Universidad de los Andes

Análisis espacial de datos y sus aplicaciones en Python

Profesor: Germán González

Sesión 7: Autocorrelación y regresión espacial

Universidad de los Andes

Índice

- Conceptos básicos e interpretación.
- Estadísticos de auto correlación espacial globales y locales.
- Regresiones lineales espaciales.
- Aplicación: Prevalencia e incidencia de enfermedades.

Índice

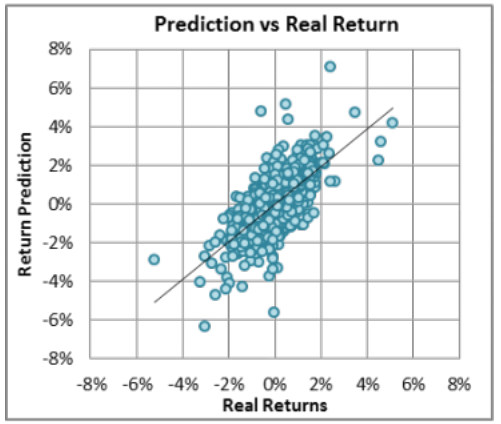
- **Conceptos básicos e interpretación.**
- **Estadísticos de auto correlación espacial globales y locales.**
- **Regresiones lineales espaciales.**
- **Aplicación: Prevalencia e incidencia de enfermedades.**

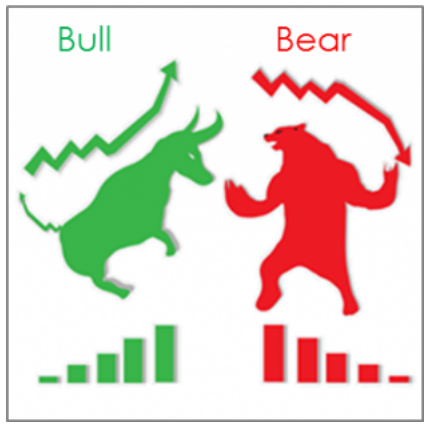
Machine Learning

Modelos supervisados:

El aprendizaje supervisado utiliza como entrenamiento un set de datos que contiene una marca o etiqueta en los datos. En este tipo de aprendizaje es claro la distinción entre las variables independientes y la variable dependiente.

Regression vs Classification





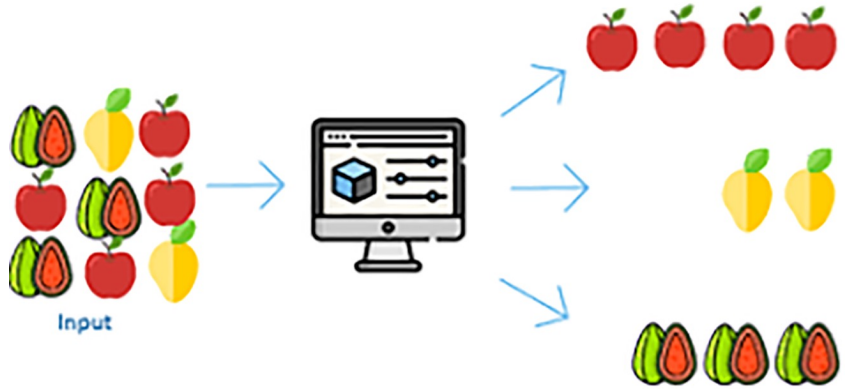
Regresión:
Estimación continua

Clasificación:
Estimación discreta

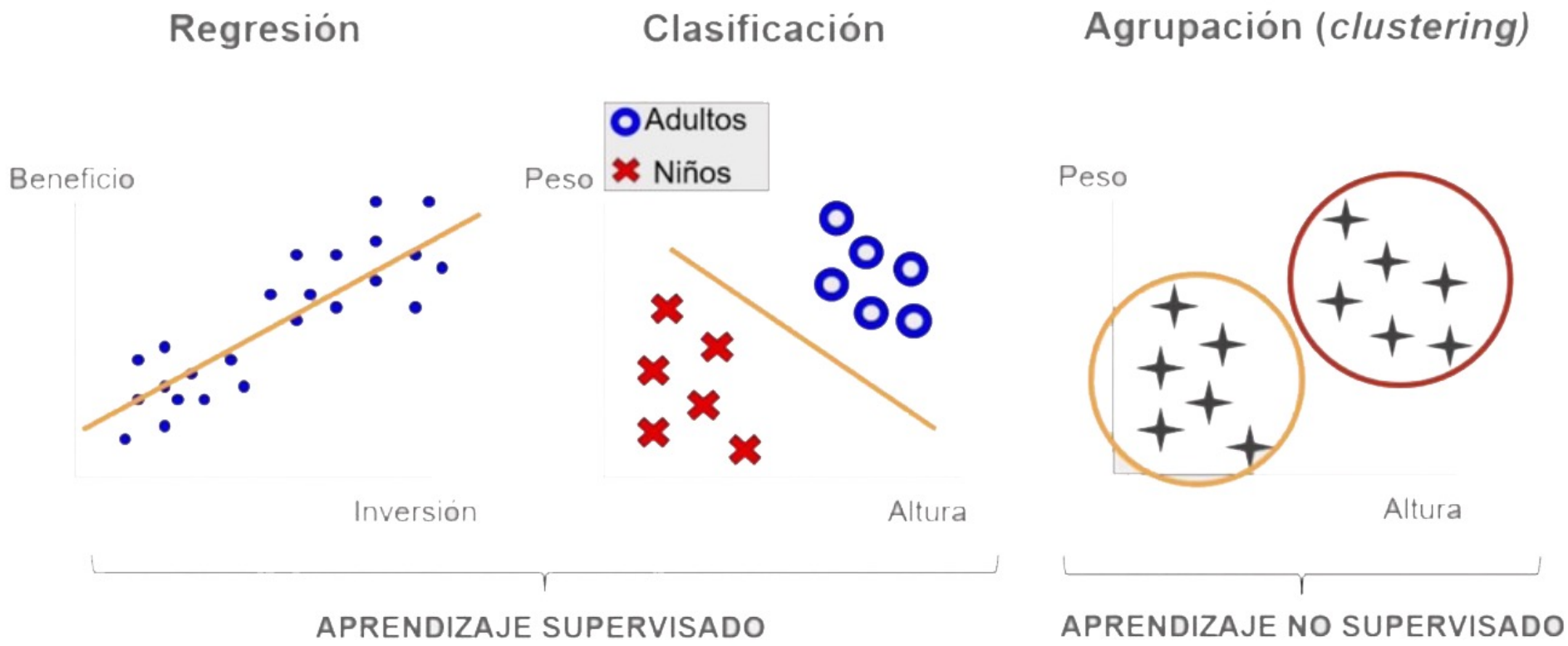
Machine Learning

Modelos no supervisados

El aprendizaje no supervisado no requiere de una etiqueta previa de los datos, utiliza toda la información para realizar asociaciones entre los datos o agrupaciones de estos.

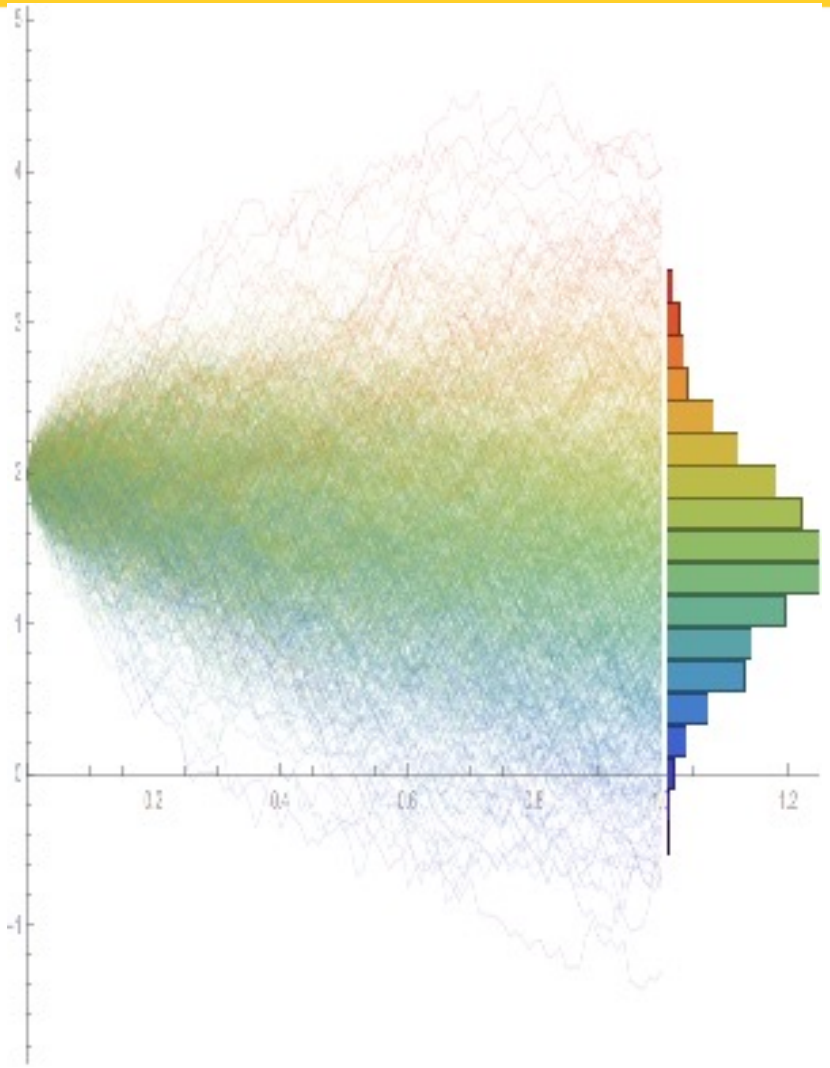


Machine Learning

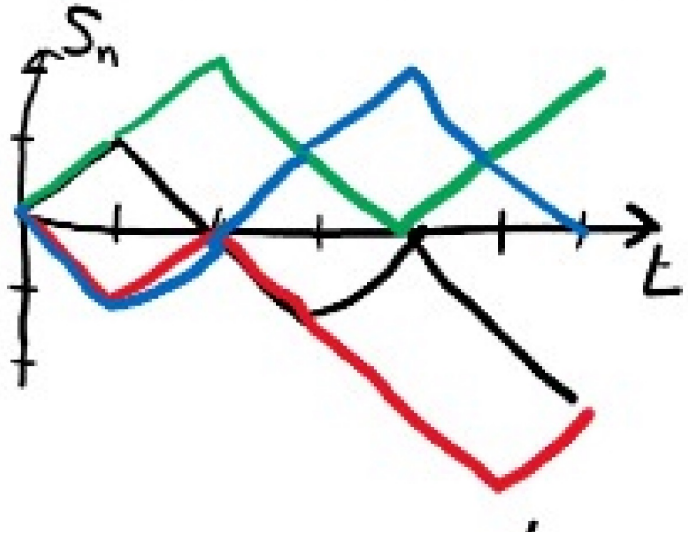


Los datos observados en una muestra realmente provienen de distribuciones conjuntas de una secuencia de variables aleatorias de las cuales se postula que x_t es una realización.

Problema: No sé conoce con certeza cuál será la próxima realización de una serie en particular.



Variable aleatoria: variable medible que toma valores dado una distribución de probabilidad que describe su comportamiento. La distribución de probabilidad va ser la encargada de definir todos los valores posibles de la variable junto con la probabilidad de que cada uno ocurra.



Conjunto de información

¿Qué información disponible se conoce?

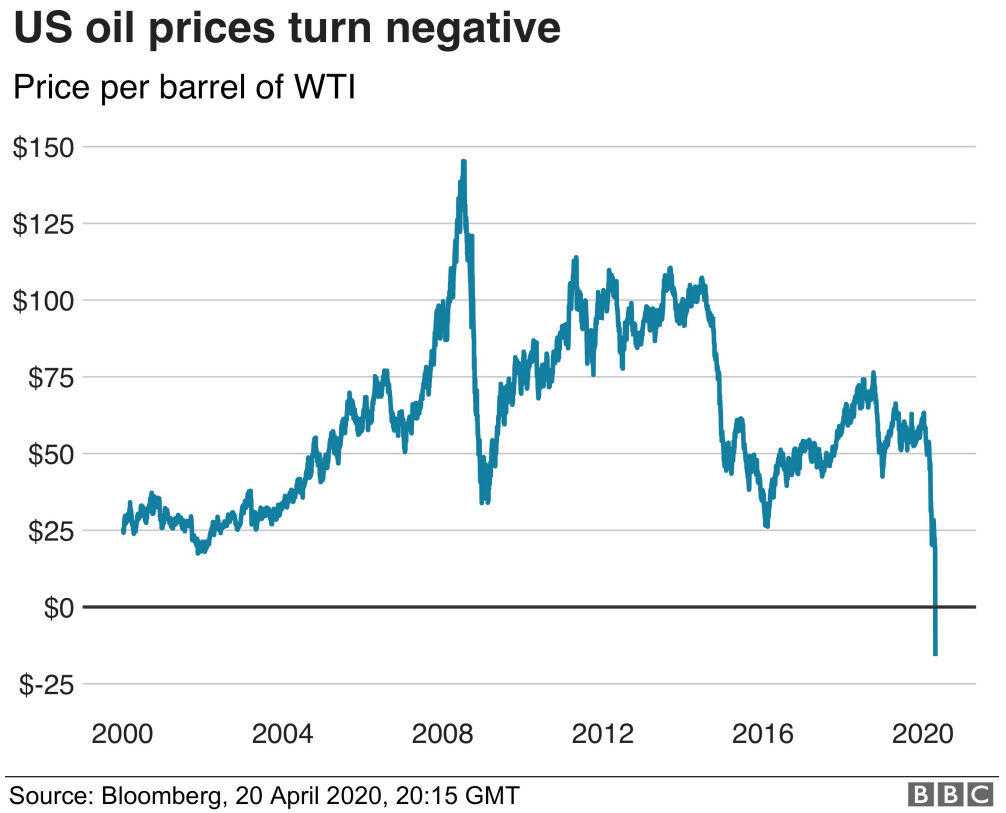
Conjunto de información: conjunto que contiene toda la información disponible de la serie de tiempo hasta un momento T.

Ejemplo: Perspectivas teóricas, opinión de expertos

¿Se puede mejorar el pronóstico aumentando el conjunto de información?

Cambios estructurales

- ¿Comportamientos constantes?
- ¿Cuáles cambian?
- ¿A qué velocidad cambian?



Motivación

Algunos beneficios son:

- Generar un mejor entendimiento de las correlaciones entre las diferentes variables de análisis.
- Simulación de escenarios.
- Anticipar escenarios.

Principio de parsimonia

- Mundo real es tremendamente complejo
- Modelos simples son mejores a modelos complejos:
  - Estimación de parámetros
  - Facilidad de interpretación y escrutinio de anomalías
  - Comunicación
  - Evita ajustar el modelo a idiosincrasias

Incertidumbre y mejoras

- Ningún modelo es el verdadero proceso generador de datos!!!
- Ningún modelo es correcto, pero algunos son útiles
- El mejor modelo depende de cada situación
- Un modelo que funcionaba bien, puede dejar de funcionar

Siempre y continuamente diagnosticar el desempeño empírico de los modelos y consistencia con la teoría.

Índice

Correlación

- Recordemos:
  - Covarianza:

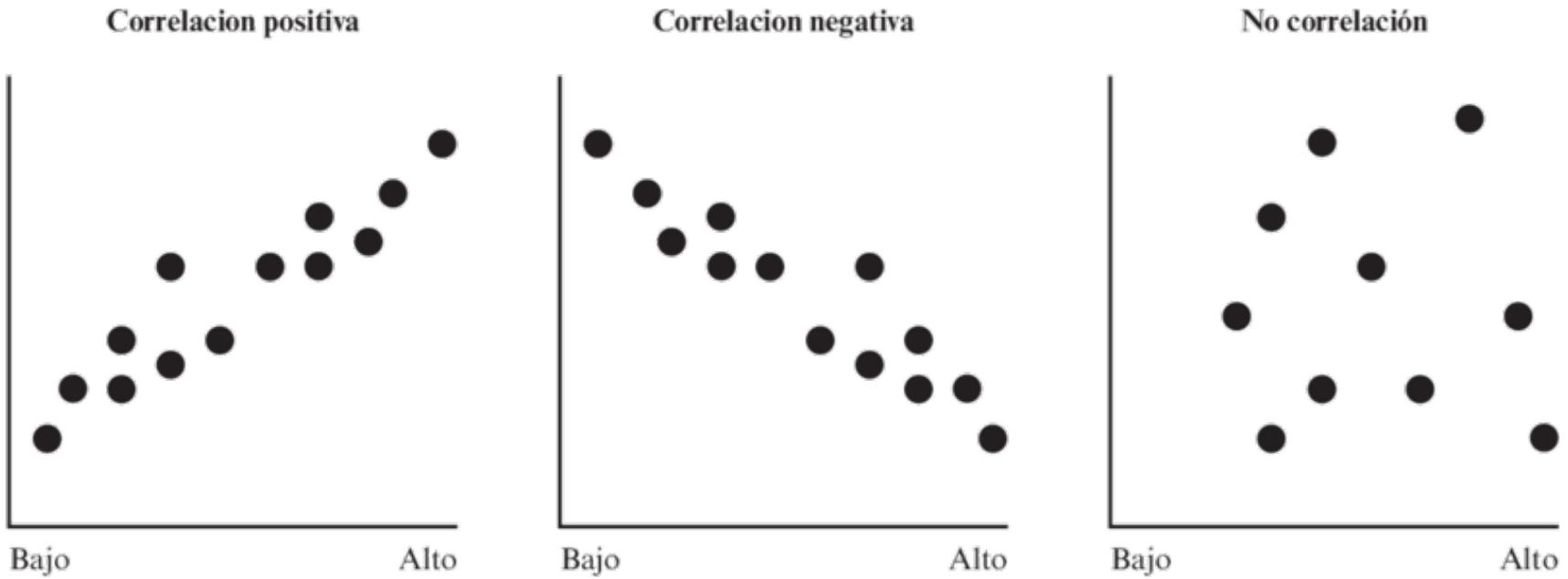
$$Cov(x,y) = E\left(\left(X - E(X)\right) * \left(Y - E(Y)\right)\right)$$

$$Cov(x,y) = \frac{1}{n} * \sum_{i=1}^{N} (x_i - \bar{x}) * (y_i - \bar{y})$$

  - Correlación de Pearson:

$$\rho = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

Correlación



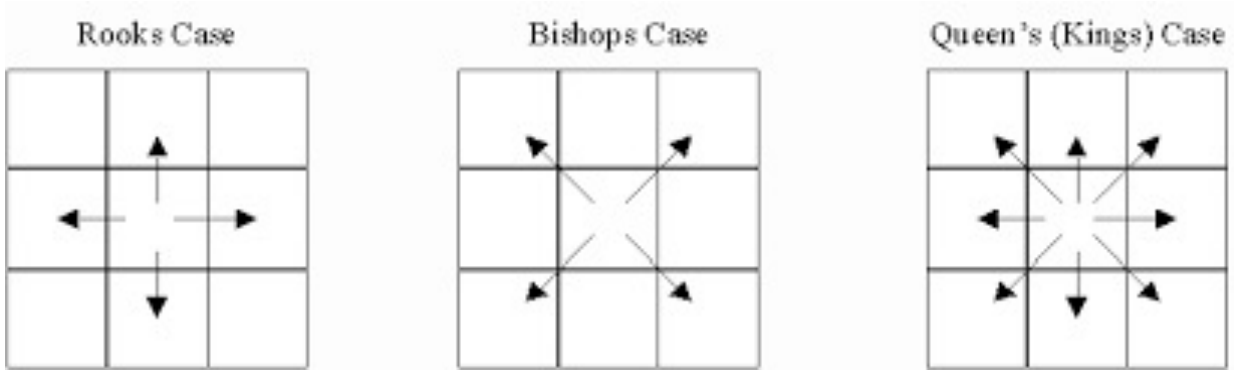
Correlación Espacial

Primero se debe construir la matriz de pesos espaciales W. Para esto hay diferentes alternativas:

Identificar el número de elementos continuos al objeto de análisis. (Similar a la matriz de adyacencia)

Binarios:

- Contiguidad de Rook (torre- Polígonos)
- Contiguidad de Bishop (alfil - Polígonos)
- Contiguidad de Queen (reina - Polígonos)
- Vecinos más cercanos (Punto)
- Distancia máxima (Punto)

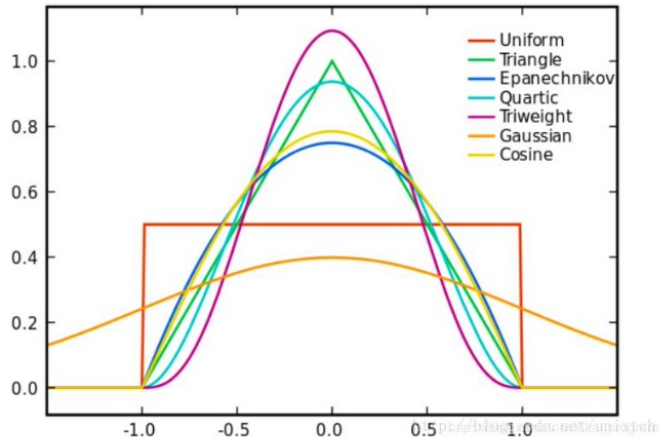


Por ejemplo: todos los municipios que cumplan con una de estas métricas van a tener un valor binario que lo cuente cómo vecino.

Correlación Espacial

Primero se debe construir la matriz de pesos espaciales W. Para esto hay diferentes alternativas:

Identificar el número de elementos continuos al objeto de análisis. (Similar a la matriz de adyacencia)



Continua:

- Funciones Kernel con distancia euclideana o haversine.

Por ejemplo: todos los municipios que cumplan con una de estas métricas van a tener un valor continuo que asigne una magnitud de qué tan vecino es.

Correlación Espacial

La interdependencia de los atributos de un lugar y su ubicación geográfica se conoce como auto correlación espacial global, y puede ser de tres tipos:

Positiva: Lugares cercanos espacialmente tienen a su vez valores similares de sus características.
Negativa: Lugares cercanos espacialmente tienen a su vez valores diferentes de sus atributos.
No hay: La cercanía espacial entre dos lugares no afecta la relación entre los valores de las características entre los lugares.

$$Lag\ Espacial = \sum_j w_j x_j$$

$$Lag\ Espacial = \frac{\sum_{vecinos} prevalencia}{\#vecinos}$$

Valor	Yopal	Mani	Tauramena	Leticia	Bogotá	Total
Aguazul		1	1	1	0	0	3
Normalizado		1/3	1/3	1/3	0	0	1
Covid Prop población		100	200	150	600	1000	
W*Prevalencia		33	67	50	0	0	150/3

Índice Gamma

Ayuda a entender la “correlación” que existe una variable para dos o más unidades geográficas castigando espacialmente.

$$\Gamma = \sum_i w_{ij} a_{ij}$$

Para el cual hay múltiples formas de definir α, el cual captura qué tanto se parecen los valores del atributo (prevalencia de covid) de dos municipios.

$$a_{ij} = |x_i - x_j| \qquad a_{ij} = \frac{|x_i - x_j|}{std(x)}$$

$$a_{ij} = \frac{(x_i - \bar{x})(x_j - \bar{x})}{var(x)} \qquad a_{ij} = (x_i - x_j)^2 \qquad a_{ij} = \frac{(x_i - x_j)^2}{var(x)}$$

Índice Gamma

La autocorrelación espacial se caracteriza por la correlación de una variable entre otras regiones en el espacio

Usando contiguidad normalizada (suma de filas = 1):

$$I = \frac{\sum_i (x_i - \bar{x})(\sum_j w_{ij} x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

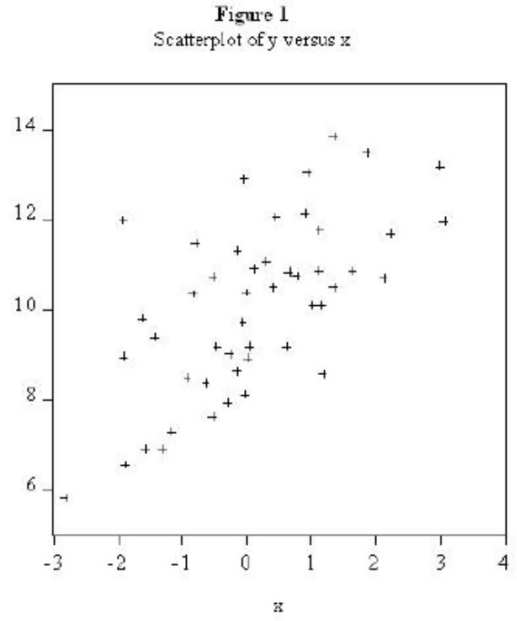
Valores positivos indican autocorrelación espacial positiva (mayor si cerca a 1), negativos autocorrelación espacial negativa (mayor si cerca a -1) y cercanos a 0 no autocorrelación espacial.

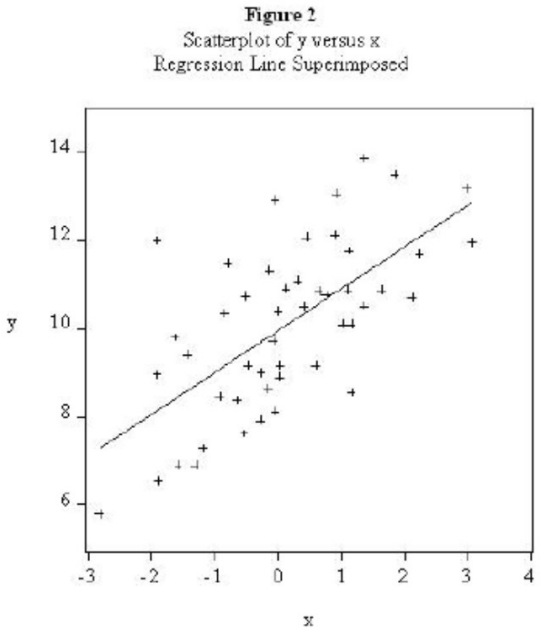
Ejemplo:

Qué tanto un lugar i con casos de COVID, se diferencia del promedio nacional (+,-)
Qué tanto el promedio de sus vecinos se diferencia del promedio nacional (+,-)

Mínimos cuadrados ordinarios (OLS)

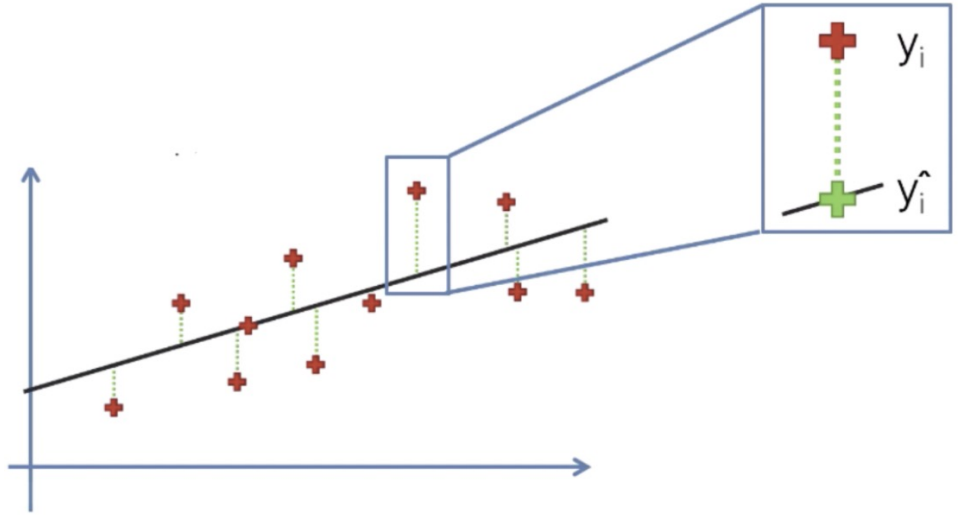
La función lineal de x que mejor se ajusta a y





Mínimos cuadrados ordinarios (OLS)

En una regresión lineal se quiere modelar la relación entre una variable dependiente Y y una o más variables explicativas X1,X2,...,Xk.



Una regresión lineal con n observaciones y k variables se expresa como:

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{i,k} + \epsilon_i; \quad i = 1, \ldots, n$$

Mínimos cuadrados ordinarios (OLS)

Ejemplo:

El gobierno quiere conocer cuál es el impacto de la inflación, el ingreso per cápita y área del país (millas cuadradas) sobre el peso de las importaciones como porcentaje del PIB. Para esto tiene una base de datos que contiene información sobre las dinámicas comerciales de 114 países para 1980.

Variable dependiente:

- importaciones como porcentaje del PIB (Open)

Variables independiente:

- Nivel de inflación para el año reportado (inf)
- Ingreso per cápita (pcinc)
- Área de cada país en millas cuadradas (land)

$$Openi = \beta_0 + \beta_1 inf_i + \beta_2 pcinc_i + \beta_3 land_i + \varepsilon_i$$

Mínimos cuadrados ordinarios (OLS)

Teniendo n observaciones y k variables, este modelo también se puede ver como:

$$Y = X^T \beta + \epsilon$$

Dónde:

$$Y = [Y_1, \ldots, Y_n]^T$$

$$\beta = [\beta_0, \beta_1, \ldots, \beta_k]^T$$

$$\epsilon = [\epsilon_1, \ldots, \epsilon_n]^T$$

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n3} & \cdots & x_{nk} \end{bmatrix}$$

Mínimos cuadrados ordinarios (OLS)

Teniendo n observaciones y k variables, este modelo también se puede ver como:

$$\text{OPEN} = \text{XB} + \text{U}$$

Dónde:

$$Open = \begin{bmatrix} Open1 \\ Open2 \\ Open3 \\ Open4 \\ \vdots \\ \vdots \\ \vdots \\ Open114 \end{bmatrix} = \text{XB} \begin{bmatrix} 1 & inf1 & pcinc1 & land1 \\ 1 & inf2 & pcinc2 & land2 \\ 1 & inf3 & pcinc13 & land3 \\ 1 & \vdots & \vdots & \vdots \\ 1 & inf114 & pcinc114 & land114 \end{bmatrix} x \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ \vdots \\ u_{114} \end{bmatrix}$$

Tamaño de los vectores:

OPEN= (114x1)
X = (114x4)
B = (4X1)
XB = (114 x 1)
U = (114 x 1)

Mínimos cuadrados ordinarios (OLS)

¿Cómo sabemos si el modelo de regresión lineal se ajusta adecuadamente a nuestros datos?

Necesitamos una **función de pérdida:**

Pero $\widehat{Y_i} = X\hat{\beta}$

$$L(\mathrm{B}) = \sum_{i=1}^{n} \left(Y_i - \widehat{Y_i}\right)^2$$

$$L(.) = \left(Y - X\hat{\beta}\right)^T \left(Y - X\hat{\beta}\right)$$

Al hacerlo matricialmente

$$L(.) = \left(Y' - \hat{\beta}'X'\right)(Y - X\hat{\beta})$$

$$L(.) = (Y'Y - Y'X\hat{\beta} - \hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta})$$

$$L(.) = (Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta})$$

Queremos encontrar β que minimiza la función de pérdida

$$\frac{\partial S}{\partial \hat{\beta}} = -2X'Y + 2X'X\hat{\beta} = 0$$

$$\hat{\beta} = [X'X]^{-1}X'Y$$

Matriz Simétrica es aquella matriz cuadrada que es igual a su traspuesta

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Por esto:

$$\beta'X'Y = Y'X\beta$$

Mínimos cuadrados ordinarios (OLS)

En nuestro ejemplo se vería algo así:

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

$$\hat{\beta} = \begin{bmatrix} \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ inf_1 & inf_2 & inf_3 & \cdots & inf_{114} \\ pcinc1_1 & pcinc1_2 & pcinc1_3 & \cdots & pcinc1_{114} \\ land_1 & land_2 & land_3 & \cdots & land_{114} \end{bmatrix} \begin{bmatrix} 1 & inf1 & pcinc1 & land1 \\ 1 & inf2 & pcinc2 & land2 \\ 1 & inf3 & pcinc13 & land3 \\ 1 & \vdots & \vdots & \vdots \\ 1 & inf114 & pcinc114 & land114 \end{bmatrix} \end{bmatrix}^{-1} \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ inf1 & inf2 & inf3 & \cdots & inf114 \\ pcinc1 & pcinc2 & pcinc3 & \cdots & pcinc114 \\ land1 & land2 & land3 & \cdots & land114 \end{bmatrix} \mathrm{X} \begin{bmatrix} Open1 \\ Open2 \\ Open3 \\ Open4 \\ \vdots \\ \vdots \\ \vdots \\ Open114 \end{bmatrix}$$

Variables	Valores estimados
Explicada - open	
fila1 - $(\widehat{\beta_0})$	40.954539
inf - $(\widehat{\beta_1})$	-0.15645901
pcinc - $(\widehat{\beta_2})$	0.00055232
land - $(\widehat{\beta_3})$	-0.00001093

Supuestos

- El modelo de regresión es lineal en los coeficientes y el término de error.
- Los errores se distribuyen normal con media 0:

$$\mathbb{E}(\epsilon_i \mid X_{1i}, \ldots, X_{ki}) = 0$$

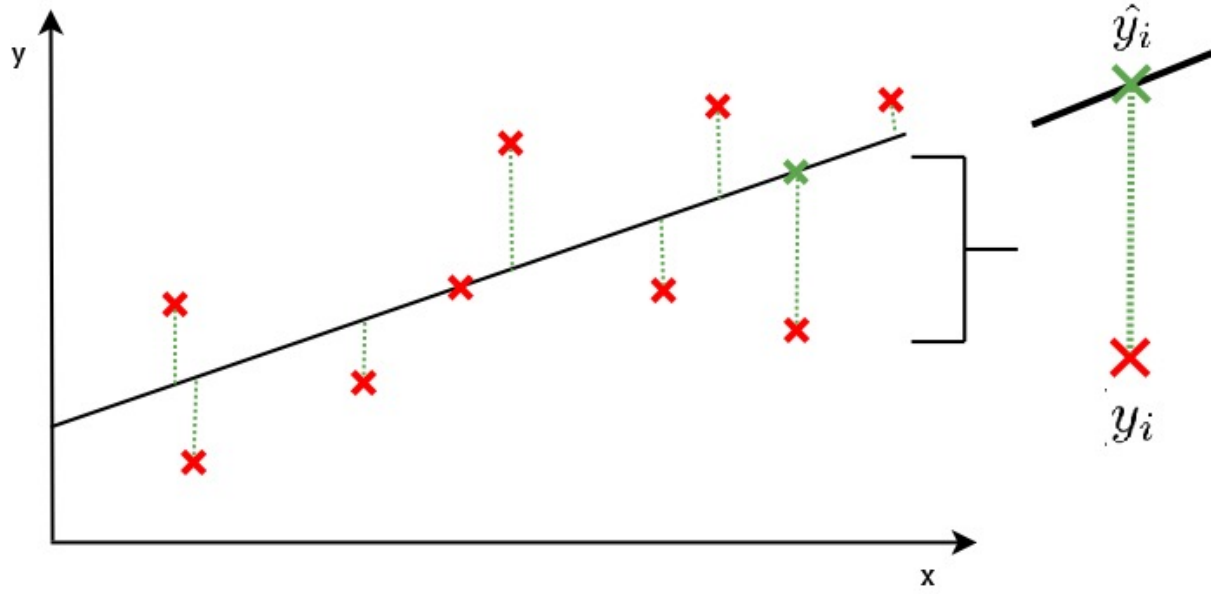
- Los errores tienen varianza constante (no heteroscedasticidad = homoscedasticidad)
- Los errores se distribuyen normalmente
- Los regresores X1, ..., Xk con i.i.d. (independientes e idénticamente distribuidos)
- Las variables son independientes del error
- Las observaciones de los errores son independientes
- Ninguna variable independiente es una función lineal perfecta de otras variables explicativas (Multicolinealidad)

Suma de los Cuadrados Residuales

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

y_i = Valores observados de la variable dependiente

$\hat{y}_i$ = Valores estimados por el modelo



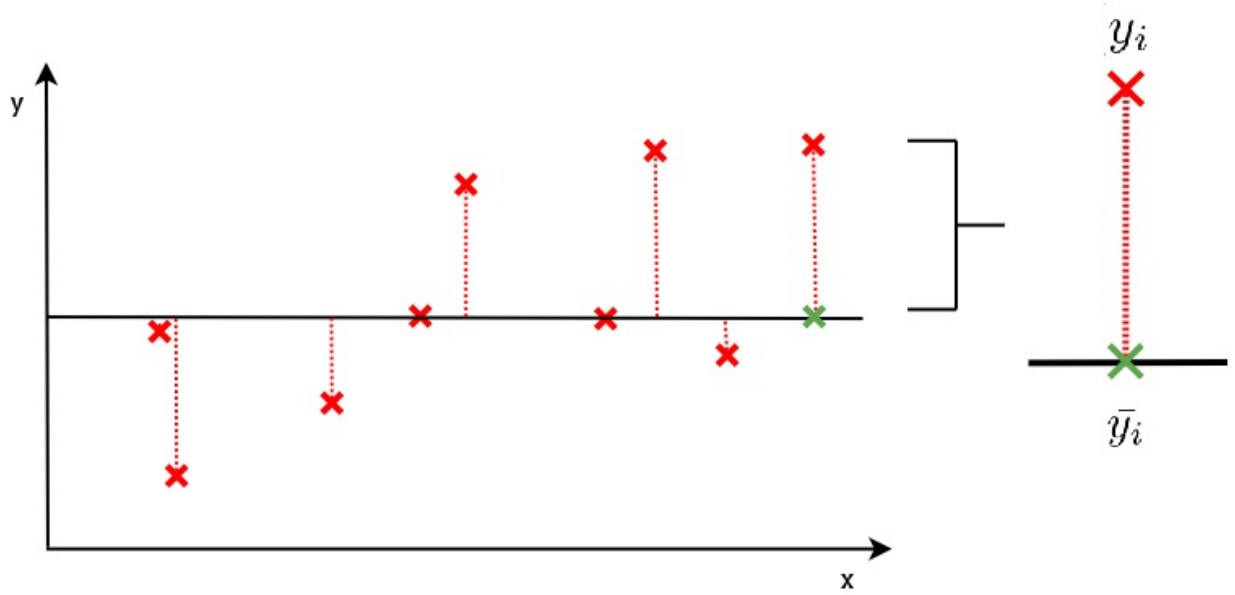
La parte que las variables independientes no son capaces de explicar sobre la variabilidad de la variable dependiente.

Suma Total de los Cuadrados

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

y_i = Valores observados de la variable dependiente

$\bar{y}$ = Valor promedio de los datos recogidos



Error cuadrático medio

Objetivo: comparar un valor predicho y un valor observado o conocido:

$$MSE = \frac{\sum_{t=1}^{T}(y_t - \hat{y}_t)^2}{T} = \frac{\sum_{t=1}^{T} e_t^2}{T}$$

Entre más pequeño el valor del criterio este es mejor:

Una mayor diferencia se produce debido a que en el modelo existe mayor aleatoriedad o porque el estimador no tiene en cuenta la información que podría producir una estimación más precisa.

R – Cuadrado (**Coeficiente de determinación**)

El R^2 (coeficiente de determinación) nos dice cual es la fracción de la varianza muestra de y_i explicada por x_t

$$R^2 = 1 - \frac{\text{SSR}}{SST}$$

En otros términos, muestra el ajuste de la línea del modelo de regresión lineal con respecto a la línea del promedio entre los valores de la variable dependiente que se está calculando.

Si SSR ↑ → R^2 ↓

Si SSR ↓→ R^2. ↑ Deseado

R – Cuadrado ajustado

El R^2 asume que cada variable individual explica la variación en la variable dependiente.

El R^2 ajustado indica el porcentaje de variación explicado solo por las variables independientes que realmente afectan la variable dependiente.

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

A valores más altos de k el $\bar{R}^2$ será bajo que el del R Cuadrado normal.

Métricas de desempeño

- **Error cuadrático medio (MSE):** la métrica más utilizada que otorga una penalización más alta a los errores grandes y viceversa, $[0, \infty)$.

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

- **Raíz cuadrática media (RMSE):**

$$\text{RMSE} = \sqrt{MSE}$$

RMSE se puede interpretar como la desviación estándar del error de medición. (Raíz para tener unidades en mismo valor)

Métricas de desempeño

- **Error absoluto medio (MAE):** Métrica fácilmente interpretable dado que tiene las mismas unidades que la variable en cuestión.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\, y_i - \widehat{y_i}|$$

- **Error absolute porcentual medio (MAPE):** Igual que el MAE pero en porcentaje con respecto a la variable de interés.

$$MAPE = \frac{100}{n}\sum_{i=1}^{n} \frac{|y_i - \widehat{y_i}|}{y_i}$$