



1970
2020

FICH


- UNIVERSIDAD NACIONAL DEL LITORAL
- **FACULTAD DE INGENIERÍA Y CIENCIAS
HÍDRICAS**

• **ESTADÍSTICA**

- **INGENIERÍA EN INFORMÁTICA**
- *MG. SUSANA VANLESBERG*



ANÁLISIS DE REGRESIÓN Y CORRELACIÓN

- 
- **RETOMAMOS EL ESTUDIO DE VARIABLES QUE SE DISTRIBUYEN DE FORMA CONJUNTA.**
 - **La asociación entre variables se estudia a través de dos aspectos:**


- ***ANÁLISIS DE REGRESIÓN:*** permite encontrar el modelo que vincula a las variables en cuestión, brindando un mecanismo de pronóstico.

- ***ANÁLISIS DE CORRELACIÓN:*** determina la medida del grado de exactitud de la relación establecida entre las variables.

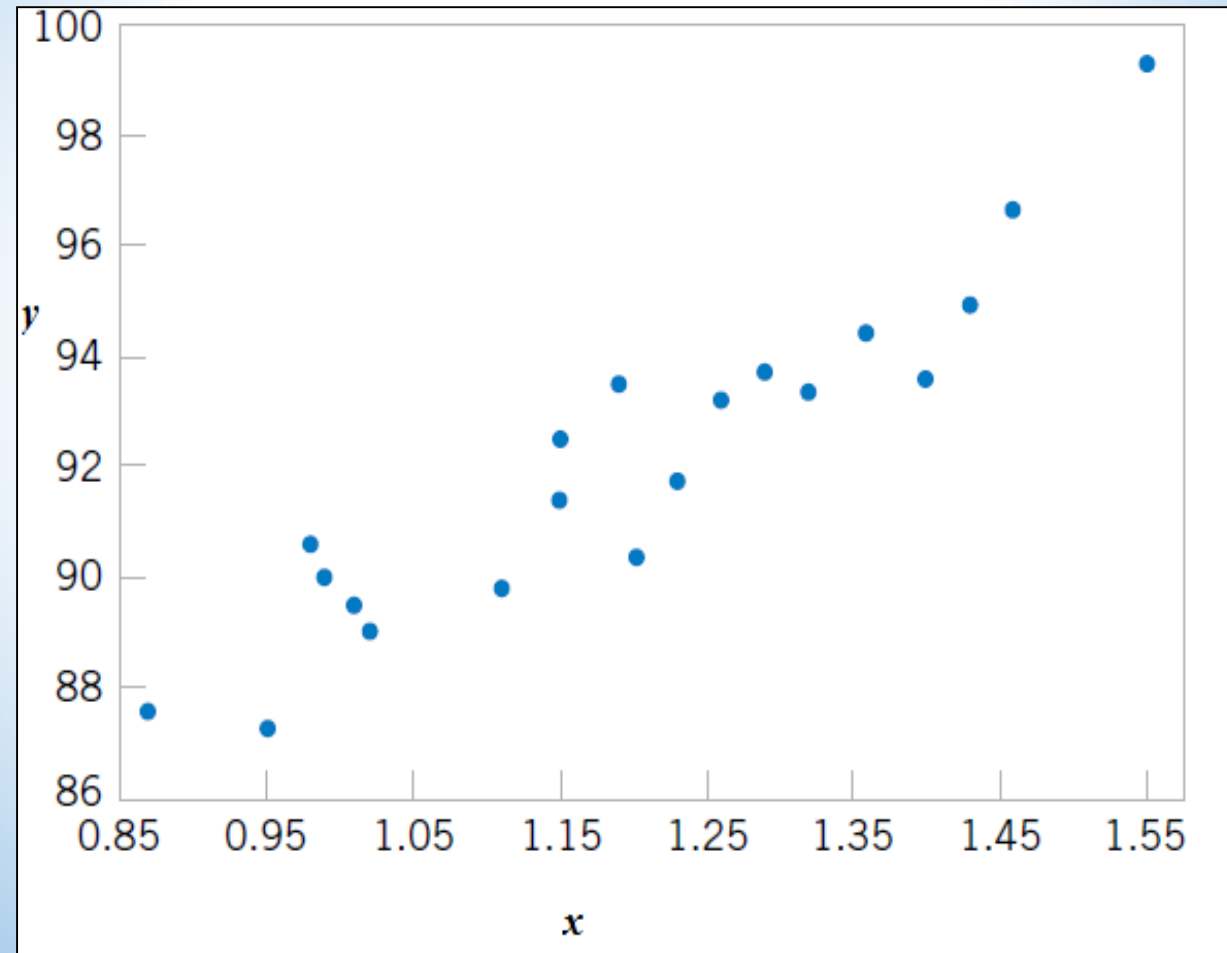
- Ejemplo: una empresa de servicios públicos establece la relación entre la temperatura diaria y la demanda de electricidad, el objetivo, predecir la necesidad del consumo de electricidad considerando las temperaturas diarias que se esperan para el mes siguiente.
- Algunas veces los directivos de empresas se apoyan en la intuición para juzgar la relación entre dos variables.
- Sin embargo, cuando los datos están disponibles, puede emplearse el procedimiento llamado *análisis de regresión* para obtener una ecuación que establezca la relación entre las variables.



Análisis de regresión

- 
- Se comienza realizando el gráfico que permite visualizar los valores de las variables, es lo que se denomina **Dispersiograma**, ya que muestra la variabilidad o dispersión existente entre ambas variables.

DISPERSIOGRAMA






REGRESIÓN LINEAL SIMPLE

- Es el análisis en el que se estudia la relación en la que interviene una variable independiente y una variable dependiente y que se aproxima mediante una línea recta.
- Al análisis en el que intervienen dos o más variables independientes se le llama ***análisis de regresión múltiple***.

ECUACIÓN DE REGRESIÓN

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$


$$Y_i = \frac{\alpha + \beta X_i}{\text{I}} + \frac{\varepsilon_i}{\text{II}}$$

- 
- ***I parte sistemática***
 - ***II parte aleatoria***
 - ***Debido a la parte aleatoria, el proceso de obtención del modelo no es como la determinación del ajuste de una función matemática a una serie de puntos.***

α y β parámetros del modelo, que deberán ser estimados

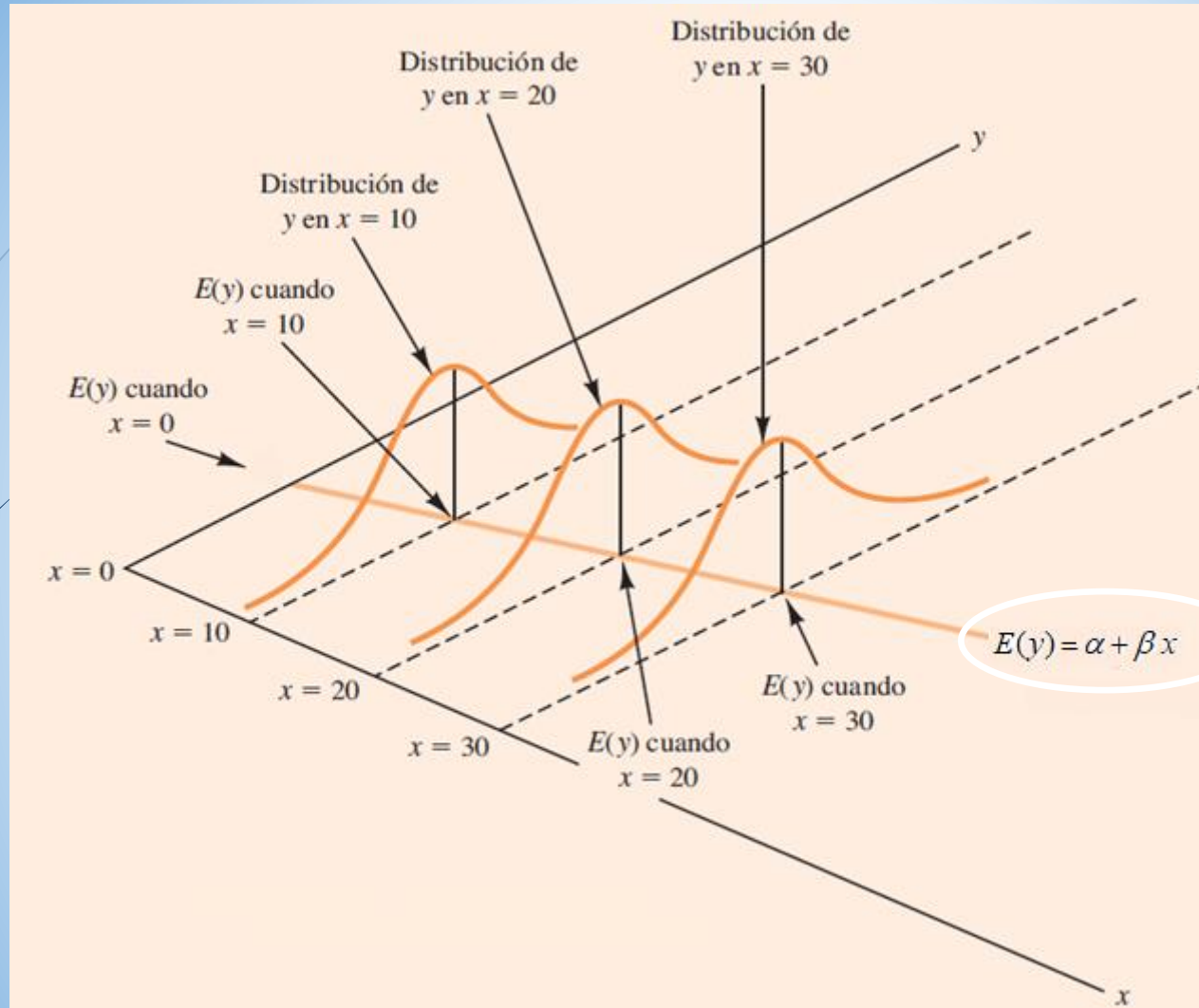
X_i variable independiente, fija, conocida, variable explicativa.

ε término de error aleatorio

Supuestos del modelo de regresión

- -La varianza de ϵ , es la misma para todos los valores de x .
$$\text{Var}(\epsilon) = \sigma^2$$
- **Implicancia.** La varianza de Y respecto al modelo de regresión es igual a σ^2 y es la misma para todos los valores de x .
- - Los valores de ϵ son independientes.
- **Implicancia.** El valor de ϵ correspondiente a un determinado valor de x no está relacionado con el valor de ϵ para cualquier otro valor; por tanto, el valor de y correspondiente a un valor particular de x no está relacionado con el valor de y de ningún otro valor de x .
- - El término del error ϵ es una variable aleatoria distribuida normalmente con valor esperado cero: $E(\epsilon) = 0$.
- **Implicancia.** Como Y es una función lineal de ϵ , también será una variable aleatoria distribuida normalmente.

- Lo que se determina es que para cada valor fijo de x existen distintos valores de la variable dependiente, ya que ella es una variable aleatoria y eso provoca que se tengan subpoblaciones para cada valor de x .
- Cada una de estas distribuciones tiene su propia media o valor esperado.
- A la ecuación que describe la relación entre el valor esperado de y , que se simboliza $E(y)$, y x se le llama **ECUACIÓN DE REGRESIÓN**.



Ecuación
de
Regresión

Significado de los parámetros

α : intercepción de la línea de regresión con el eje Y.

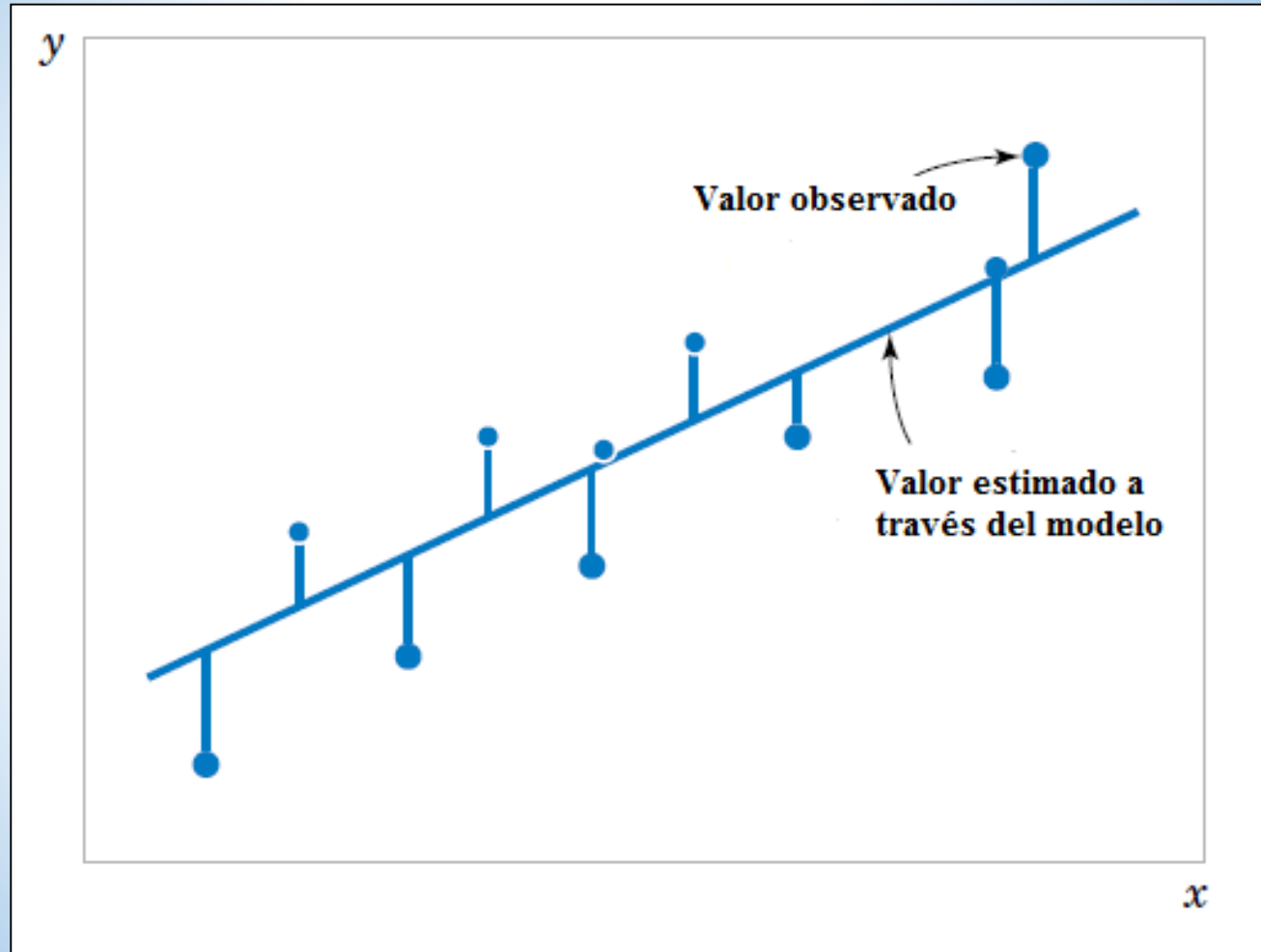
β : pendiente de la recta, proporción de cambio en la media de la distribución de probabilidad de Y por unidad de cambio de X.

ESTIMACIÓN DE LOS PARÁMETROS DEL MODELO

- **Método de Mínimos cuadrados**

Debido a que se pretende encontrar el mejor modelo que ajuste a la nube de puntos se utiliza este método.

- ***Se parte de considerar que la subpoblación de Y es normal, y que la suma de los cuadrados de las desviaciones de las observaciones***



Se parte de considerar las distancias entre valores observados y estimados a través del modelo de regresión

$$S = \sum_{i=1}^n [Y_i - (\alpha + \beta X_i)]^2$$

$$S = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad \hat{Y}_i = a + bX_i$$

$$\text{luego} \quad S = \sum_{i=1}^n [Y_i - (a + bX_i)]^2$$

Los estimadores a y b de los parámetros serán aquellos que minimicen el valor de S:

$$\frac{\partial S}{\partial \alpha} = 0$$


$$\frac{\partial S}{\partial \beta} = 0$$

$$\frac{\partial S}{\partial \alpha} = -2 \sum_{i=1}^n (Y_i - \alpha - \beta X_i)$$


$$\frac{\partial S}{\partial \beta} = -2 \sum_{i=1}^n X_i (Y_i - \alpha - \beta X_i)$$

Luego:

$$\begin{cases} \sum_{i=1}^n Y_i - na - b \sum_{i=1}^n X_i = 0 \\ \sum_{i=1}^n X_i Y_i - a \sum_{i=1}^n X_i - b \sum_{i=1}^n X_i^2 = 0 \end{cases}$$




$$b = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$b = \frac{m_{1,1}}{S_x^2} = \frac{\text{COV}}{S_x^2}$$

- 
- **El análisis de regresión no puede entenderse como un procedimiento para establecer una relación de causa y efecto entre las variables. Sólo indica cómo o en qué medida las variables están relacionadas una con otra. Cualquier conclusión acerca de una relación causa y efecto debe basarse en los conocimientos de los especialistas en la aplicación de que se trate.**
 - ***Hay que tener cuidado con el uso de la ecuación de regresión estimada para hacer predicciones. Fuera del rango de valores de la variable independiente no puede asegurarse que esta relación siga siendo válida.***



Varianza de la regresión

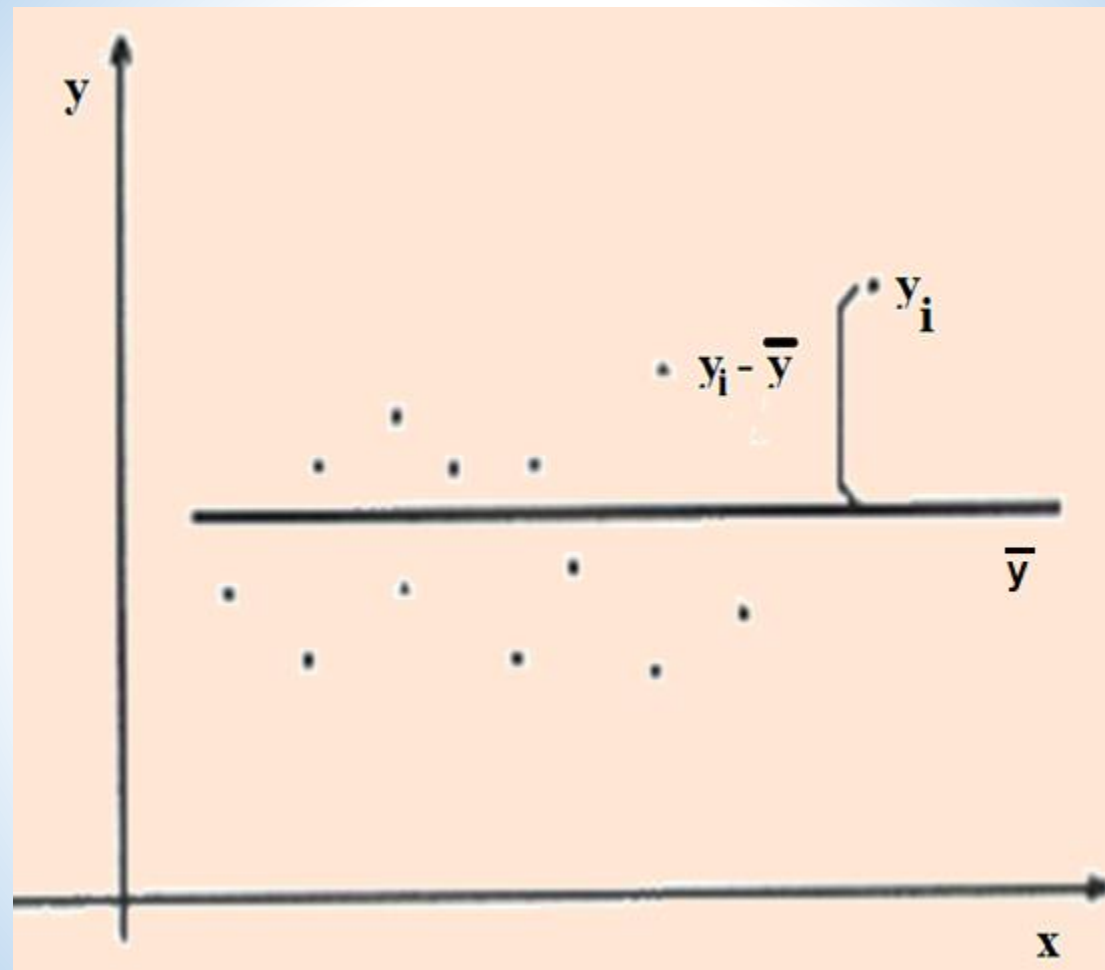
- 
- 
- Se suele llamar ecuación predictiva a la ecuación de regresión, ya que su principal objetivo es predecir valores medios de la variable dependiente asociados con un valor dado de la variable independiente.
 - Para saber si realmente es conveniente utilizar esta ecuación como herramienta de predicción, puede analizarse la variabilidad del valor estimado a través del modelo de regresión.

- La medida numérica de la desviación de las observaciones respecto al modelo es el estimador de la varianza de la regresión poblacional:

- $S^2_{y/x} = S_e^2$


- ***El análisis de la varianza de regresión se basa en la partición de la suma de cuadrados.***

La variación de las variables dependientes Y_i generalmente se mide en términos de las desviaciones respecto al valor medio:

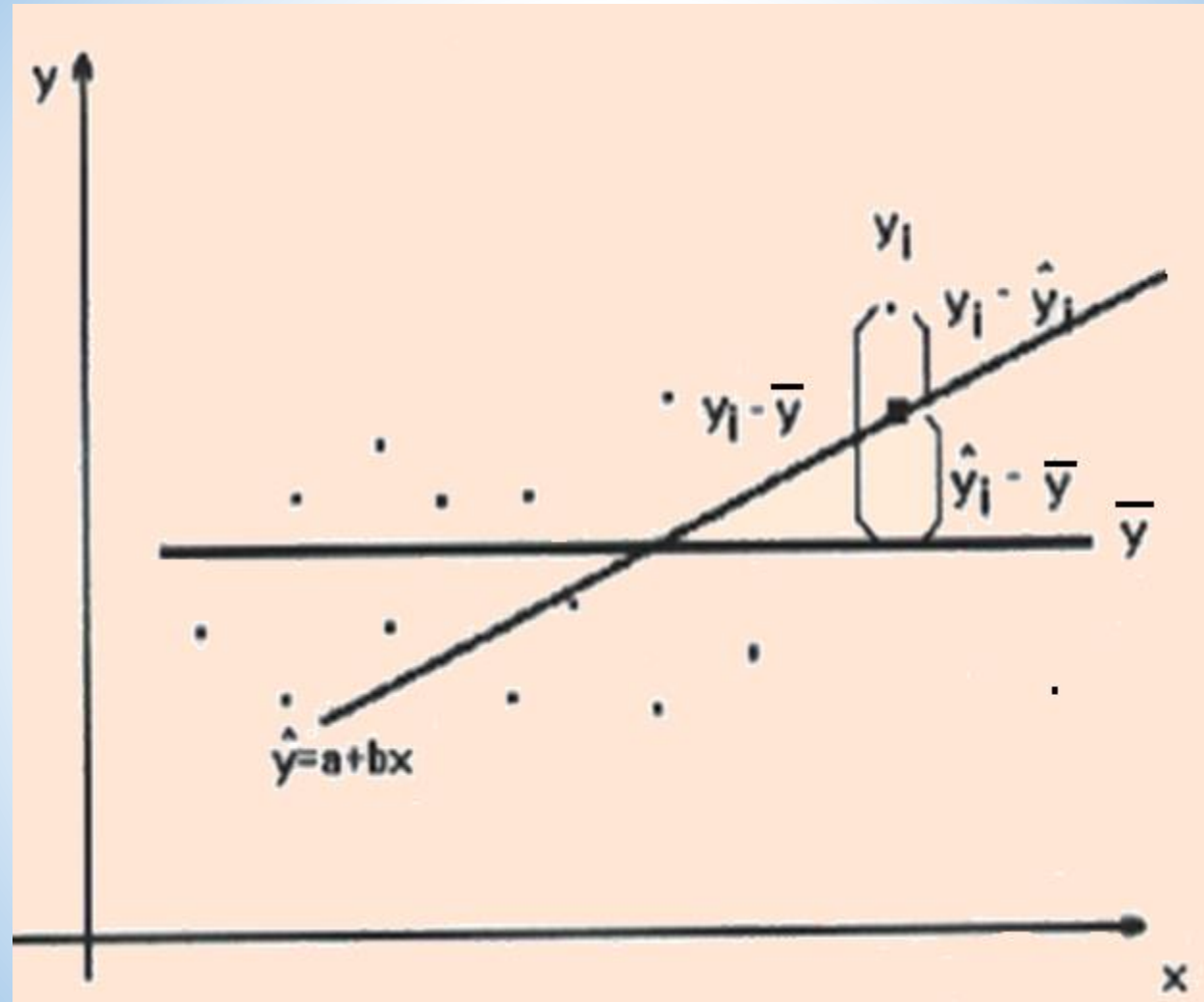


- La variación total siempre se mide respecto al valor medio:

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


- 
- **Cuanto mayor es este valor, mayor es la variación de la curva ajustada respecto a las observaciones.**

Utilizando el modelo ajustado, *la variación total queda expresada de acuerdo a la diferencia con los valores ajustados:*



- Con base en el modelo de regresión y sus supuestos, se puede decir que σ^2 , la varianza de ε , representa también la varianza de los valores de **y** respecto de la recta de regresión.
- Las desviaciones de los valores de **y** respecto de la recta de regresión estimada se denominan **residuos o residuales**.
- La suma de los cuadrados de los residuales, es una medida de la variabilidad de las observaciones reales respecto de la línea de regresión estimada: **SCE Suma de cuadrados residuales o error**.

- Si se divide **SCE** por los grados de libertad que en este caso es $(n-2)$ ya que a partir de la muestra se obtienen 2 estimadores puntuales de los parámetros, se obtiene una estimación puntual sin sesgo de la varianza de regresión σ^2 que es desconocida:


$$\hat{y}_i = a + b x_i$$

$$SCE = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - a - bx_i)^2$$

- El error estándar de estimación es la raíz de $S^2_{y/x}$

$$S_{y/x}^2 = ECM = \frac{SCE}{n - 2}$$

Análisis de la tabla de Varianza de regresión:

- Generalmente del análisis hecho con los softwares se obtiene una tabla que resume el análisis de cuadrados que permite obtener la varianza , los residuos y que sirve también para otros análisis de bondad del modelo.

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrado medio
Regresión	SCR	1	$CMR = \frac{SCR}{1}$
Error	SCE	$n - 2$	$ECM = \frac{SCE}{n - 2}$
Total	STC	$n - 1$	

- **POR EJEMPLO:**

ANÁLISIS DE VARIANZA

	Grados de libertad	Suma de cuadrados	Promedio de los cuadrados
Regresión	1	1701032,819	1701032,819
Residuos	21	1493449,89	71116,66145
Total	22	3194482,71	

Varianza estimada de la regresión

Interpretación


- Regresión, error y total son las etiquetas de las tres fuentes de variación, y **SCR**, **SCE** y **STC** son las sumas de cuadrados correspondientes que aparecen en la columna 2.
- En la columna 3 se indican los grados de libertad 1 para SCR ya que en este caso la variable independiente es 1, $n - 2$ para SCE y $n - 1$ para STC.
- ECM es el cuadrado medio debido al error y es lo que se calculó como varianza de la regresión.

Uso de la ecuación de regresión para estimación y predicción

- Si existe una relación significativa entre x e y , y se determina que la ecuación de regresión estimada es adecuada entonces es útil para usarla para estimación y predicción.



Análisis de correlación

- 
- **Brinda medidas que dicen cuan fuerte o importante es la asociación encontrada entre las variables**

- 
- **Se analizan dos coeficientes:**
 - Correlación**
 - Determinación**

Coeficiente de Correlación

1 - Las variables X e Y son variables aleatorias, esto significa que no es fijo decir variable dependiente o independiente, cualquiera de las dos puede ser la variable independiente o a la inversa.

2 - Las variables proceden de una población Normal bivariada, o sea X e Y están distribuidas conjuntamente como normal.

3 - X e Y tienen cada una distribución Normal

4 - La relación entre X e Y es lineal ; este supuesto implica decir que las medias de Y para valores de X caen sobre la recta $Y_i = a + \beta X_i$, de la misma manera que para $X_i = a + \beta Y_i$

5 - Si las dos rectas de regresión (con X dependiente o con Y dependiente) son iguales, quiere decir que la relación es perfecta.

Coeficiente de Correlación poblacional:

$$\rho = \frac{E[(X - \mu_x)(Y - \mu_y)]}{\sqrt{E(X - \mu_x)^2 E(Y - \mu_y)^2}}$$

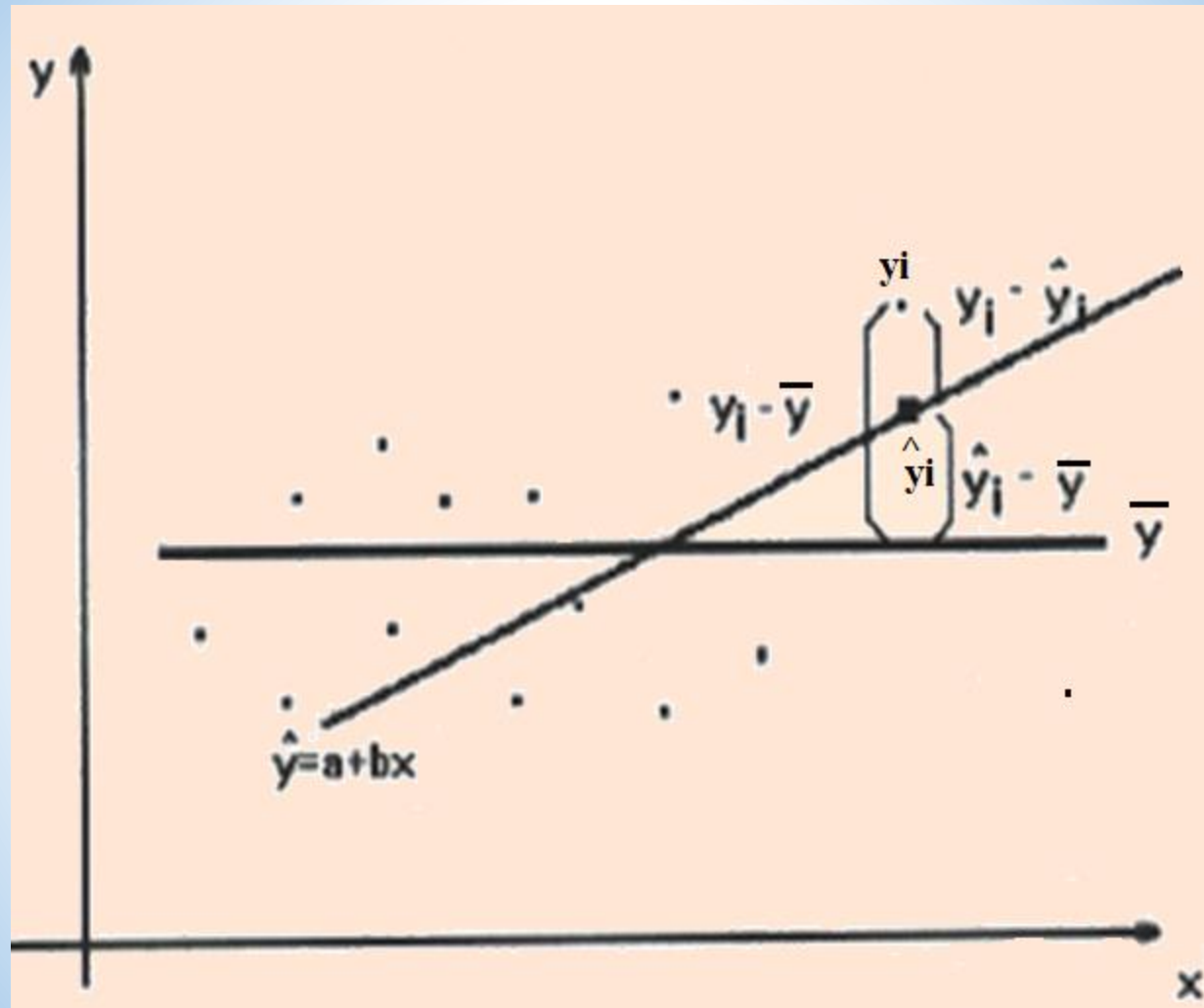
$$\rho = \frac{\text{COV}}{\sigma_x \sigma_y}$$


Coeficiente de correlación muestral

$$r = \hat{\rho} = \frac{\sum_i (x_i - \bar{x}) (y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$
$$= \frac{m_{1,1}}{S_x S_y}$$

Coeficiente de determinación

- Se parte del mismo análisis realizado para la Varianza de regresión, la partición de cuadrados:




$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2$$

$$\sum_i (y_i - \bar{y})^2 = (SCTotales)$$

$$\sum_i (\hat{y}_i - \bar{y})^2 = (SCR \text{ debidos a la regresión})$$


$$\sum_i (y_i - \hat{y}_i)^2 = (SCError)$$


$$SCE = SCT - SCR$$

$$\frac{SCE}{SCT} = \frac{SCT}{SCT} - \frac{SCR}{SCT}$$

$$1 = \frac{SCR}{SCT} + \frac{SCE}{SCT}$$


$$r^2 = 1 - \frac{SCE}{SCT} = \frac{SCR}{SCT}$$



r^2 varía entre 0 y 1, ya que SCR es menor o igual que SCT.

Algunos comentarios:

Si $SCE = 0$, implica que $SCR = SCT$, luego r^2 es igual a 1. Esto significa que todos los puntos están sobre la recta estimada.



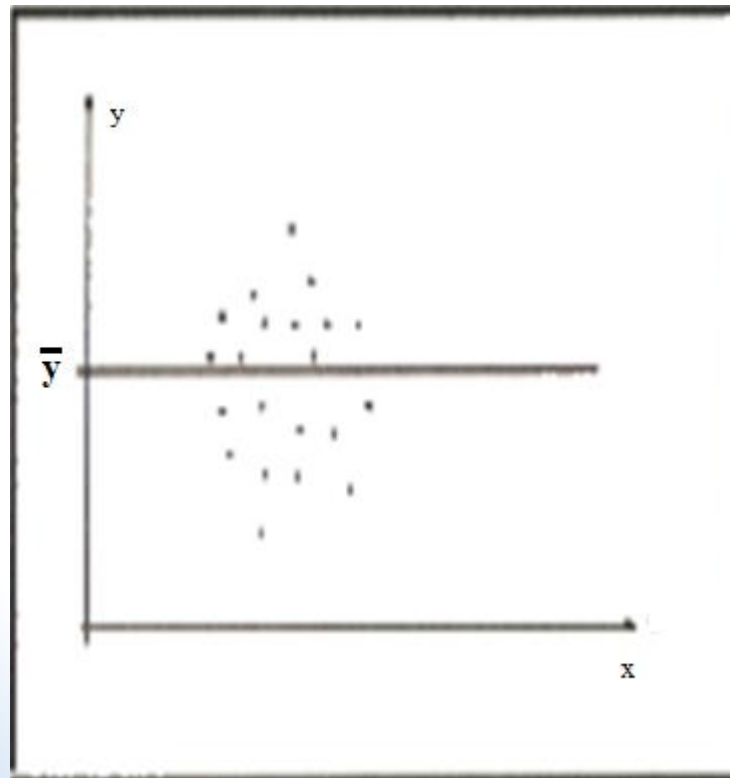
Si $SCR = 0$, implica que $SCE = SCT$, con lo cual

$$r^2 = 0$$

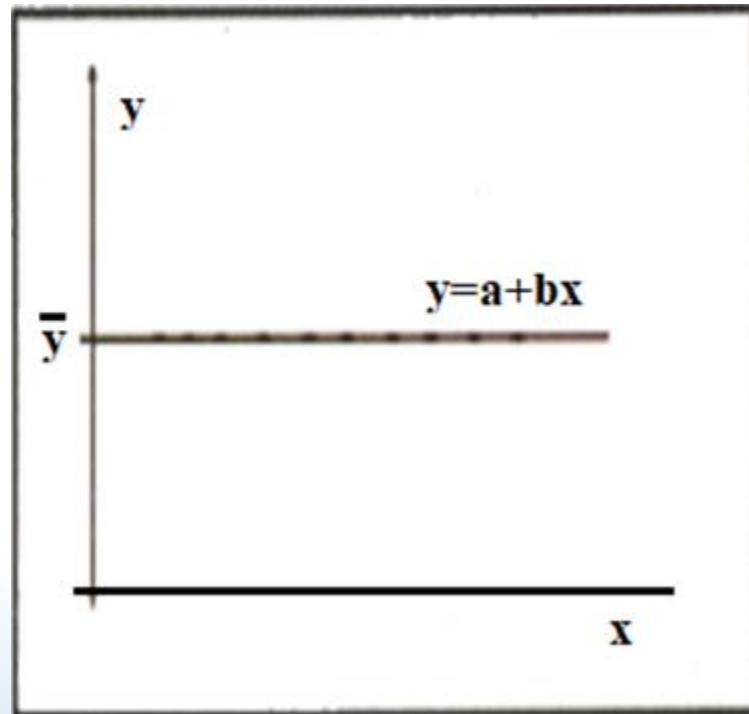
Esto significaría que la pendiente de la recta es igual a cero. Esto puede deberse a que la línea de regresión sea horizontal.

Esto puede ser relacionado a distintas causas:

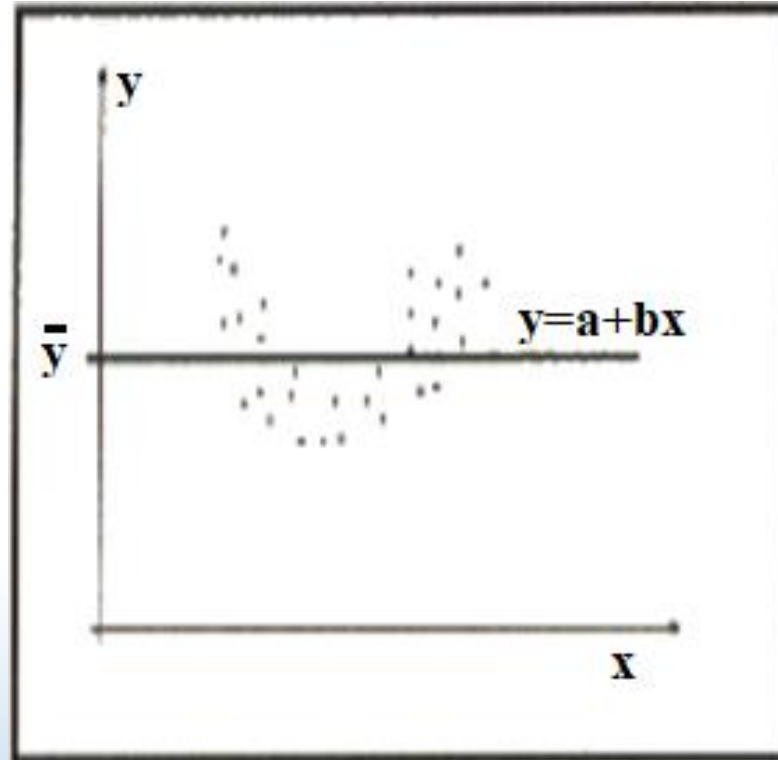
las observaciones se dispersan alrededor del valor medio en forma aleatoria.



- todas las observaciones tienen el mismo valor, cualquiera sea el valor de x




- las observaciones se dispersan alrededor de una curva tal que la línea mejor ajustada es una línea recta horizontal

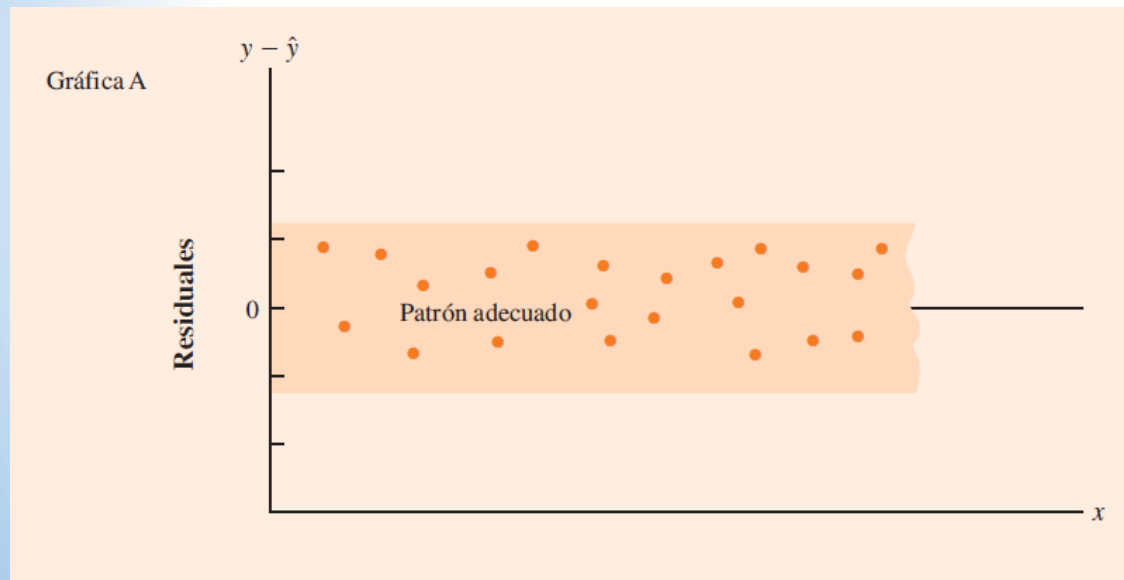


Análisis de residuos: confirmación de los supuestos del modelo

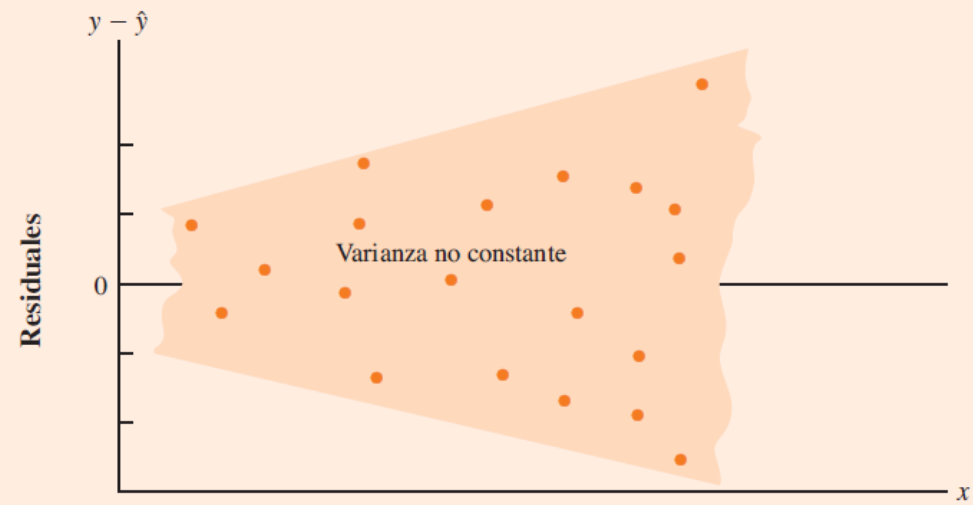
- Otro análisis interesante y que ayuda a confirmar si el modelo es adecuado es el ***análisis de residuos***.
- Como ya se indicó, el *residuo* de la observación i es la diferencia entre el valor observado de la variable dependiente y_i y el valor estimado de ella usando el modelo de regresión \hat{y}_i

- 
- Se plantearon al comienzo los siguientes supuestos para el término del error ε :
 - **1.** $E(\varepsilon) = 0$
 - **2.** La varianza de ε , σ^2 , es la misma para todos los valores de x .
 - **3.** Los valores de ε son independientes.
 - **4.** El término del error ε tiene una distribución Normal.

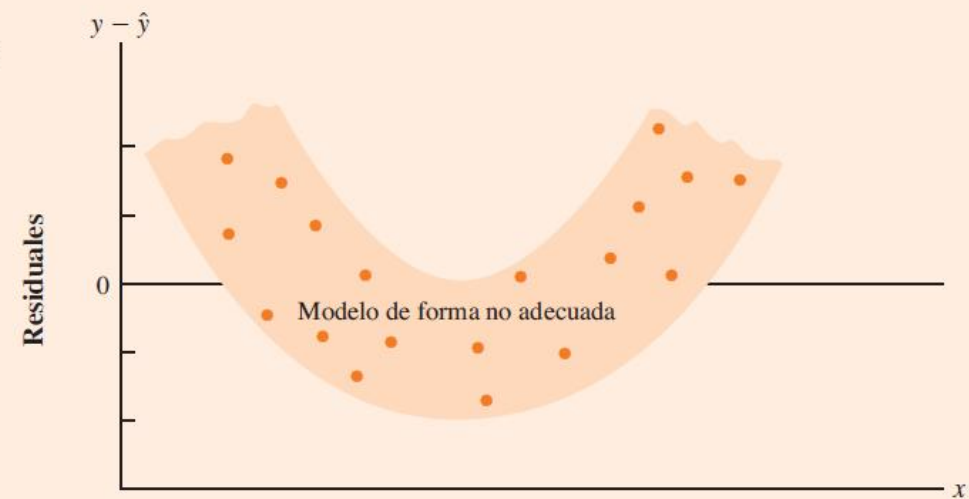
Se puede hacer el análisis de los residuos para saber si se cumplen esos supuestos:



Gráfica B



Gráfica C



distribución de probabilidad normal de los residuos

- Para determinar la validez del supuesto de que el término del error tiene una distribución normal puede hacerse un histograma de los residuos, un diagrama box plot, o bien verificar las características del modelo Normal de estos residuos: coincidencia de los valores de la media , mediana y moda, y el valor de asimetría aproximadamente igual a cero y la kurtosis próxima a tres.