Estimación e Interpolación

Esperanza Ayuga Téllez
(2008)

Imagen de la NASA

Una aplicación importante de los SIG es la **estimación** de los valores que puede tomar la variable Z sobre el terreno.

Las fuentes de datos son las observaciones o mediciones de Z obtenidas en puntos dispersos distribuidos sobre el terreno (red de muestreo).

A partir de los datos se estiman los valores de Z en puntos intermedios mediante interpolación o métodos geoestadísticos

Contenidos

Estimación puntual de Z con los datos de la red de muestreo.

Estimación de valores intermedios mediante métodos globales de interpolación.

- Análisis de la varianza (ANOVA)
- Métodos de regresión simple.
- Métodos de regresión polinomial.
- Métodos de regresión múltiple.

Contenidos

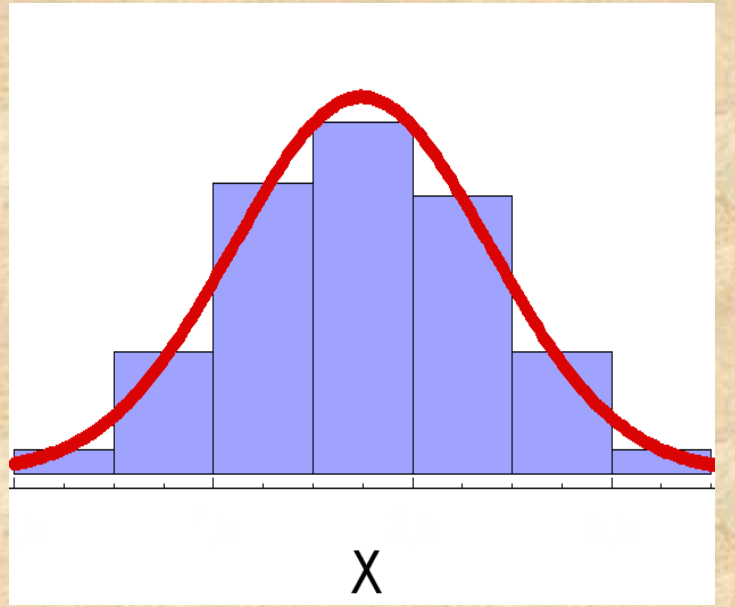
Estimación de valores intermedios mediante métodos locales de interpolación.

- Polígonos de Voronoi
- Ponderación de la inversa de la distancia (IDW).
- Funciones no lineales o funciones de base radial (FBR).

ESTIMACIÓN PUNTUAL

Una variable aleatoria puede tomar diferentes valores y cada valor o conjunto de valores tiene asociada una probabilidad de ocurrencia.

Si la variable toma valores discretos la probabilidad es una **función de masa** y si es continua se obtiene una función de densidad.



Función de masa de la distribución Binomial.
Función de densidad de la distribución Normal.

ESTIMACIÓN PUNTUAL

Si X tiene función de masa o densidad $f(x,\theta)$, donde θ es un parámetro desconocido, se llama estimador de θ a un **estadístico** que puede alcanzar valores próximos al valor desconocido de θ.

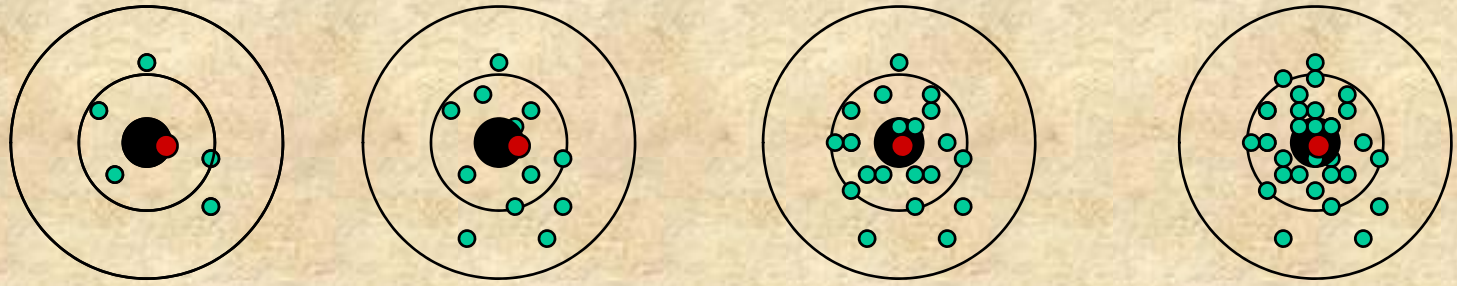
Una vez obtenida la m.a.s. $x_1, x_2, ..., x_n$, el valor numérico que se asigna al parámetro recibe el nombre de **estimación de** θ y se calcula con la muestra

$$\widehat{\theta} = g(x_1, x_2, \ldots, x_n)$$

ESTIMACIÓN PUNTUAL

Los estimadores deben tener buenas propiedades.

Consistencia: se aproxima al valor real del parámetro desconocido a medida que aumenta n.



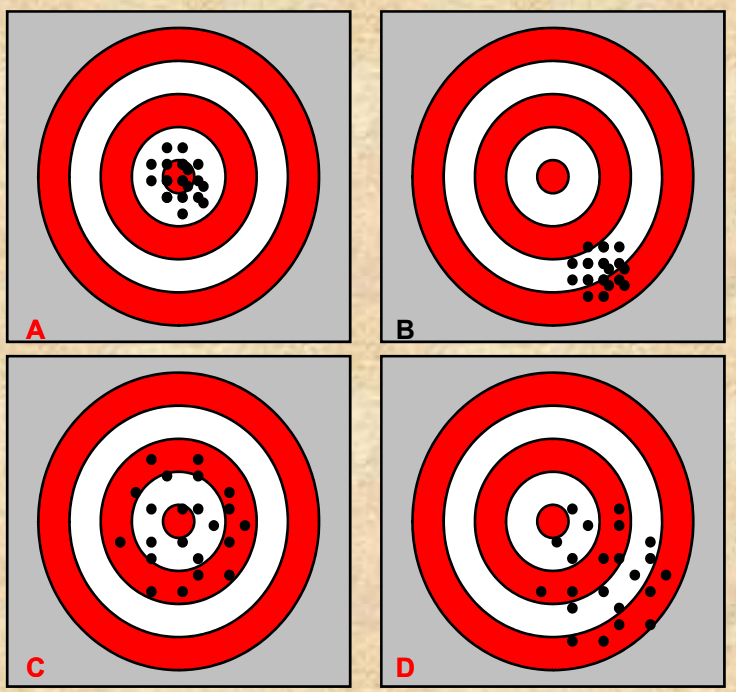
n = 5 n = 12 n = 21 n = 34

ESTIMACIÓN PUNTUAL

Insesgado o centrado: Su media coincide con θ.

Si es centrado tomará valores alrededor del verdadero valor. Si es sesgado, se desviará sistemáticamente de θ.

Eficiente: De todos los estimadores de θ el más eficiente será el que tenga **menor varianza**.



A: Estimador centrado y eficiente;
B: Estimador sesgado y eficiente
C: Estimador centrado e ineficiente;
D: Estimador sesgado e ineficiente

ESTIMACIÓN PUNTUAL

Error cuadrático medio: valor esperado del cuadrado de la diferencia entre el estimador $\hat{\theta}$ y el parámetro θ que trata de estimar.

$$ECM(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = V(\hat{\theta}) + [\theta - E(\hat{\theta})]^2$$

De lo anterior se concluye que el ECM está compuesto por dos cantidades no negativas, que son:

La varianza del estimador θ.

El cuadrado del sesgo del estimador.

ESTIMACIÓN PUNTUAL

Obtención de estimadores:

Mediante el método de los **mínimos cuadrados**.

El estimador es el valor que minimiza la suma de las diferencias al cuadrado entre los datos y el parámetro.

NOTA: En enero de 1801 el asteroide Ceres desapareció de la vista. Carl Friedrich Gauss, mediante el método de los mínimos cuadrados, consiguió calcular su órbita y en diciembre de ese mismo año el asteroide fue redescubierto muy cerca de la posición predicha.

Reverso del premio Gauss

ESTIMACIÓN PUNTUAL

- Los estimadores son valores obtenidos con la muestra y próximos a los parámetros de la población.
- Un buen estimador DEBE SER centrado, eficiente, consistente y con mínimo error cuadrático medio...
- Se escoge el estimador con mejores propiedades.
- Métodos de obtención de estimadores: Método de los mínimos cuadrados

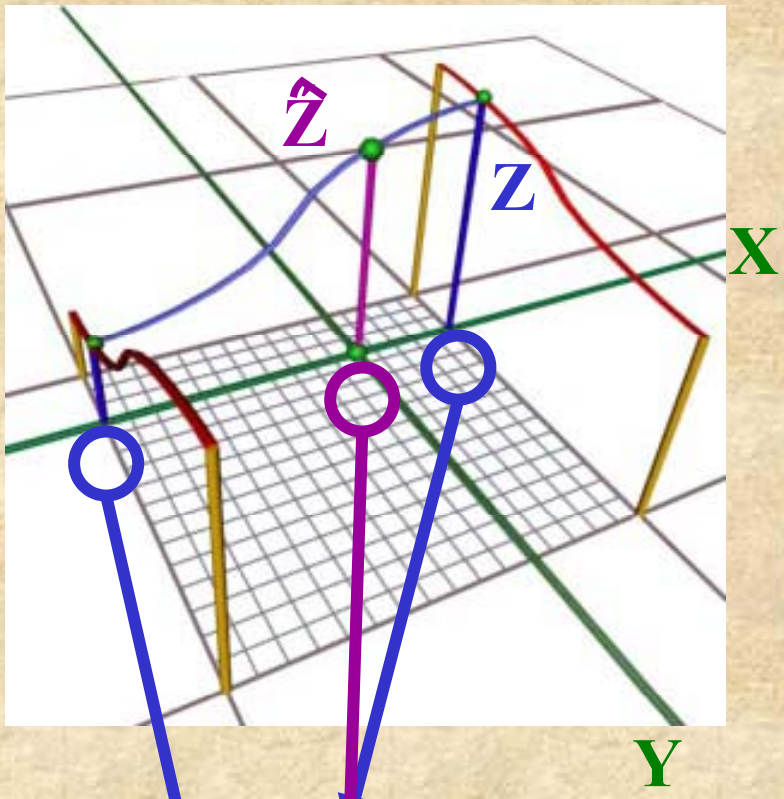
ESTIMACIÓN PUNTUAL

No olvidemos que un buen estimador no garantiza una buena estimación.

De una muestra no representativa se obtiene una mala estimación.

MÉTODOS DE INTERPOLACIÓN GLOBALES

El proceso de interpolación espacial consiste en la estimación de los valores que alcanza una variable $\hat{Z}$ en un conjunto de puntos definidos por un par de coordenadas (**X**,**Y**), partiendo de los valores de **Z** medidos en una muestra de puntos situados en el mismo área de estudio.



Adaptado de **pgrafica.webideas4all.com/perlin.html**

MÉTODOS DE INTERPOLACIÓN GLOBALES

Cuando se trabaja con un SIG la interpolación espacial suele utilizarse para obtener capas raster que representan la variable a interpolar. En esos casos cada celdilla de la capa raster constituye un punto en el que hay que realizar la interpolación.

MÉTODOS DE INTERPOLACIÓN GLOBALES

Los estimadores globales utilizan todos los datos muestrales para proporcionar predicciones en todo el área de interés.

Se asume la dependencia de la variable a interpolar de otras variables de apoyo.

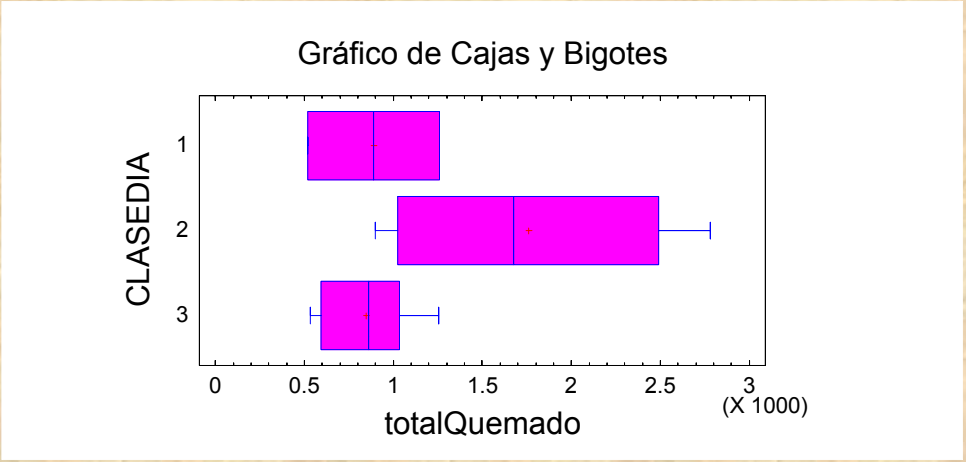
MÉTODOS DE INTERPOLACIÓN GLOBALES

Se usan para examinar y evaluar los efectos en las variaciones globales de las tendencias o clasificar grupos de terrenos distintos que pueden tener diferencias en las medias

PREDICCIÓN GLOBAL MEDIANTE CLASIFICACIÓN

La variable de apoyo es cualitativa (usos del suelo, tipos de suelo o roca, etc).

Se asume que la variable adopta, en cada punto, el valor medio correspondiente al valor de la variable de apoyo en ese punto.



PREDICCIÓN GLOBAL MEDIANTE CLASIFICACIÓN

1. Las variaciones de Z dentro de las diferentes clases de V (j) sean aleatorias y no autocorrelacionadas espacialmente.

2. $Z \sim N(\mu_j, \sigma)$ μ_j la misma en todas las j y σ igual en todas y cada una de las j.

3. Los cambios de Z en las fronteras entre clases se producen de forma brusca.

PREDICCIÓN GLOBAL MEDIANTE CLASIFICACIÓN

El resultado es equivalente a una reclasificación que produce un mapa en el que los diferentes valores de V se transforman en valores de Z según el modelo lineal:

$$Z(x_j) = \mu + \alpha_i + \varepsilon_j$$

Donde Z es el valor del atributo en la localización x_j,

μ es la media general sobre el dominio de interés,

α_i es la diferencia entre la media general y la media de la clase i a la que pertenece la localización x_j

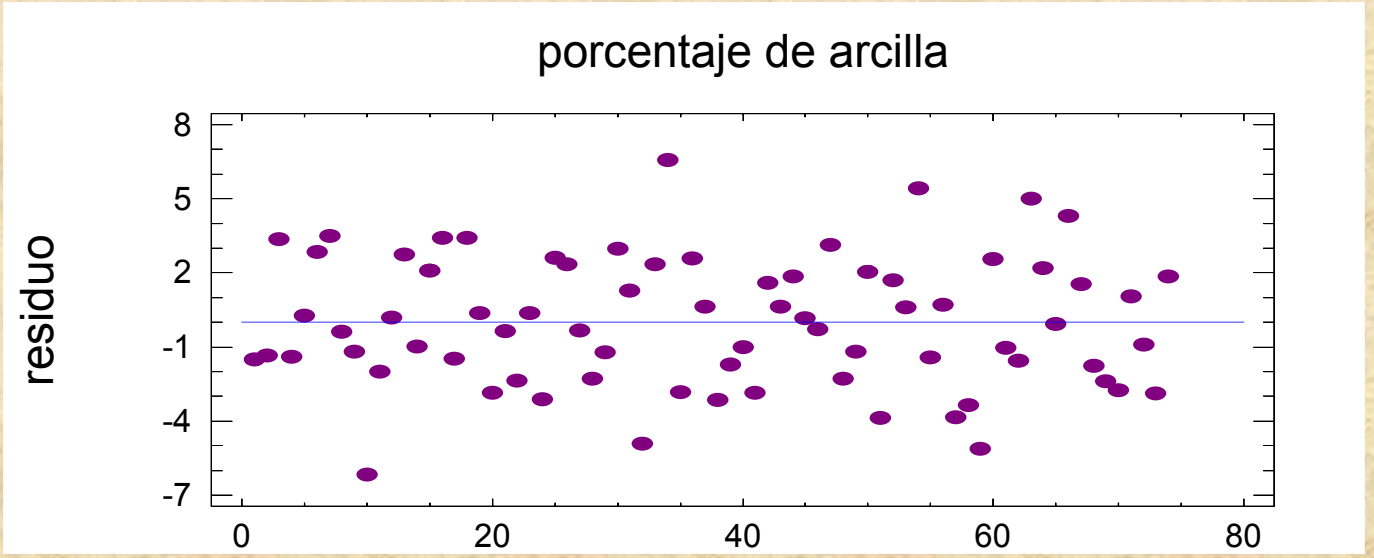
y ε_j es el residuo o perturbación del modelo.

PREDICCIÓN GLOBAL MEDIANTE CLASIFICACIÓN

Las hipótesis sobre los residuos son:

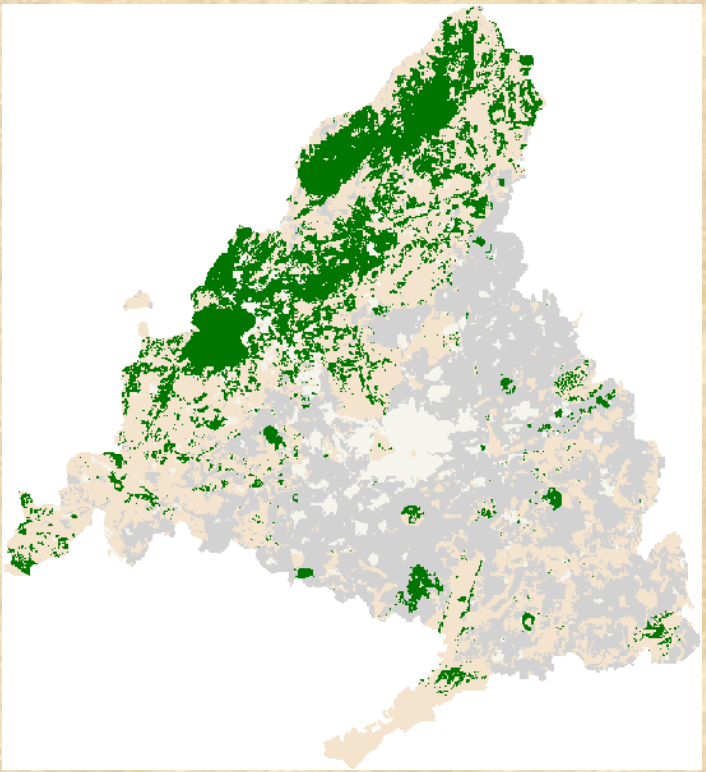
1. $\varepsilon_j \sim N(0, \sigma^2)$, con σ^2 cte. e independiente de x.

2. ε_j también indep. entre sí, es decir, $E[\varepsilon_i \varepsilon_j]=0$ con $i \neq j$



PREDICCIÓN GLOBAL MEDIANTE CLASIFICACIÓN

Se utiliza únicamente al cartografiar el suelo y el paisaje para definir unidades homogéneas



PREDICCIÓN GLOBAL MEDIANTE REGRESIÓN LINEAL

El OBJETIVO del modelo de regresión simple es establecer una relación sencilla entre dos variables numéricas y poder predecir los valores de una, conocido el valor de la otra.

p.e. si el valor de un atributo ambiental z ha sido medido a lo largo de los puntos x_1, x_2,..., x_n en un transecto.

Si Z se incrementa linealmente con la localización x, su variación puede aproximarse con un modelo de regresión.

REGRESIÓN LINEAL

Buscamos, por tanto, encontrar una función de X muy simple (lineal) que nos permita aproximar Z mediante ella.

El modelo completo de regresión lineal simple es

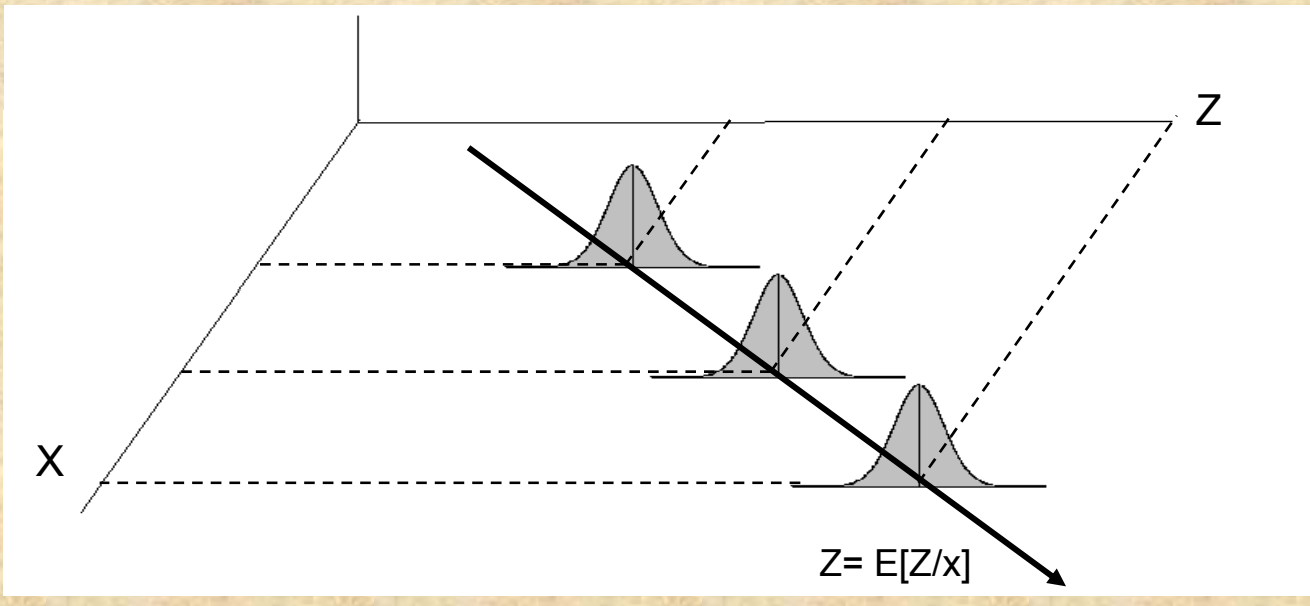
$$Z(X) = \alpha + \beta X + \varepsilon$$

α = intersección en el origen, ordenada en el origen o constante

β = pendiente de la recta

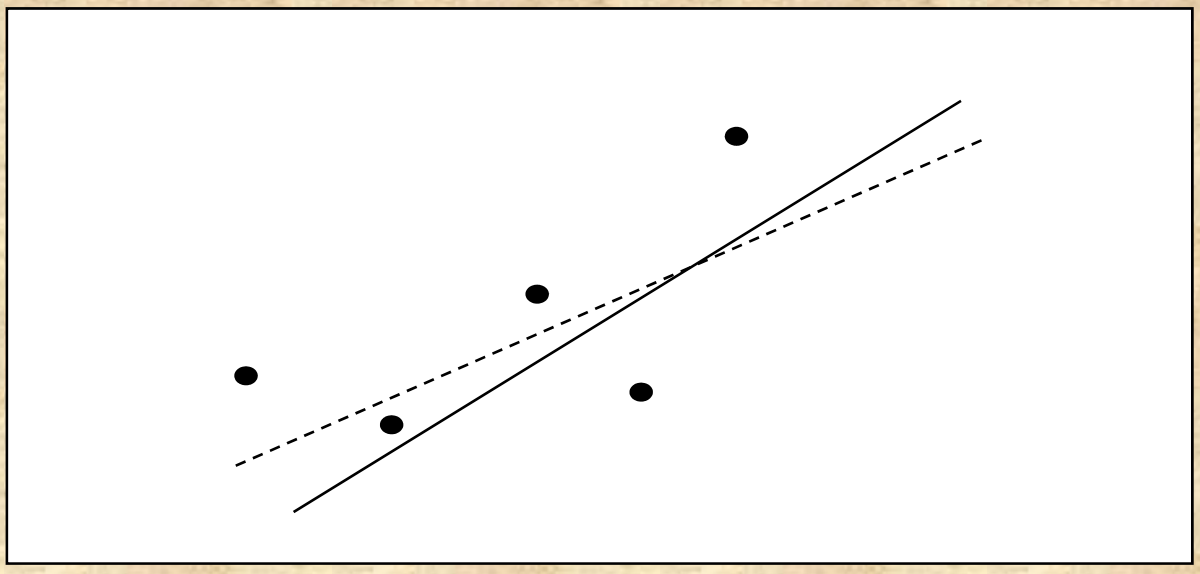
REGRESIÓN LINEAL

Para cualquier ecuación de regresión de esta forma:



Z es la v. dependiente, X es la v. independiente o explicativa y ε el término del error que representa la varianza de y para un valor dado de X.

REGRESIÓN LINEAL



El error se presenta casi siempre ya que rara vez coincidirán, por muy bueno que sea el modelo de regresión, el verdadero valor de Z y el obtenido con el modelo .

REGRESIÓN LINEAL

Para estimar el modelo recurrimos a una m.a. de datos:
En $x_1, x_2, ..., x_n$ se mide $y_1 = Z(x_1), ..., Z(x_n)$ y obtenemos la expresión:

$$\mathbf{y_i = \alpha + \beta\, x_i + \varepsilon_i}$$

las hipótesis del modelo son:

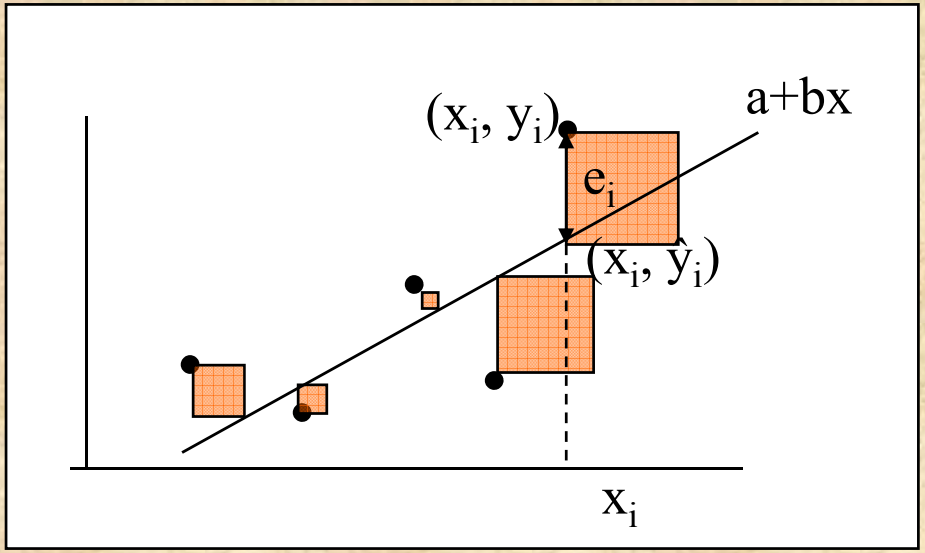
1. $\varepsilon_i \sim N(0, \sigma^2)$, con σ^2 cte. e independiente de x.

2. Las perturbaciones también son indep. entre sí, es decir, $E[\varepsilon i\ \varepsilon j]=0$ con $i \neq j$

3. Para que la relación sea lineal $\beta \neq 0$.

REGRESIÓN LINEAL

Los parámetros del modelo que debemos estimar con la muestra son tres: α, β y σ.



El método de estimación que se emplea es el de los mínimos cuadrados: Se buscan los valores que minimicen la cantidad:

$$E[(Y-f(X))^2] = E[(Y-(\alpha+\beta X))^2] = E\left[\varepsilon^2\right]$$

REGRESIÓN LINEAL

$$b = \hat{\beta} = \frac{Cov(X,Y)}{V(X)}$$

$$a = \hat{\alpha} = E[Y] - \frac{Cov(X,Y)}{V(X)} E(X)$$

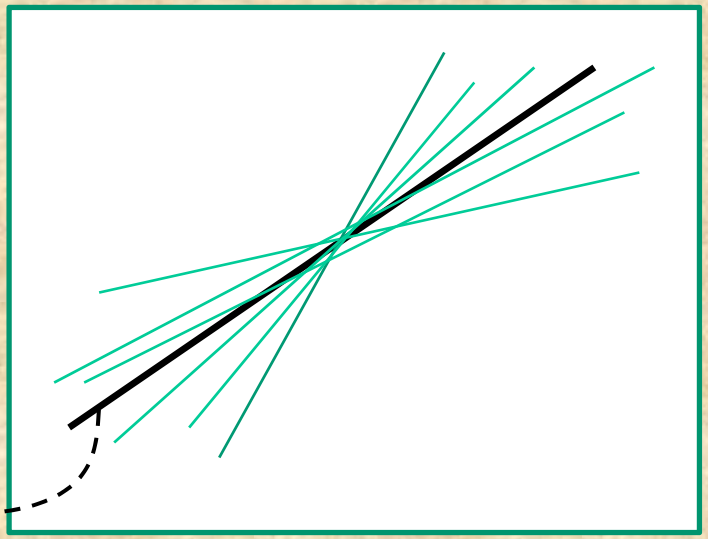
$$s_R = \hat{\sigma} = \sqrt{\frac{\sum_i e_i^2}{n-2}}$$

Los errores, residuos o perturbaciones se estiman con el modelo y las observaciones:

$$e_i = \hat{\varepsilon}_i = y_i - (a + b\, x_i)$$

REGRESIÓN LINEAL

Como siempre, muestras diferentes proporcionan conjuntos de datos diferentes, lo cual genera rectas de regresión diferentes. Estas rectas se distribuyen alrededor de $Y = \alpha + \beta X + \varepsilon$.



Entonces la pregunta es ¿cómo se distribuyen a y b alrededor de α y β, respectivamente? Y ¿cómo construiremos los intervalos de confianza y el contraste de hipótesis?

Covarianza de dos variables X e Y

- La covarianza entre dos variables, $\mathbf{S_{xy}}$, nos indica si la posible relación entre dos variables es directa o inversa.
  - Directa: $S_{xy} > 0$
  - Inversa: $S_{xy} < 0$
  - Incorreladas: $S_{xy} = 0$

$$S_{xy} = \frac{1}{n}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

- El signo de la covarianza nos dice si el aspecto de la nube de puntos es creciente o no, pero no nos dice nada sobre el grado de relación entre las variables.

Correlación lineal de Pearson

- **La coeficiente de correlación lineal de Pearson de dos variables, r, nos indica si los puntos tienen una tendencia a disponerse en línea (excluyendo rectas horizontales y verticales).**

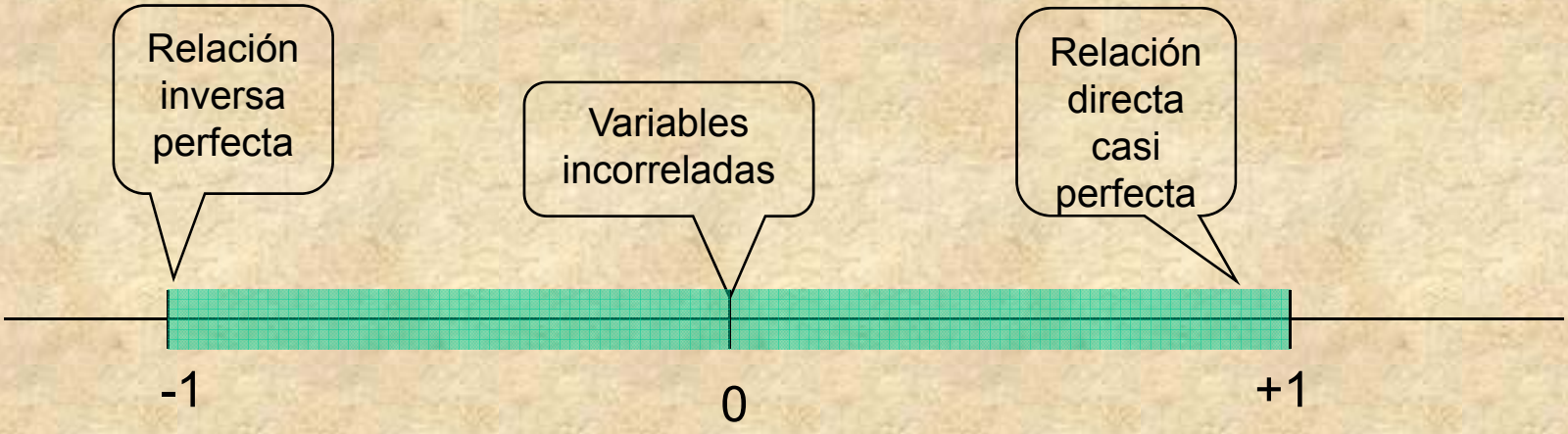
- **tiene el mismo signo que S_{xy} por tanto su signo indica si la posible relación es directa o inversa.**

$$r = \frac{S_{xy}}{S_x S_y}$$

- **r es útil para determinar si hay relación lineal entre dos variables, pero no servirá para otro tipo de relaciones (cuadrática, logarítmica,...)**

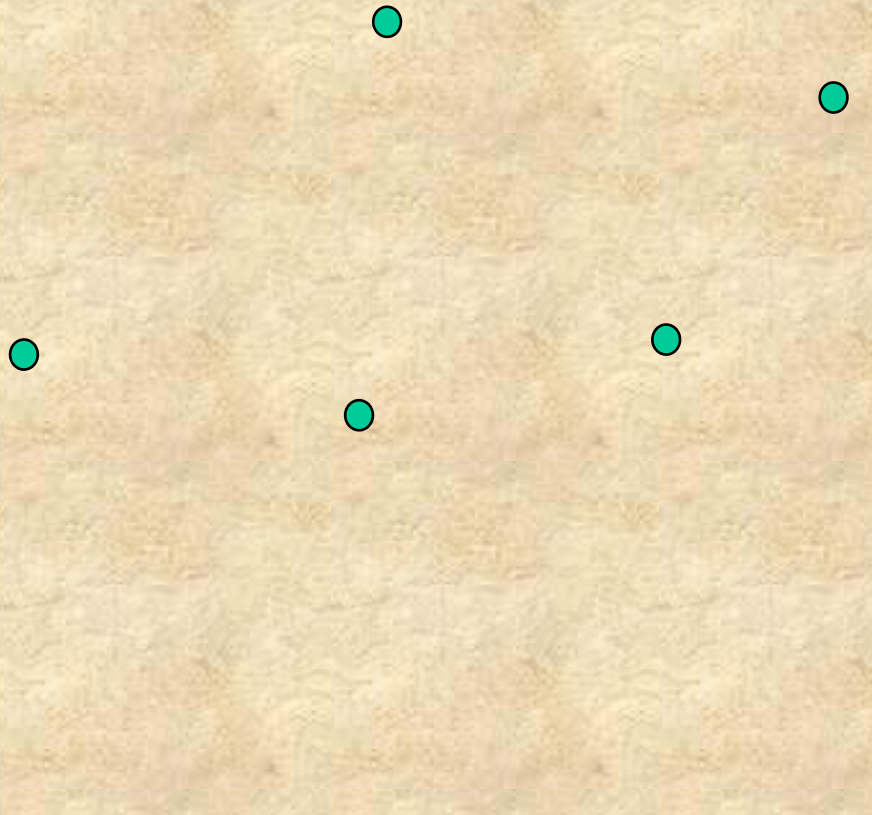
Propiedades de r

- Es adimensional
- Sólo toma valores en [-1,1]
- Las variables son incorreladas $\Rightarrow$ r =0
- Relación lineal perfecta entre dos variables $\Leftrightarrow$ r =+1 o r =-1
  - Excluimos los casos de puntos alineados horiz. o verticalmente.
- Cuanto más cerca esté r de +1 o -1 mejor será el grado de relación lineal.
  - Siempre que no existan observaciones anómalas.



¿Cómo medir la bondad de la regresión?

Imaginemos un diagrama de dispersión, y vamos a tratar de comprender en primer lugar qué es el error residual, su relación con la varianza de Y, y de ahí, cómo medir la bondad de un ajuste.

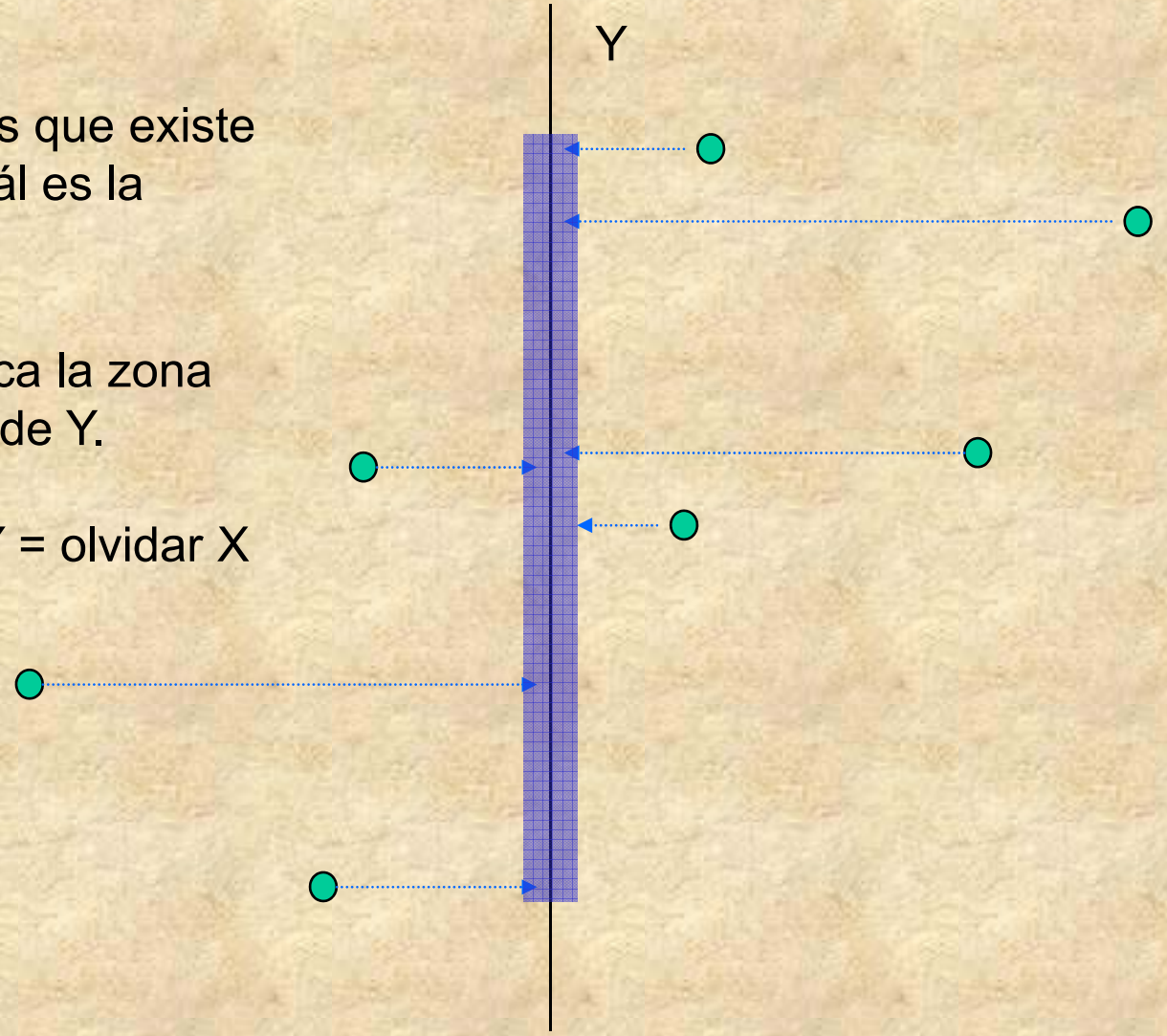


Interpretación de la variabilidad en Y

En primer lugar olvidemos que existe la variable X. Veamos cuál es la variabilidad en el eje Y.

La franja sombreada indica la zona donde varían los valores de Y.

Proyección sobre el eje Y = olvidar X

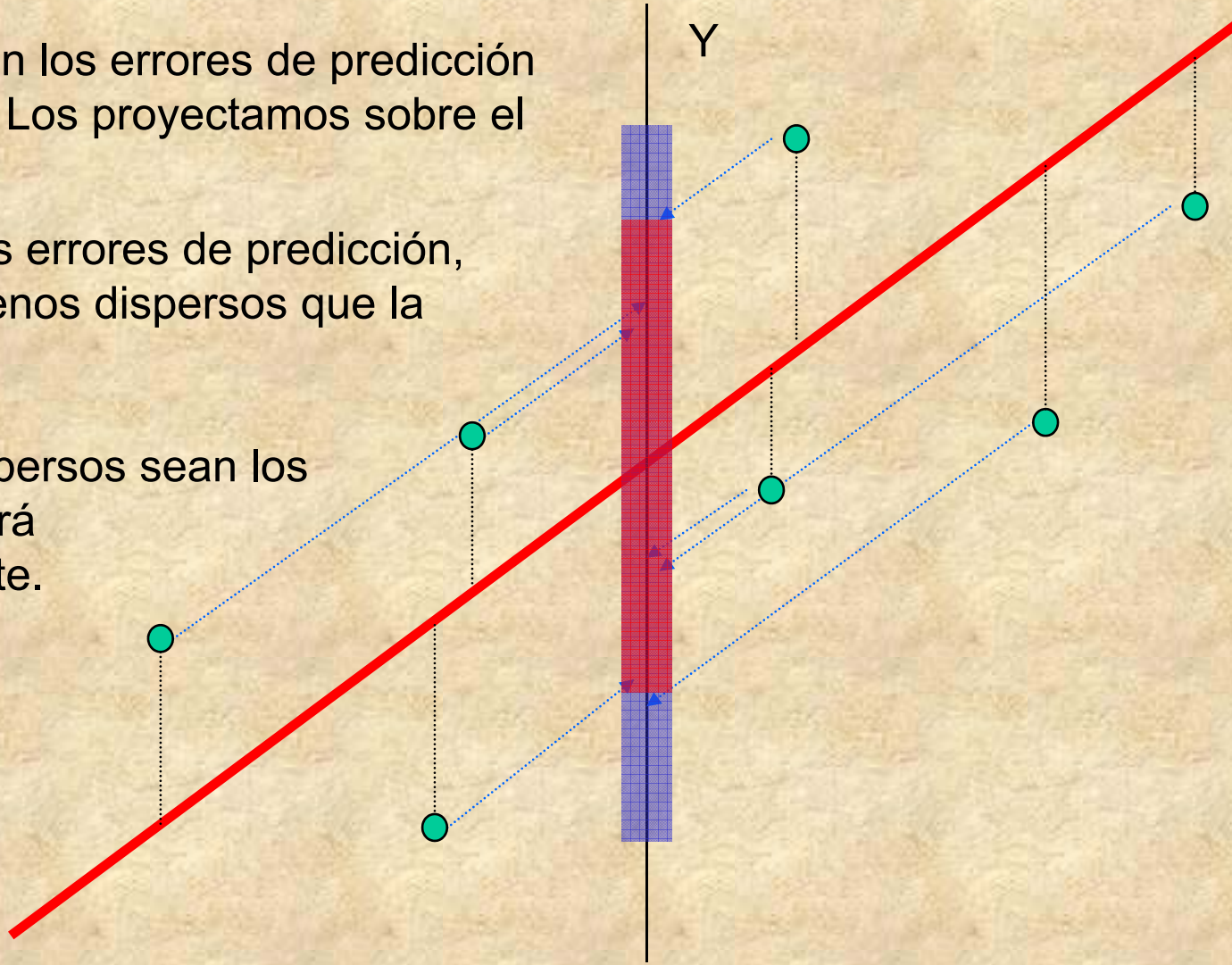


Interpretación del residuo

Fijémonos ahora en los errores de predicción (líneas verticales). Los proyectamos sobre el eje Y.

Se observa que los errores de predicción, residuos, están menos dispersos que la variable Y original.

Cuanto menos dispersos sean los residuos, mejor será la bondad del ajuste.



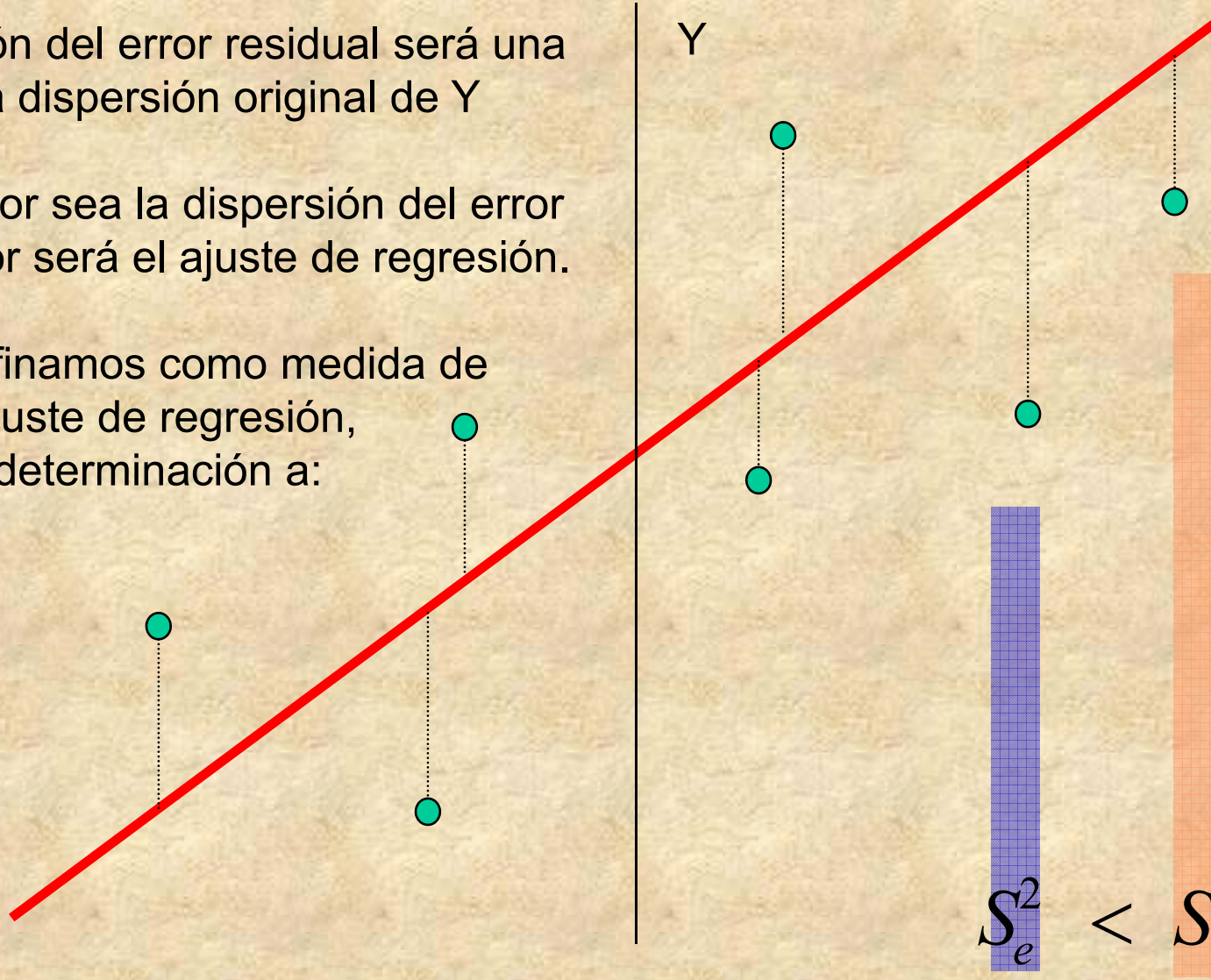
Bondad de un ajuste

• La dispersión del error residual será una fracción de la dispersión original de Y

•Cuanto menor sea la dispersión del error residual mejor será el ajuste de regresión.

Eso hace que definamos como medida de bondad de un ajuste de regresión, o coeficiente de determinación a:

$$R^2 = 1 - \frac{S_e^2}{S_Y^2}$$



Resumen sobre bondad de un ajuste

- La bondad de un ajuste de un modelo de regresión se mide usando el coeficiente de determinación R^2
- R^2 es una cantidad adimensional que sólo puede tomar valores en [0, 1]
- Cuando un ajuste es bueno, R^2 será cercano a 1. Cuando un ajuste es malo R^2 será cercano a 0.
- A R^2 también se le denomina porcentaje de variabilidad explicado por el modelo.
- En el modelo lineal simple, la expresión es de lo más sencilla: $\mathbf{R^2=r^2}$

MODELOS POLINÓMICOS

La importancia de este tipo de modelos es que pueden representar localmente cualquier relación no lineal.

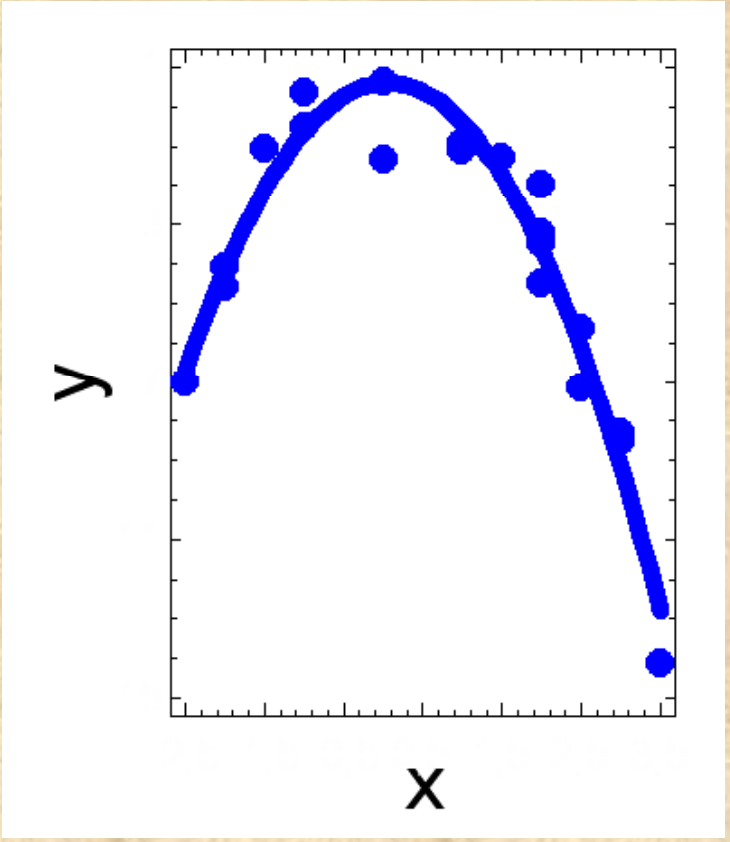
Esto se debe a la conocida expresión matemática del desarrollo de Taylor.

Aumentando adecuadamente el grado del polinomio, en teoría, se puede conseguir una muy buena aproximación.

REGRESIÓN NO LINEAL

A veces los datos dibujan una curva no lineal y se utilizan varios trucos para usar técnicas de regresión lineal para problemas no lineales. Lo más fácil es escribir Y como una polinomial.

Tratando a x y a x^2 como variables independientes en un modelo lineal.



$$Y = \beta_0 + \beta_1 X + \beta_2 X^2$$

MODELOS POLINÓMICOS

Por ejemplo un modelo con dos variables y grado del polinomio 2 seguirá la expresión:

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

Un modelo con una variable y grado k será:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k$$

Los modelos de 2 variables (X e Y) son los más utilizados para obtener por interpolación valores de superficies continuas.

REGRESIÓN MÚLTIPLE

El objetivo de este modelo es establecer una relación estadística sencilla entre un grupo de variables independientes y una variable dependiente Z.

p.e. Permite predecir la temperatura, la precipitación y la humedad mediante el día, mes, altitud, latitud y longitud de una localidad.

DEFINICIÓN:

Es la extensión del modelo lineal simple a k variables explicativas. Se usan aquellas que miden el efecto más importante y se agregan las restantes en el efecto aleatorio (ε_i)

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

HIPÓTESIS BÁSICAS:

- sobre las perturbaciones $\varepsilon_i \approx N(0, \sigma)$ y $E[\varepsilon_i \varepsilon_j] = 0$
- además $n \geq k+1$ y
- las x_i son linealmente **independientes**

MÉTODOS DE INTERPOLACIÓN LOCALES

Los **métodos locales** de estimación utilizan la interpolación empleando sólo la información de los puntos más cercanos.

Los polígonos de Voronoi

Es uno de los métodos de interpolación más simples, basado en la distancia euclidiana, apropiada para datos cualitativos.

Se crean al unir los puntos entre sí, trazando las mediatrices de los segmento de unión. Las intersecciones de estas mediatrices determinan una serie de polígonos en un espacio bidimensional alrededor de los puntos de control, de manera que el perímetro de los polígonos generados sea equidistante a los puntos vecinos y designando su área de influencia.

Los polígonos de Voronoi

Una formación teselada Voronoi es una estructura de celdas en la que el interior de cada celda está compuesto por todos los puntos cercanos a un punto del entramado particular, más que a cualquier otro punto.

Las estructuras de celdas de Voronoi son polígonos de forma irregular; el número y la ubicación de las celdas pueden ajustarse para coincidir con la densidad y la ubicación de los datos espaciales.

Ponderación por distancia

Estima los puntos del modelo realizando una asignación de pesos a los datos del entorno en función inversa a la distancia que los separa del punto en cuestión. De esta forma, se acepta que los puntos más próximos al centroide intervienen de manera más relevante en la obtención del valor definitivo de Z para ese punto.

La formula general para la interpolación por IDW es :

$$Z(x,y) = \frac{\sum_{i=1}^{n} \frac{z_i}{d_i^p}}{\sum_{i=1}^{n} \frac{1}{d_i^p}}$$

donde Z : punto problema; z_i : punto del entorno; p : exponente de ponderación; d_{ij} = distancia entre los puntos

Ponderación por distancia

La elección del exponente de ponderación (p) determina la contribución de los puntos circundantes al punto problema: cuanto mayor es p, más contribuyen los puntos próximos.

Es necesario contar con muchos puntos para la interpolación, ya que de lo contrario se obtienen zonas aterrazadas en exceso.

Funciones no lineales o "spline".

También se conocen por funciones de base radial (FBR). Este algoritmo ajusta una curva suave a un conjunto de puntos conocidos,

$$C(u) = \frac{\sum_{i=0}^{i=n} w_i N_{i,k}(u) P_i}{\sum_{i=0}^{i=n} w_i N_{i,k}(u)}$$

donde,
u es un parámetro.
Ni,k es la función base de grado k.
Pi son los puntos de control.
wi son los pesos.

Funciones no lineales o "spline".

El método consiste en interpolar los valores de gris faltantes con polinomios de grado k en cada dimensión i y promediar el resultado de cada uno de ellos.

La particularidad que tienen estas funciones es que imponen continuidad y suavidad de la función interpolante (derivada primera continua) en los puntos frontera de la región que se desea interpolar.

Funciones óptimas que emplean la autocorrelación

Correlación de una variable consigo misma, cuando las observaciones son consideradas con una diferencia en el tiempo (autocorrelación temporal) o en el espacio (autocorrelación espacial).

Funciones óptimas que emplean la autocorrelación

Si la presencia de un atributo en una parte de un territorio hace su presencia en las zonas vecinas más o menos probable, existe un efecto de contigüidad en la estructura espacial, y en tal caso el fenómeno muestra una autocorrelación espacial.

El método de interpolación basado en optimizar funciones usando autocorrelación espacial es el krigeado

FUENTES

Ayuga Téllez E. (2008). Aplicaciones de Inferencia Estadística a la Ingeniería del Medio Natural. Fundación Conde del Valle de Salazar, ETSIM (UPM), Madrid.

Peña Llopis, J. (2006) Sistemas de Información Geográfica aplicados a la gestión del territorio. Editorial Club Universitario, Madrid.