.4	9.5	10.4	12.7	10.3	7.0	5.1

- Haga un diagrama de dispersión.
  - Determine la ecuación de la recta y represéntela en el diagrama anterior.
  - ¿Cuál es la cantidad de precipitado estimada si se usa 7.1 de reactivo?
  - Pruebe la hipótesis al 1 % de significación de que  $\beta = 0$  y de que  $\beta = 2$
2. La Tabla siguiente muestra los resultados de las medidas de resistividades eléctricas en Ohm-cm \*  $10^{-6}$  del platino a diferentes temperaturas ( $^{\circ}\text{K}$ ):

Temperatura	100	200	300	400	500
Resistividad	4.1	8.0	12.6	16.3	19.4

- Haga un diagrama de dispersión.
  - Determine la ecuación de la recta y represéntela en el diagrama anterior.
  - ¿Cuál es la estimación de la resistividad del platino cuando la temperatura es  $350^{\circ}\text{K}$ ?
  - Pruebe la hipótesis al 1 % de significación de que  $\beta = 0$
3. La Tabla siguiente 'X' es la fuerza de tracción aplicada a una probeta de acero en miles de libras; 'Y' es la elongación resultante en milésimas de pulgada. Suponiendo que existe una relación lineal entre ambas variables, calcular los parámetros de la línea de regresión con un intervalo de confianza a un nivel de 0.95 para  $\beta$ .

X	1	2	3	4	5	6
Y	15	35	41	63	77	84

4. Un estudiante obtuvo los siguientes datos sobre la cantidad de bromuro de potasio que se puede disolver en 100 g de agua a distintas temperaturas

Temperatura $^{\circ}\text{C}$	0	10	20	30	40	50
g.	52	60	64	73	76	81

- Calcular la ecuación de regresión.
  - Probar la hipótesis nula  $\beta = 0.5$  a un nivel de significación de 0.05.
5. Los datos de concentración de licor verde  $\text{Na}_2\text{S}$  y la producción de papel de una máquina son:

Nro de observación	1	2	3	4	5	6	7	8	9	10	11	12	13
$\text{Na}_2\text{S}$ , g/l	40	42	49	46	44	48	46	43	53	52	54	57	58
Producción, ton/día	825	830	890	895	890	910	915	960	990	1010	1012	1030	1050

- Ajuste el modelo de regresión lineal simple con la concentración de licor verde como Y y la producción como X. Dibuje el diagrama de dispersión de los datos y del modelo ajustado de dichos datos.
- Encuentre el valor ajustado de Y que corresponde a  $X = 910$  así como el residuo correspondiente.
- Encuentre la concentración promedio de licor verde cuando la tasa de producción es de 950 toneladas por día.
- Pruebe la significancia de la regresión con  $\alpha = 0.05$ . encuentre el valor  $p$  de esta prueba.
- Pruebe  $H_0 : \beta_0 = 0$  contra  $H_a : \beta_1 \neq 0$  con  $\alpha = 0.05$ . ¿Cuál es el valor  $p$  de esta prueba?



f. Haga una gráfica de los residuos de  $y$  calculado contra  $x$ . Comente las gráficas.

6. En una planta se destila aire líquido para producir oxígeno, nitrógeno y argón. Se cree que el porcentaje de impureza del oxígeno está linealmente relacionado con la cantidad de impurezas que hay en el aire, medida mediante el "conteo de contaminación" en partes por millón (ppm). Los datos son los siguientes:

Pureza, %	93.3	92.0	92.4	91.7	94.0	94.6	93.6	93.1	93.2	92.9	92.2	91.3	90.1	91.6	91.9
Contam., ppm	1.10	1.45	1.36	1.59	1.08	0.75	1.20	0.99	0.83	1.22	1.47	1.81	2.03	1.75	1.68

- Ajuste un modelo de regresión lineal a los datos.
- ¿Parece razonable la relación lineal entre la pureza y el conteo de la contaminación?
- Construya el intervalo de confianza de 95% para la pendiente y la ordenada en el origen del modelo de regresión lineal.
- Pruebe la significancia de la regresión con  $\alpha = 0.05$ . encuentre el valor  $p$  de esta prueba.
- Pruebe  $H_0 : \beta_0 = 0$  contra  $H_a : \beta_0 \neq 0$  con  $\alpha = 0.05$ . ¿Cuál es el valor  $p$  de esta prueba?
- Haga una gráfica de los residuos de  $y$  calculado contra  $x$ . Comente las gráficas.

7. La resistencia de papel utilizado en la fabricación de cajas de cartulina ( $y$ ) está relacionada con la concentración de madera dura en la pulpa original ( $x$ ). Bajo condiciones controladas, una planta piloto fabrica 16 muestras, cada una con un lote diferente de pulpa, y mide la resistencia a la tensión. Los datos obtenidos son los siguientes:

Y	101.4	117.4	117.1	106.2	131.9	146.9	146.8	133.9
X	1.0	1.5	1.5	1.5	2.0	2.0	2.2	2.4

Y	111.0	123.0	125.1	145.2	134.3	144.5	143.7	146.9
X	2.5	2.5	2.8	2.8	3.0	3.0	3.2	3.3

- Ajuste un modelo de regresión lineal simple con los datos.
- Pruebe la significancia de la regresión con  $\alpha = 0.05$ .
- Construya el intervalo de confianza del 90% para la pendiente del modelo.
- Pruebe la falta de ajuste con  $\alpha = 0.05$ . ¿Cuál es el valor  $p$  de esta prueba?

8. Los datos siguientes muestran la salida de CD de un generador de viento ( $y$ ) y la velocidad del viento ( $x$ ):

Observación #	1	2	3	4	5	6	7	8	9	10	11	12	13
Vel. Viento, mph	5.00	6.00	3.40	2.70	10.00	9.70	9.55	3.05	8.15	6.20	2.90	6.35	4.60
Salida de CD.	1.582	1.822	1.057	0.500	2.236	2.386	2.294	0.558	2.166	1.866	0.653	1.930	1.562

Observación #	14	15	16	17	18	19	20	21	22	23	24	25
Vel. Viento, mph	5.80	7.40	3.60	7.85	8.80	7.00	5.45	9.10	10.20	4.10	3.95	2.45
Salida de CD	1.737	2.088	1.137	2.179	2.112	1.800	1.501	2.303	2.310	1.194	1.144	0.123

- Dibuje un diagrama de dispersión para estos datos. ¿Qué tipo de relación parece ser la más apropiada entre  $y$  y  $x$ ?
  - Ajuste un modelo de regresión lineal simple para estos datos.
  - Pruebe la significancia de la regresión utilizando  $\alpha = 0.05$  ¿Qué conclusiones puede obtenerse?
  - Haga una gráfica de residuos del modelo de regresión lineal simple contra valor calculado y contra la velocidad del viento. ¿Qué puede concluirse con respecto a la adecuación del modelo?
  - Dibuje un diagrama de dispersión de  $y$  contra  $1/x$  ¿Qué tipo de relación parece ser la más apropiada?
  - Ajuste el modelo de regresión lineal simple que relaciones  $y$  contra  $1/x$ . Pruebe la significancia de la regresión utilizando  $\alpha = 0.05$  y haga una gráfica de los residuos de este modelo contra el valor calculado de  $y$  y  $1/x$ . ¿Qué conclusiones puede obtenerse sobre la adecuación del modelo?
9. El rendimiento de una reacción química depende de la concentración del reactivo y de la temperatura de operación. Los datos obtenidos son los siguientes:

Rendimiento	81	89	83	91	79	87	84	90
Concentrac.	1.00	1.00	2.00	2.00	1.00	1.00	2.00	2.00
Temperatura	150	180	150	180	150	180	150	180

- Ajuste el modelo lineal de regresión múltiple a los datos.
  - Pruebe la significancia del modelo de regresión
10. Ajuste el modelo de regresión polinomial de segundo orden utilizando los siguientes datos:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \varepsilon$$

Y	26	24	175	160	163	55	62	100	26	30	70	71
$X_1$	1.0	1.0	1.5	1.5	1.5	0.5	1.5	0.5	1.0	0.5	1.0	0.5
$X_2$	1.0	1.0	4.0	4.0	4.0	2.0	2.0	3.0	1.5	1.5	2.5	2.5

11. Se piensa que la potencia consumida al mes por una planta química está relacionada con la temperatura ambiente promedio ( $X_1$ ), el número de días al mes, ( $X_2$ ), la pureza producto del producto ( $X_3$ ) y las toneladas del producto producidas ( $X_4$ ). Los datos correspondientes al año pasado son:

Y	240	236	290	274	301	316	300	296	267	276	288	261
$X_1$	25	31	45	60	65	72	80	84	75	60	50	38
$X_2$	24	21	24	25	25	26	25	25	24	25	25	23
$X_3$	91	90	88	87	91	94	87	86	88	91	90	89
$X_4$	100	95	110	88	94	99	97	96	110	105	100	98

- Ajuste un modelo de regresión lineal múltiple a los datos.
- Prediga el consumo de potencia para un mes en el que  $X_1 = 75$  °F,  $X_2 = 24$  días,  $X_3 = 90\%$  y  $X_4 = 98$  toneladas.
- Pruebe la significancia de la regresión utilizando  $\alpha = 0.01$ . ¿Cuál es el valor  $P$  de esta prueba?
- Utilice la prueba  $t$  para evaluar la contribución al modelo de cada variable de regresión. Si se emplea  $\alpha = 0.01$ , ¿qué conclusiones se puede obtener?
- Calcule  $R^2$ . Interprete esta cantidad.
- Haga una gráfica de residuos contra valores calculados. Interprete la gráfica.

12. Considere los datos siguientes, los cuales son resultado de un experimento para determinar el efecto de  $X$  = tiempo de prueba en horas a una temperatura particular sobre  $Y$  = cambio en la viscosidad del aceite:

Y	-1.42	-1.39	-1.55	-1.89	-2.43	-3.15	-4.05	-5.15	-6.43	-7.89
X	0.25	0.5	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50

- Ajuste los datos a un polinomio de segundo orden.
  - Pruebe con  $\alpha = 0.05$  la significancia de la regresión.
  - Pruebe con  $\alpha = 0.05$  la hipótesis de que  $\beta_{11} = 0$
13. A continuación se presentan los datos sobre solubilidad de la fracción molar de un soluto  $Y$ , a temperatura constante y los parámetros de Hansen de solubilidad parcial por dispersión, dipolo y enlace de hidrógeno respectivamente:

# Obs.	1	2	3	4	5	6	7	8	9
Y	0.2220 0	0.39500	0.42200	0.43700	0.42800	0.46700	0.44400	0.37800	0.49400
$X_1$	7.3	8.7	8.8	8.1	9.0	8.7	9.3	7.6	10.0
$X_2$	0.0	0.0	0.7	4.0	0.5	1.5	2.1	5.1	0.0
$X_3$	0.0	0.3	1.0	0.2	1.0	2.8	1.0	3.4	0.3

# Obs.	10	11	12	13	14	15	16	17	18
Y	0.4560 0	0.45200	0.11200	0.43200	0.10100	0.23200	0.30600	0.09230	0.11600
$X_1$	8.4	9.3	7.7	9.8	7.3	8.5	9.5	7.4	7.8
$X_2$	3.7	3.6	2.8	4.2	2.5	2.0	2.5	2.8	2.8
$X_3$	4.1	2.0	7.1	2.0	6.8	6.6	5.0	7.8	7.7

# Obs.	19	20	21	22	23	24	25	26
Y	0.0764 0	0.43900	0.09440	0.11700	0.07260	0.04120	0.25100	0.00002
$X_1$	7.7	10.3	7.8	7.1	7.7	7.4	7.3	7.6
$X_2$	3.0	1.7	3.3	3.9	4.3	6.0	2.0	7.8
$X_3$	8.0	4.2	8.5	6.6	9.5	10.9	5.2	20.7

- Ajuste el modelo:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{33} X_3^2 + \varepsilon$$

- Pruebe la significancia de la regresión con  $\alpha = 0.05$
- Haga una gráfica de los residuos y comente sobre la adecuación del modelo.