Medidas de asociación lineal y el modelo lineal con dos variables

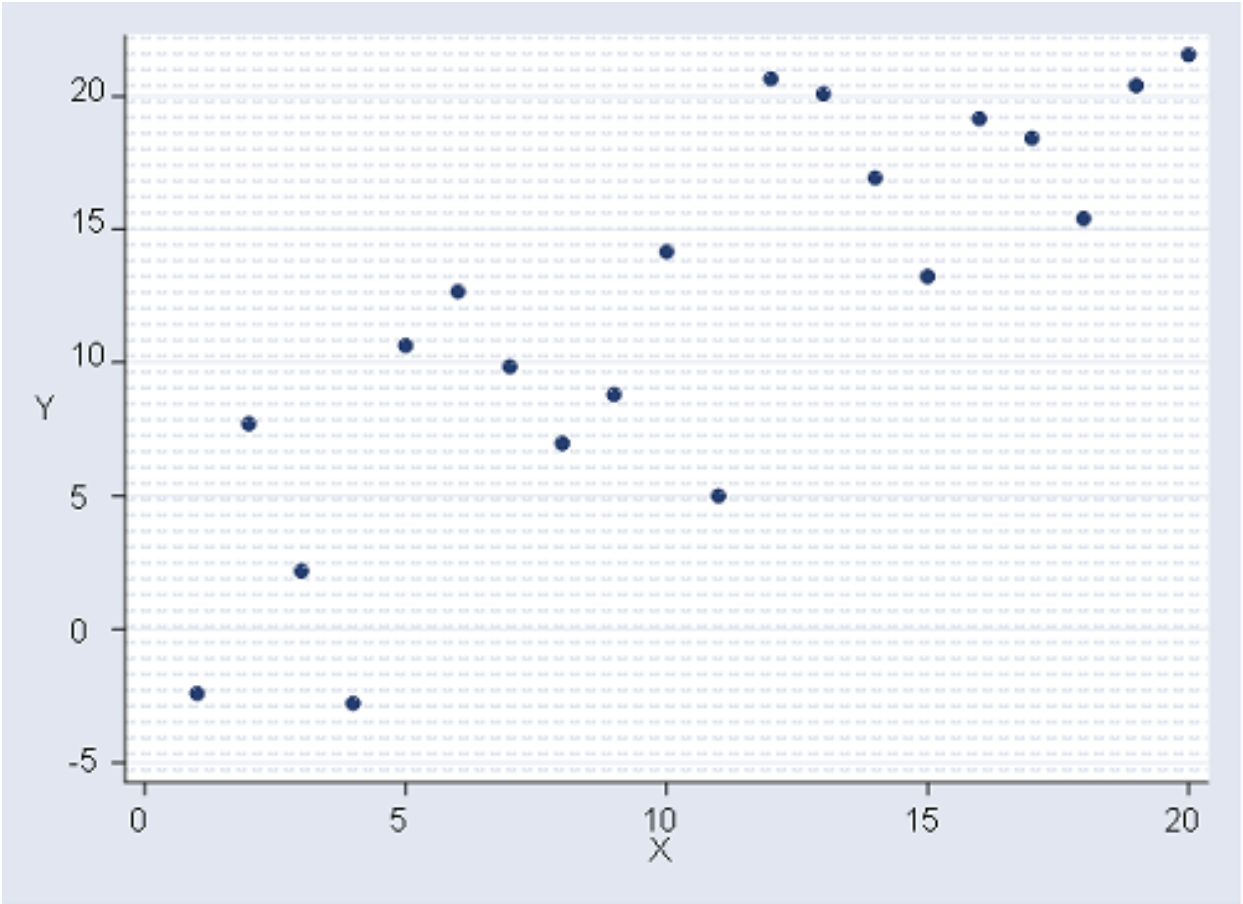
Mariana Marchionni
marchionni.mariana@gmail.com

Econometría I - FCE - UNLP
www.econometria1unlp.com

Objetivo de esta clase

- **Nos interesa la relación entre dos variables económicas**
- Recordemos la naturaleza social de los fenómenos económicos:
  - Relaciones no exactas
  - Fenómenos complejos (muchas unidades decisorias, alto grado de interacción)
  - Datos observacionales ($\neq$ experimentales)

La relación entre 2 variables económicas gráficamente



¿En qué sentido decimos que hay una relación entre Y y X?

¿Cómo podemos caracterizarla?

En la clase de hoy

1. Medidas de asociación lineal
2. El modelo lineal con dos variables

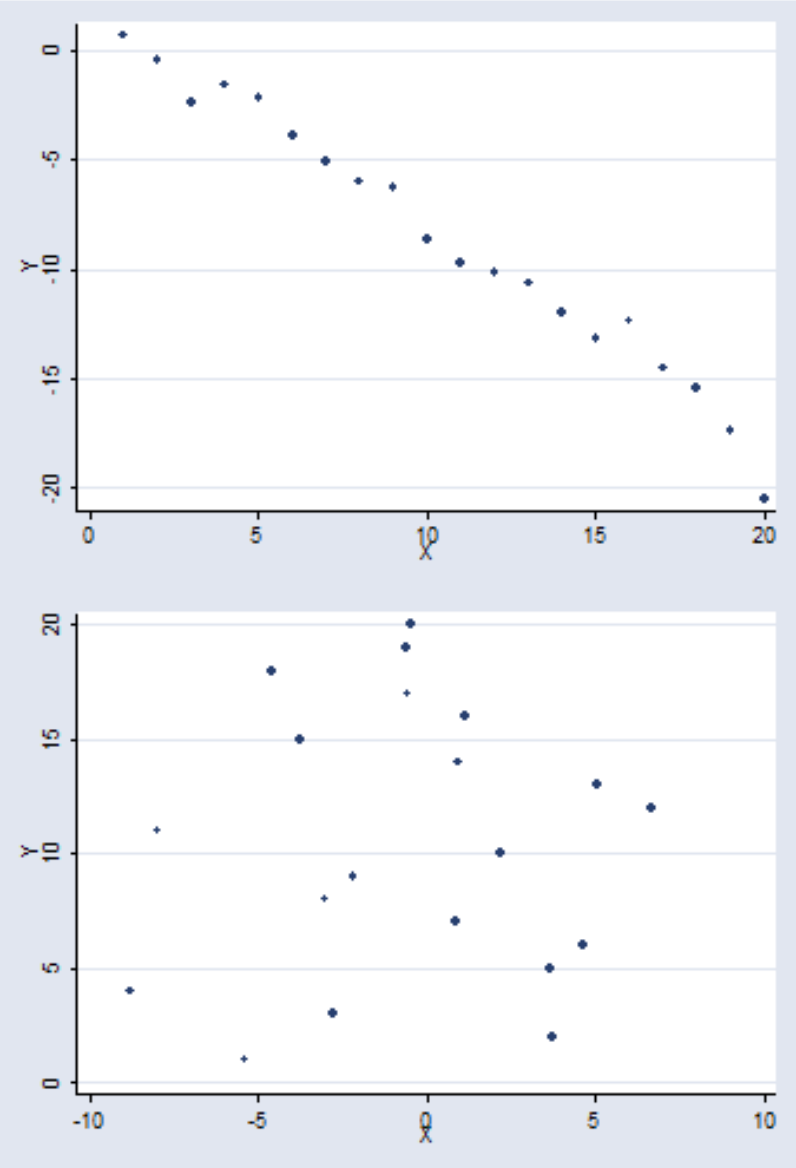
Medidas de asociación lineal: covarianza y correlación

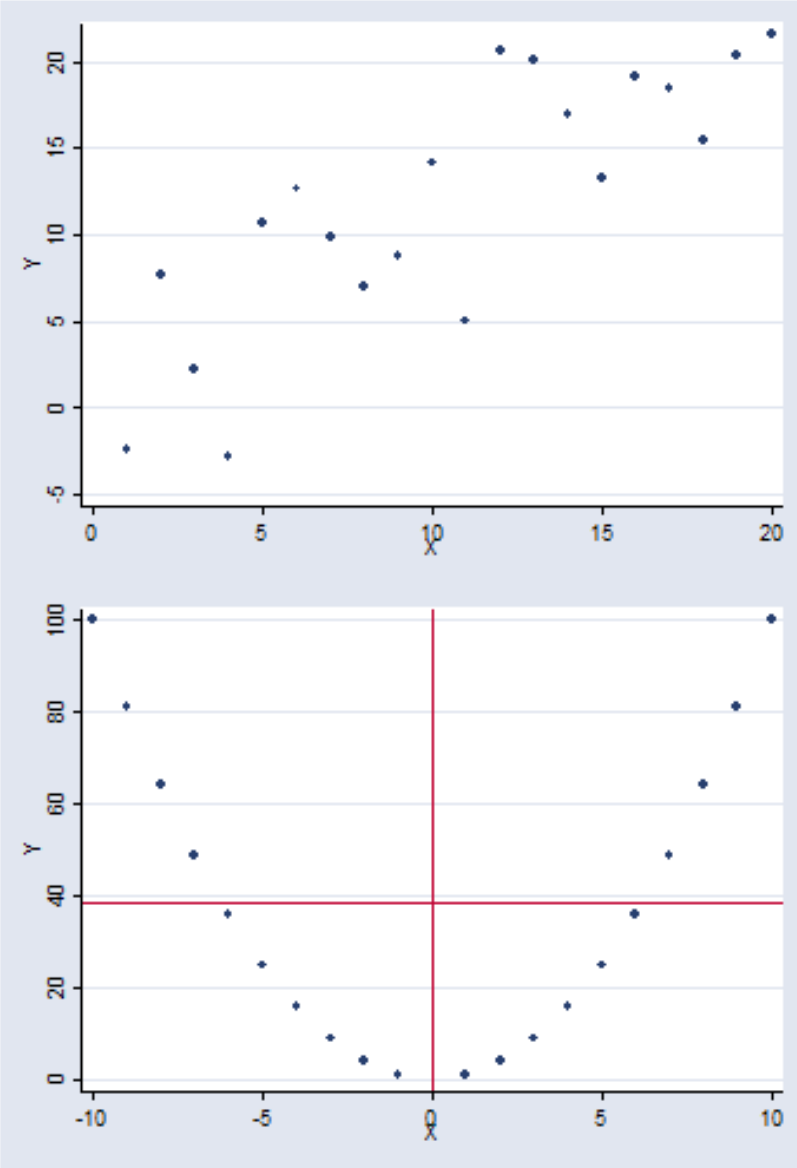
Objetivos:

- Caracterizar la relación entre dos variables Y y X.
  - determinar la *dirección* de esa relación (es decir, si es positiva o negativa)
  - y medir la *fuerza* o *intensidad* de esa relación.

- Análisis a partir de una muestra aleatoria (Y_i, X_i) con $i = 1, ...N$.
  - en realidad, en la práctica contaremos con una única realización de una muestra aleatoria. Estos son los datos.
  - Recordar: la muestra aleatoria es un conjunto de variables aleatorias; los datos son números!

Ejemplos





Definiciones: covarianza y correlación (poblacional)

Sean Y y X dos variables aleatorias, con esperanzas $E(Y) = \mu_Y$ y $E(X) = \mu_X$, y varianzas $V(Y) = \sigma_Y^2$ y $V(X) = \sigma_X^2$

- La covarianza entre Y y X se define como

$$E[(Y - \mu_Y).(X - \mu_X)]$$

- La correlación entre Y y X se define como

$$\frac{E[(Y - \mu_Y).(X - \mu_X)]}{\sigma_Y.\sigma_X}$$

- ¿Cómo se estiman estos momentos poblacionales a partir de una muestra?

Covarianza y correlación muestral

- **Covarianza muestral**

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

donde $\bar{Y}$ y $\bar{X}$ son las medias muestrales de Y y X.

- **Correlación muestral**

$$r_{Y,X} = \frac{Cov\left(Y,X\right)}{S_Y\, S_X}$$

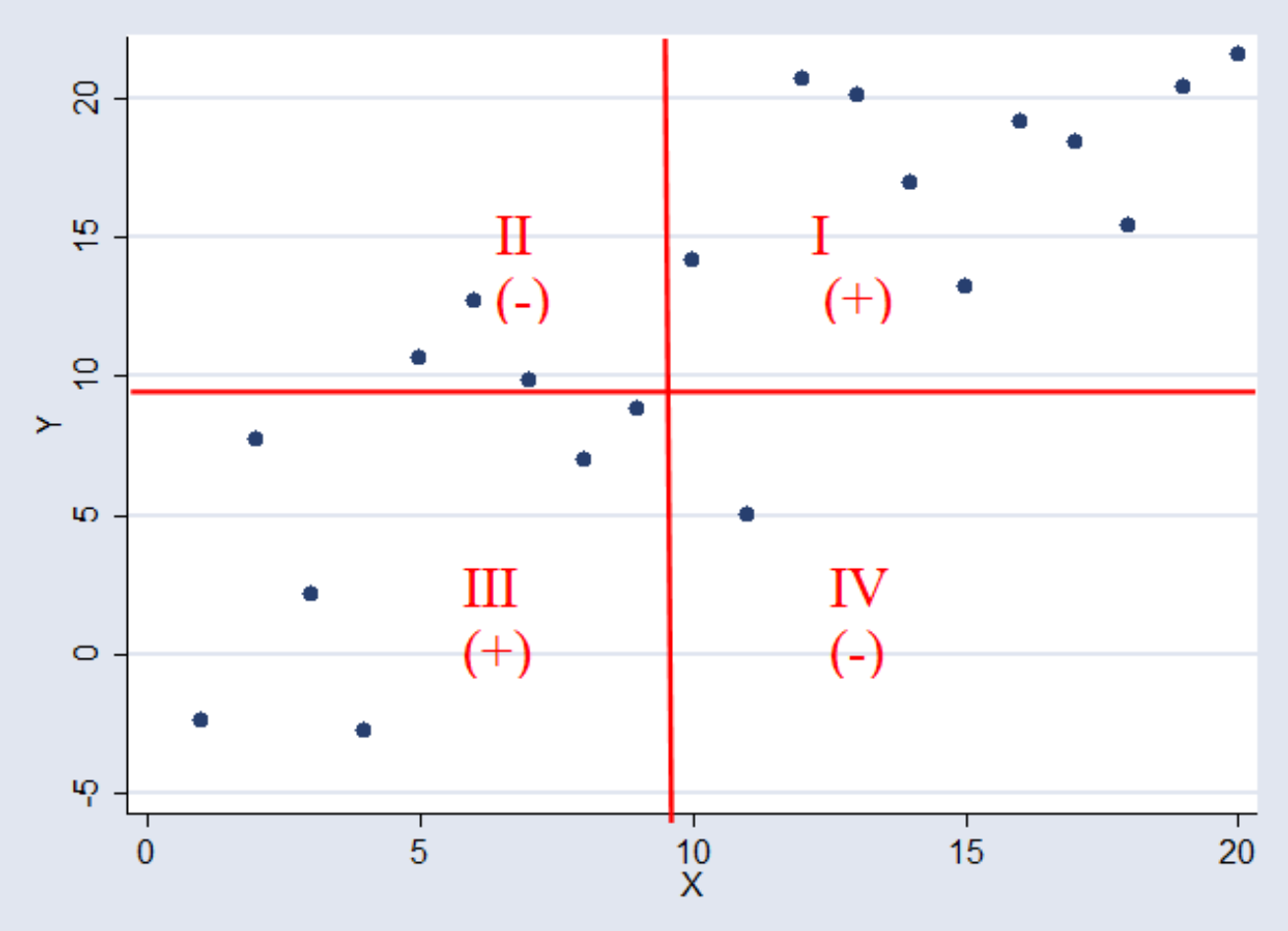
donde S_Y y S_X son los desvíos estándar muestrales de Y y X.

Propiedades de la covarianza y la correlación

$$Cov\,(Y,X) = \frac{\sum_{i=1}^{N} (Y_i - \overline{Y})\,(X_i - \overline{X})}{N-1}; \quad r_{Y,X} = \frac{Cov\,(Y,X)}{S_Y\,S_X}$$

- Son simétricas: $Cov(Y,X) = Cov(X,Y)$ y $r_{Y,X} = r_{X,Y}$
- El signo de la covarianza es igual al signo de la correlación.
- El signo indica la dirección de la asociación. ¿Por qué?

Recordar: el numerador $\sum_{i=1}^{N} (Y_i - \overline{Y})(X_i - \overline{X})$ determina el signo de la covarianza y la correlación. Notar: cada término de la sumatoria suma o resta dependiendo del cuadrante.



Propiedades de la covarianza y la correlación

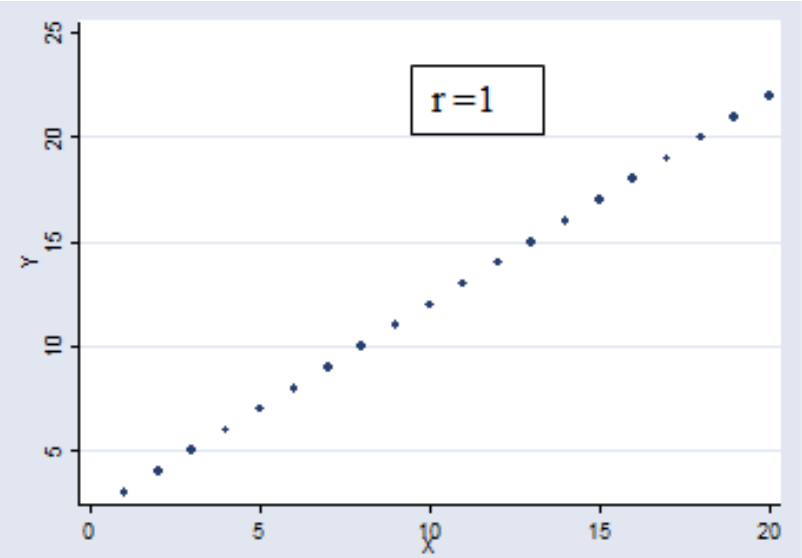
- La *covarianza* depende de las unidades de medida, pero la *correlación* no.
  - Sean a y b constantes positivas:
    - $Cov(aY, bX) = a.b.Cov(Y, X)$
    - $r_{aY,bX} = r_{Y,X}$
  - Ej.: suponer que X e Y están medidas en \$ y que $Cov(Y, X) = 50$ y $r_{Y,X} = 0.7$. Ahora reexpresamos los valores de X e Y en centavos. La correlación seguirá siendo 0.7 pero la nueva covarianza será 500.000.
- Esto implica una ventaja para interpretar el coeficiente de correlación muestral.

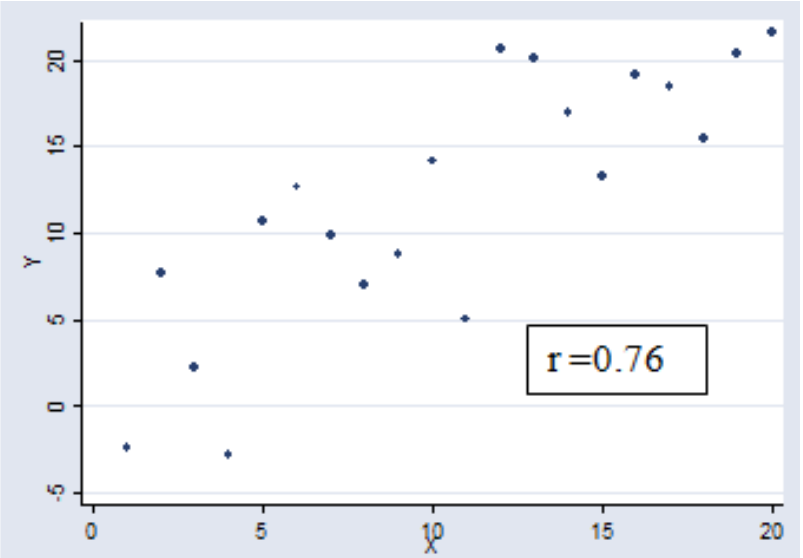
Propiedades de la covarianza y la correlación

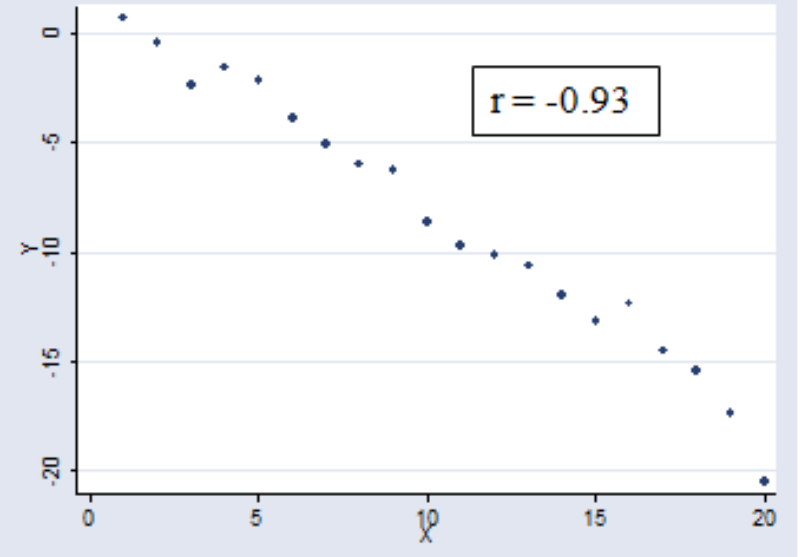
Además, el coeficiente de correlación

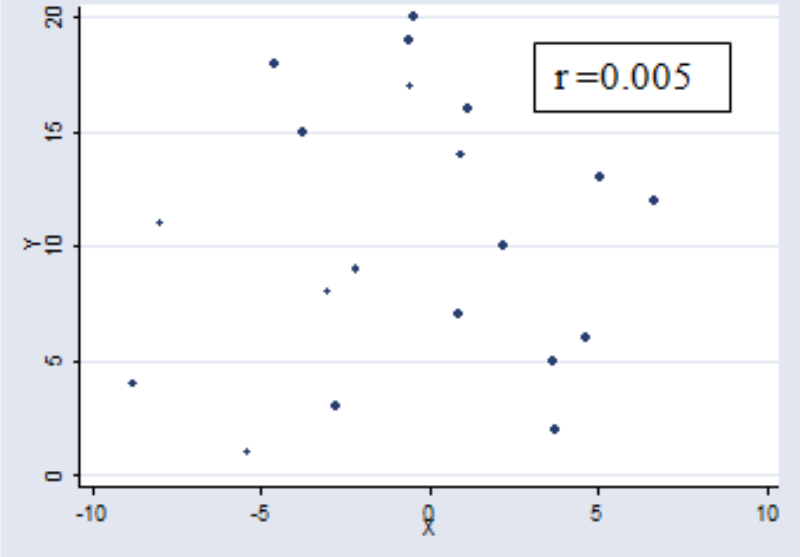
- Está acotado entre -1 y 1: $-1 \leq r_{Y,X} \leq 1$
- $r_{Y,X} = 1$ sólo cuando existe una *relación lineal exacta y directa* entre las variables Y y X:
  Para todo $i = 1, ...N$, $Y_i = \alpha + \beta X_i$ para algún $\beta > 0$.
- $r_{Y,X} = -1$ sólo cuando existe una relación lineal exacta e indirecta entre las variables Yy X:
  Para todo $i = 1, ...N$, $Y_i = \alpha + \beta X_i$ para algún $\beta < 0$.

Ejemplos


r =1
Y
X
0
5
10
15
20
25

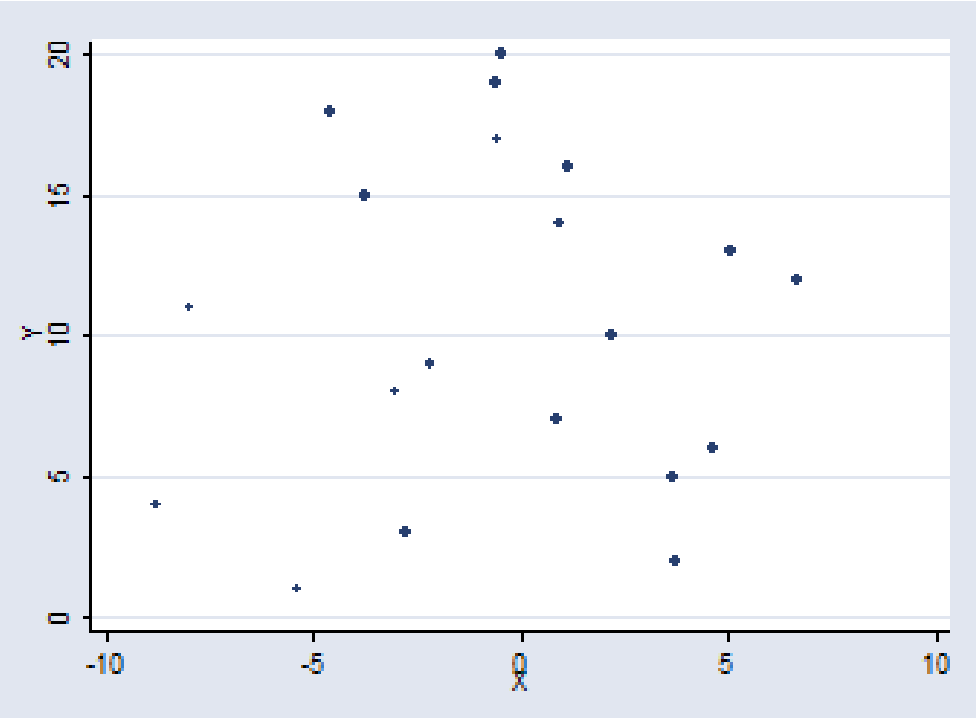

r =0.76
Y
X
-5
0
5
10
15
20


r = -0.93
Y
X
0
-5
-10
-15
-20


r =0.005
Y
X
-10
-5
0
5
10
15
20

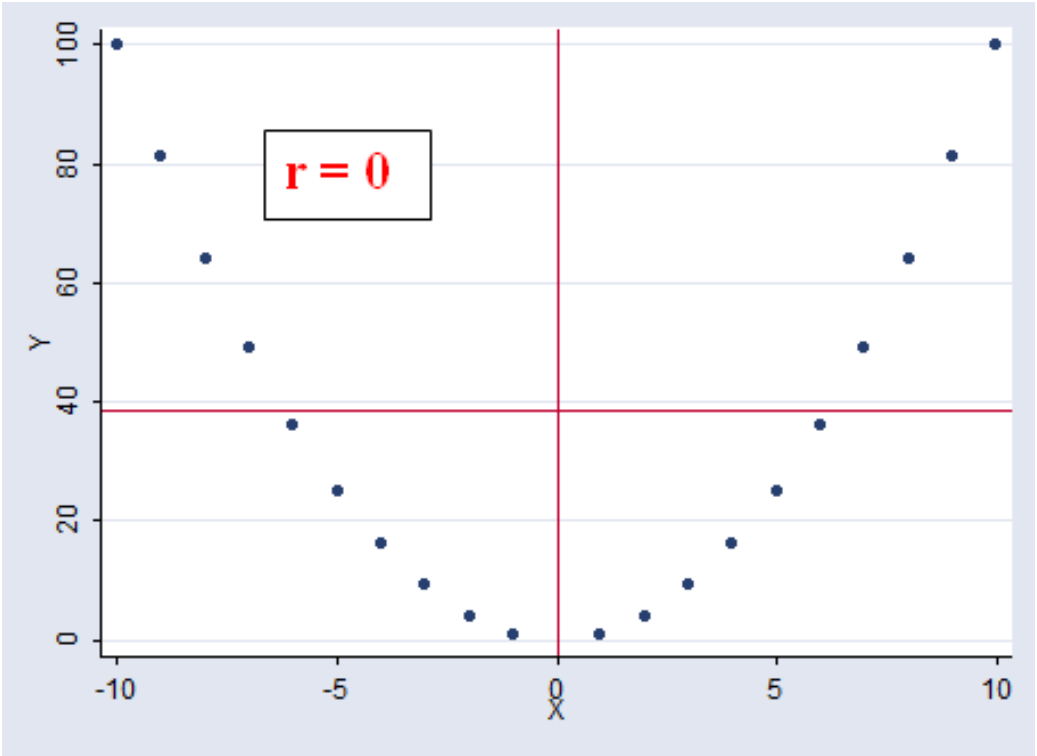
La correlación sólo mide relaciones lineales

Si no hay relación $\Rightarrow r_{Y,X} = 0$



La correlación sólo mide relaciones lineales

La recíproca no se cumple: $r_{Y,X} = 0 \not\Rightarrow$ que no hay relación



Que $r_{Y,X} = 0$ implica únicamente que **no hay relación lineal**
(pero puede haber alguna relación no lineal)

Correlación no implica causalidad

Ejemplo: relación entre la inversión en ciencia y tecnología (CyT) y el crecimiento de un país

- Resultado empírico: están correlacionadas positivamente
- ¿Hay que invertir en CyT para que el país crezca?
- ¿O cuando el país crece se invierte más en CyT?
- El coeficiente de correlación no responde.

El modelo lineal con dos variables

Objetivo: modelar una relación lineal no exacta entre Y y X.

Modelo propuesto:

$$Y_i = \alpha + \beta X_i + \mu_i \qquad i = 1, ..., N$$

- Y_i: variable explicada o dependiente. Observable.
- X_i: variable explicativa o independiente. Observable.
- α y β: parámetros desconocidos.
- μ_i: término de error que representa a todas las variables inobservables que afectan a Y_i.
  A los fines de la modelizacion lo consideramos como una variable aleatoria: término aleatorio.
- Estamos proponiendo un modelo para la muestra aleatoria $(Y_i, X_i) \quad i = 1, ..., N$. Nuestros datos serán una realización de la muestra aleatoria.

La función de regresión

$$Y_i = \alpha + \beta X_i + \mu_i \qquad i = 1, ..., N$$

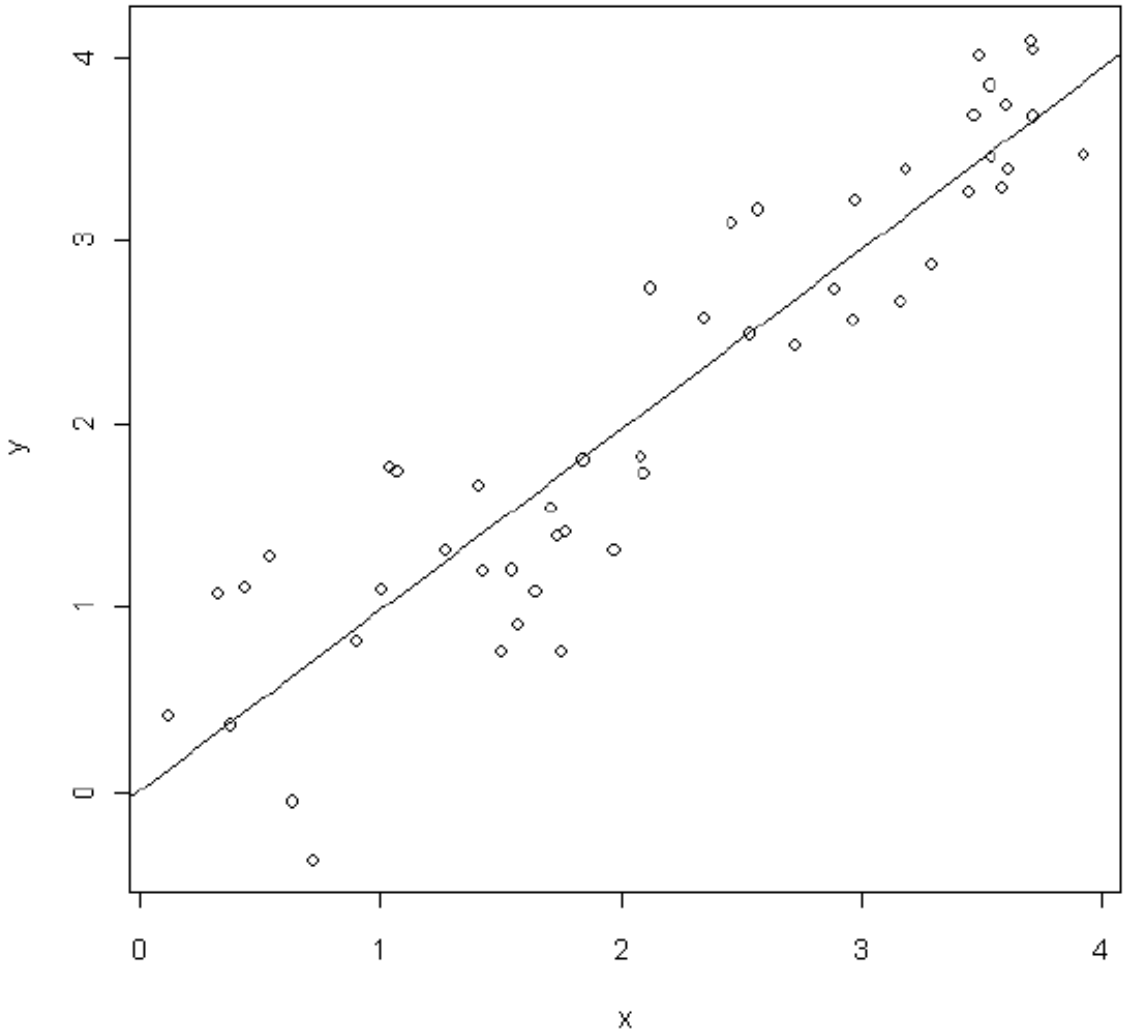
- La función $\alpha + \beta X_i$ se conoce como función de regresión y representa la parte sistemática de la relación.
- El término aleatorio μ_i representa la parte no sistemática (aleatoriedad).

Interpretación

Es como si Y_i se determinase en dos pasos.
Dado un valor de $X_i = x_0$

1. La parte sistemática es $\alpha + \beta x_0$
2. El verdadero Y_i es la parte sistemática más algo aleatorio.

Cada uno de los N puntos (Y_i, X_i) se determina por un valor sobre la recta $\alpha + \beta X_i \pm$ un shock aleatorio (μ_i)



¿Conocemos esta línea?

Ejemplo: modelo de los determinantes del consumo de las familias

$$Y_i = \alpha + \beta X_i + \mu_i \qquad i = 1, ..., N$$

- Datos para N familias.
- Y_i es el consumo de la familia i.
- X_i es el ingreso de la familia i.
- μ_i son factores distintos al ingreso que también afectan al consumo familiar: cantidad de miembros, edades, gustos, etc. En el modelo simple con una sola variable explicativa consideramos a todos esos factores como inobservables.

El rol de μ_i

- μ_i representa al conjunto de factores no observable que afectan a Y_i.
- ¿Qué sabemos sobre μ_i? Nada con certeza. Podríamos suponer cosas que parezcan razonables...
  - ¿$\mu_i = 0$? Si fuera cierto, la relación entre Y y X sería perfectamente lineal (determinística). No es un supuesto razonable para variables económicas.
  - ¿$E(\mu_i) = 0$? Si fuera cierto, el valor medio de μ_i en la poblacion sería cero. En ese caso $E(Y_i) = \alpha + \beta X_i$, es decir, la relación entre Y y X sería lineal en promedio.
    - todavía no vamos a suponer nada.
- El término aleatorio es una representación de lo no exacto de la relación.

El rol de β

- β contiene información muy importante:

$$E(Y_i) = \alpha + \beta X_i$$

- Si es posible mover X marginalmente:

$$\frac{dE(Y_i)}{dX_i} = \beta \text{ para todo } i$$

  - β contiene información cualitativa (signo) y cuantitativa (magnitud) acerca de cómo se asocian los cambios en X con los cambios esperados en Y.
  - En otras palabras: β mide en cuánto cambia Y en promedio cuando X cambia marginalmente: efecto marginal.

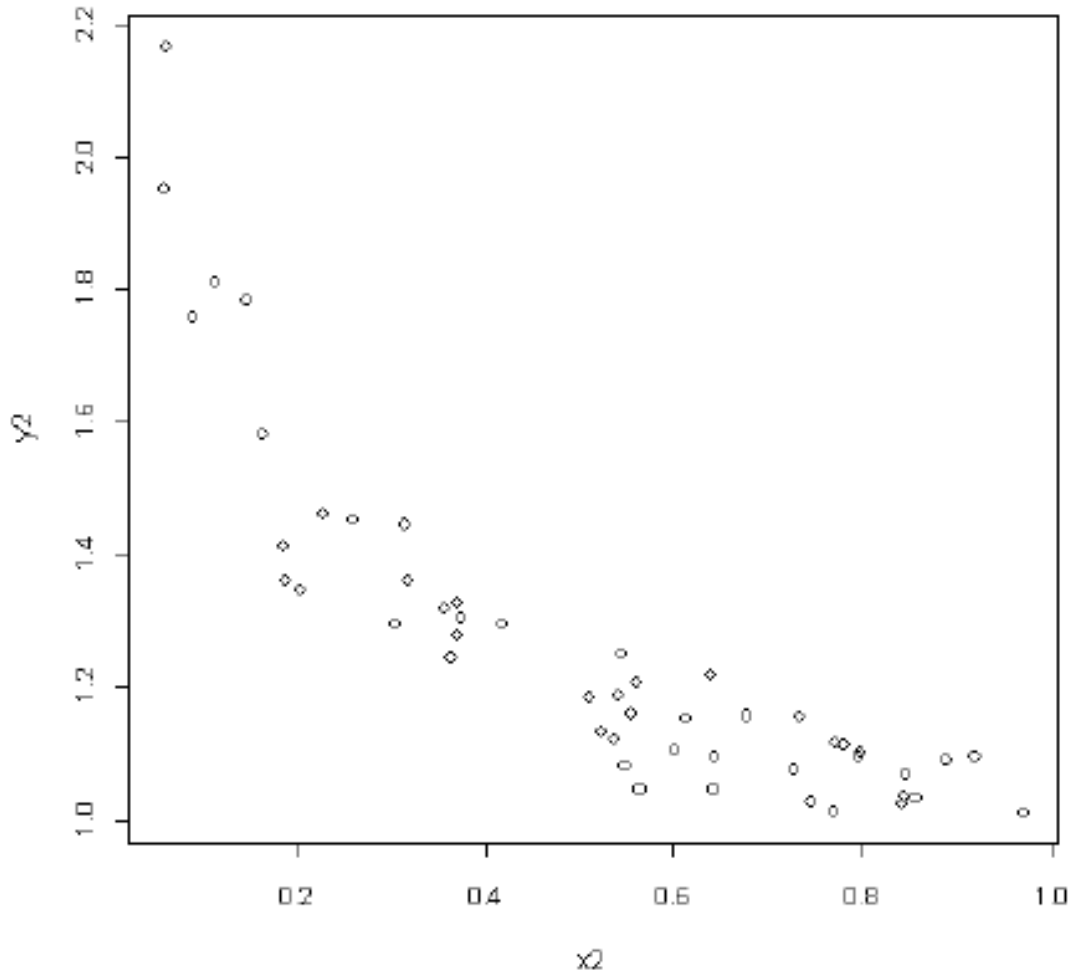
- Si X cambia en forma discreta: $\triangle E(Y_i) = \beta \triangle X_i$ (vale la "regla de tres").
- Es crucial tener en cuenta las unidades de medida de las variables para interpretar β.

Ejemplo: consumo de las familias (cont.)

- Supongamos otra vez nuestro modelo $Y_i = \alpha + \beta X_i + \mu_i$
- Supongamos que el consumo y el ingreso de las familias están medidos en miles de pesos y que $\beta = 0.8$
- Cuando el ingreso de una familia aumenta en una unidad, en este caso \$1000, *el valor esperado* de su consumo aumenta en \$800
- Errores de interpretación comunes (por omisión o trivialidad):
  - Cuando X aumenta, Y aumenta (trivial).
  - Cuando X aumenta, Y aumenta en 0.8 (¿en cuánto aumenta X? ¿0.8 qué?)

¿Vamos a estudiar solo modelos lineales?

Puede ser que los datos sugieran una relación no lineal entre las variables



Algunos modelos no lineales

- Los modelos lineales no son lo suficientemente generales. En muchas aplicaciones de la Economía se usan modelos no lineales.
- Vamos a ver 2 ejemplos:
  - cuando la variable dependiente aparece en forma logarítmica y la variable explicativa en forma lineal (modelo log-lin)
  - cuando tanto la variable dependiente como la explicativa se expresan en logaritmos (modelo log-log)

Modelo logarítmico o log-log

- Tanto variable dependiente como variable explicativa están expresadas en logaritmos

$$lnY_i = \alpha + \beta lnX_i + \mu_i$$

- Si μ_i se mantiene constante cuando X_i cambia:

$$\beta = \frac{dlnY_i}{dlnX_i} \cong \frac{\triangle \mathrm{Y_i}/\mathrm{Y_i}}{\triangle \mathrm{X_i}/\mathrm{X_i}}$$

- β es una elasticidad constante: porcentaje en el que cambia Y ante un aumento de un 1% en X.
- Notar que no importan las unidades de medida originales de Y ni de X

Modelo logarítmico o log-log

Ejemplo: modelo de demanda con elasticidad constante

$$lnQ_i = \alpha + \beta lnP_i + \mu_i$$

- Q (cantidades) está medida en unidades y P (precio) se mide en pesos
- Supongamos que $\beta = -0.5$
- Interpretación: cuando el precio aumenta en un 1%, la cantidad demandada cae un 0.5%
- En este tipo de modelo las unidades de medida no importan para la interpretación
- Notar que estamos manteniendo a μ_i constante cuando cambia P_i

Modelo log-lin o semilogarítmico

- Sólo la variable dependiente está expresada en logaritmos

$$lnY_i = \alpha + \beta X_i + \mu_i$$

- Si μ_i se mantiene constante cuando X_i cambia:

$$\beta = \frac{dlnY_i}{dX_i} \cong \frac{\triangle Y_i / Y_i}{\triangle X_i}$$

- β es una semielasticidad: $100.\beta$ es el porcentaje en el que cambia Y cuando X aumenta en una unidad.
  - Notar que multiplicamos por 100 para obtener el cambio porcentual.
- Para la interpretación no importan las unidades de medida de Y pero sí las de X.

Modelo log-lin o semilogarítmico

Ejemplo: modelo de los determinantes de los salarios

$$lnW_i = \alpha + \beta edu_i + \mu_i$$

- W (salario) está medido en miles de pesos por mes y *edu* (educación) en años
- Supongamos $\beta = 0.07$
- Interpretación: cuando la educación *aumenta en un año*, el salario mensual aumenta un 7% (0.07×100).
- Las unidades de medida de Y no importan, ya que los cambios son porcentuales, pero sí las de X.
- Notar que estamos manteniendo a μ_i constante cuando cambia edu_i

Vamos cerrando...

- Existen varios modelos no lineales, los veremos a lo largo del curso.
- También discutiremos criterios acerca de cual utilizar.
- En la próxima clase nos vamos a concentrar en cómo estimar los parámetros desconocidos del modelo lineal con dos variables.

Bibliografía para esta clase

- Notas de clase: introducción y cap. 1 secciones 1.1 y 1.2.
- Wooldridge: capítulo 1.