Econometría I (EC402)-II semestre de 2013
Clase #4 y #5 - Análisis de regresión simple, motivación y significado

Andrés M. Castaño

Ingeniería Comercial
Universidad Católica del Norte
Agosto 14 y 19 de 2013

Regresión Lineal Simple: Motivación

- **análisis de regresión:** estudio de la relación entre una variable Y y una (o más variables) X
- **regresión lineal simple:** una variable X.
- **Y:** variable dependiente ó explicada ó regresada ó respuesta ó predicha ó endógena ó resultado ó controlada.
- **X:** variable independiente ó explicativa ó regresora ó predictora ó estímulo ó exógena ó covariante ó de control.

Regresión Lineal Simple: Motivación

- Relaciones determinísticas vs relaciones estocásticas
- Regresión vs causalidad. No siempre existe una relación de causalidad $\Longrightarrow$ Importancia de la teoría.
- Regresión vs correlación.
- Escalas de medición de las variables (criterios: distancia, proporción y orden natural)
  - Escala de proporción (ejemplo: $\frac{PIB_2}{PIB_1}$).
  - Escala de intervalo $\Longrightarrow$ sólo es significativa la distancia pero no la proporción, ni el orden (Ejemplo: tiempo).
  - Escala ordinal $\Longrightarrow$ Sólo cumplen el orden natural (Ejemplo: sistema de calificación, tipos de ingresos).
  - Escala nominal $\Longrightarrow$ no cumple ninguno de los criterios (ejemplo: sexo, estado civil).

Objetivos del análisis de regresión

- Estimar la media o valor medio de Y dado X. Formalmente, estimamos la media de la distribución condicional de $Y \mid X$.

$$E(Y \mid X)$$

- Contrastar hipótesis (sugeridas por la teoría económica o de negocios).

- Predecir.

Algunos ejemplos

- Estatura padres vs estatura hijos
- Estatura vs edad
- Propensión marginal a consumir
- Elasticidad precio de la demanda
- Curva de phillips
- Mayor inflación, menor proporción de ingreso liquido
- Elasticidad de la demanda ante cambios en los gastos de publicidad
- Retornos salariales de la educación

Ejemplo

- Objeto de estudio: relación entre productividad y educación formal de los trabajadores.
- Según la teoría: más educación, más productividad. Problema: ¿cómo medir la productividad?
- Solución teórica: bajo el supuesto de mercados competitivos.

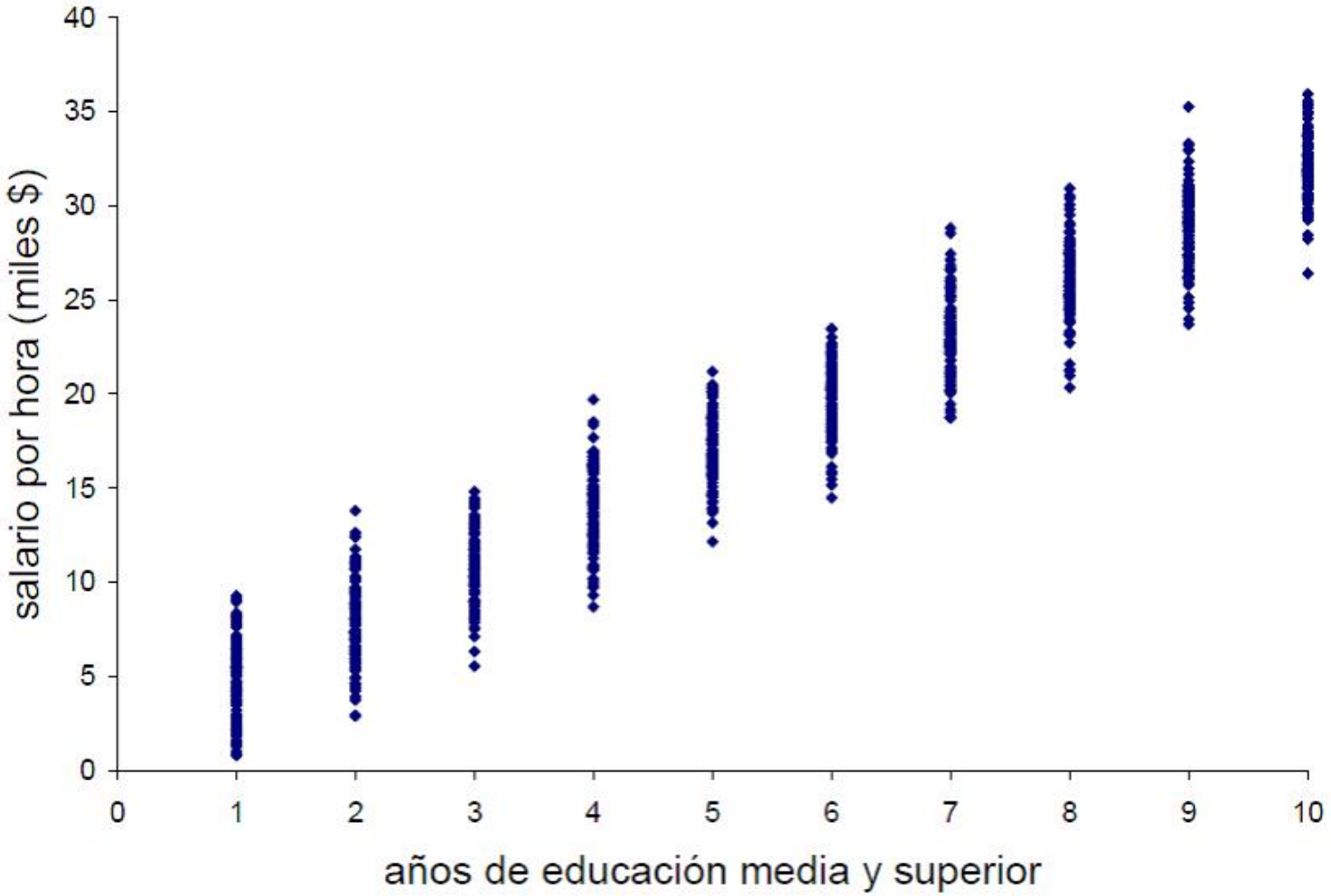
$$PML = W$$

- Estudiaremos la relación entre Y="salario por hora" y X="años de educación media y superior".

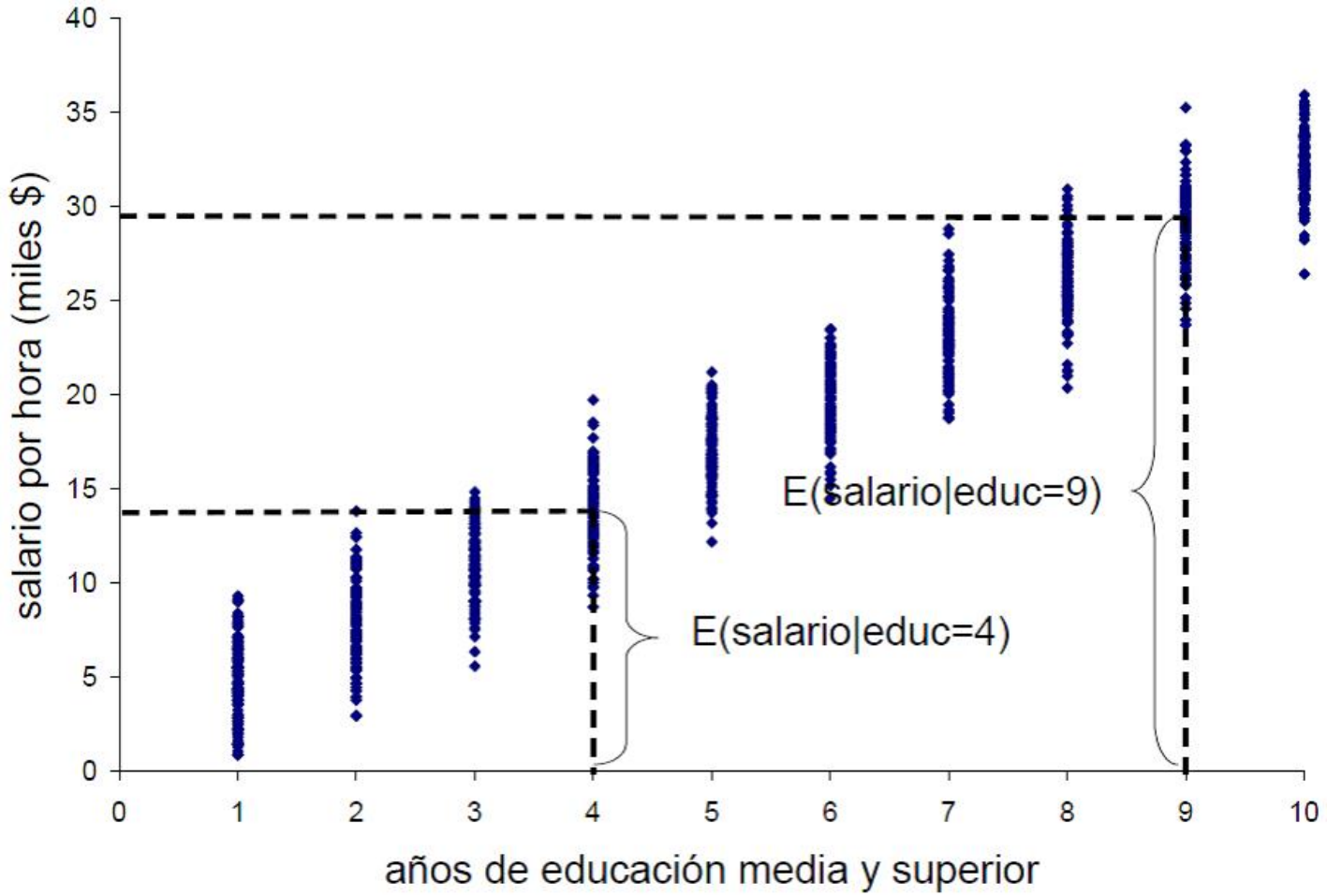
La función de Regresión Poblacional (FRP)

- Población (estadística): todos los pares posibles (educ, salario).
- Suponga que observamos todos los elementos de la población (este es un ejemplo, no un caso real!)
- Dibujamos un gráfico de puntos, cada punto $(X = educ, Y = salario)$
- Formalmente, dibujamos en el espacio (X, Y) la distribución (i.e., población) de la variable bivariante (X, Y).

Gráfico de puntos (población ó distribución de (X,Y))



Obteniendo la FRP: Para cada valor de X (educ) calculamos $E(Y \mid X = x)$



FRP

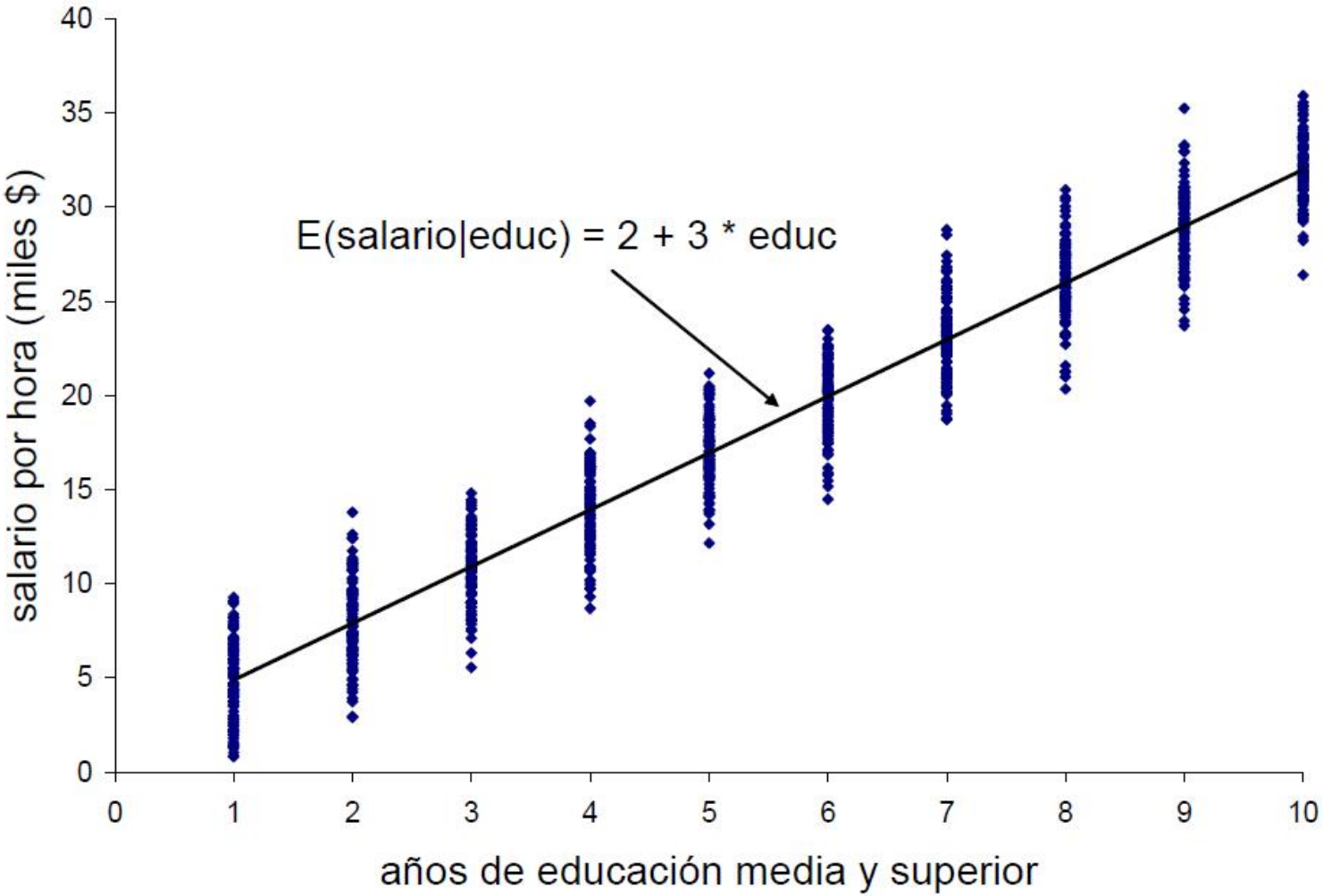
Si unimos con una línea todos los valores medios de Y condicionales a cada valor de X, obtenemos la línea de regresión poblacional. La expresión matemática de esta línea es la FRP. Si, en particular, la relación entre Y y X es lineal, la línea de regresión poblacional es una recta. Por lo tanto, la FRP es la ecuación de esa recta:

$$E(Y \mid X) = \beta_1 + \beta_2 X$$

En nuestro ejemplo:

$$E(Y \mid X) = 2 + 3X$$

FRP (ejemplo)


E(salario|educ) = 2 + 3 * educ
salario por hora (miles $)
años de educación media y superior
0
5
10
15
20
25
30
35
40
0
1
2
3
4
5
6
7
8
9
10

Especificación estocástica de la FRP

Está claro que la FRP no estocástica (como la ecuación de la recta mencionada antes) no describe exactamente la realidad de cada elemento de la población. Es decir, tal como el gráfico de puntos muestra, si tomamos un elemento (individuo) al azar, y resulta que tiene 5 años de educación, sólo por casualidad su salario será $2 + 3 * 5 = 17$.

Especificación estocástica de la FRP

Si la relación entre salario y educación fuera exacta, entonces todos los puntos se situarían sobre la recta $Y = 2 + 3 * X$ (y por lo tanto $Y = E(Y \mid X)$ para todo valor posible de X). Pero, aunque para el PROMEDIO de la población la relación es exacta, para cualquier individuo en particular la relación NO es exacta. En general, podríamos decir que para el individuo i:

$$Y_i = \beta_1 + \beta_2 X_i + \mu_i$$

Especificación estocástica de la FRP

La variable μ_i es una cantidad (aleatoria) que puede ser positiva, negativa o incluso cero. Es decir, el salario del individuo i es igual a la media del salario condicionada al valor de educacion X_i mas o menos alguna cantidad.

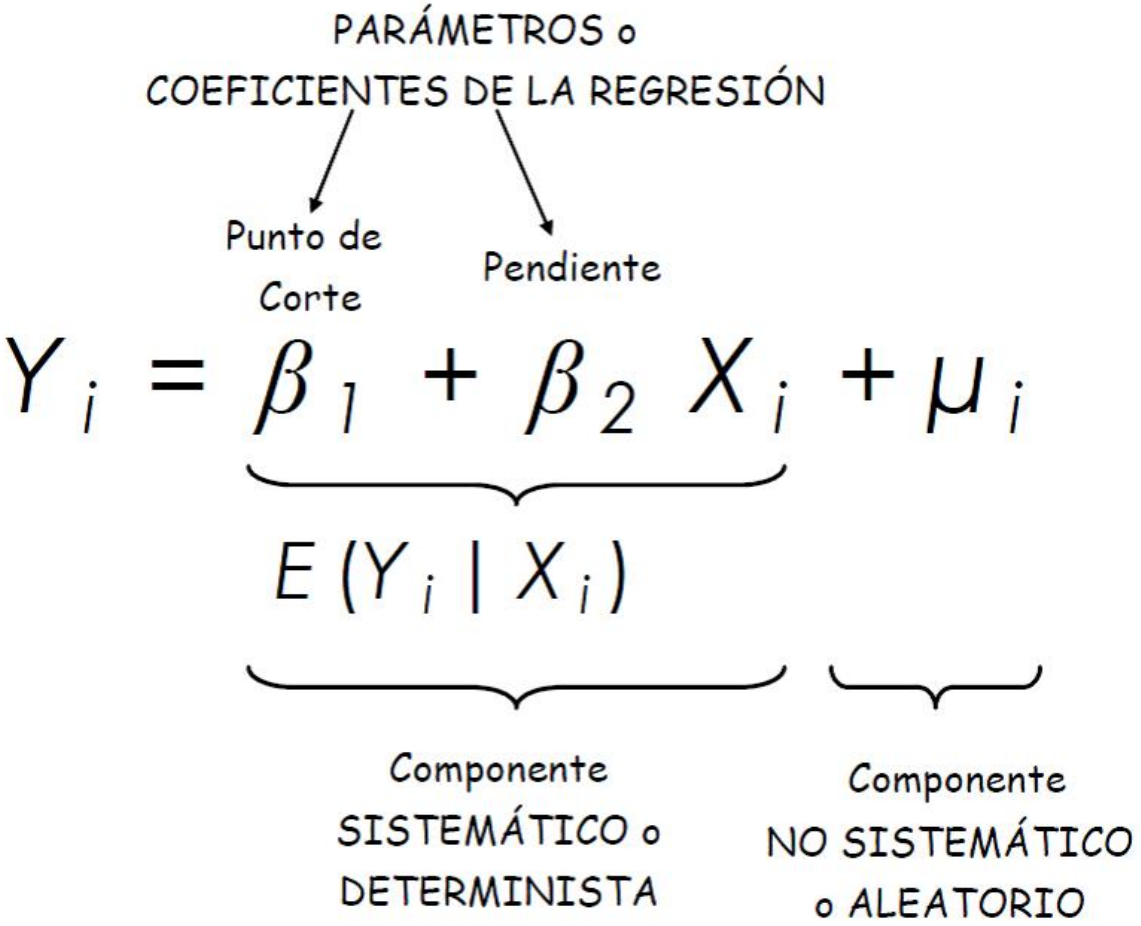
- μ_i: término de error (o perturbación) estocástico o aleatorio.
- Ecuacion $Y_i = \beta_1 + \beta_2 X_i + \mu_i$: FRP estocastica $\Longrightarrow$ nuestro objeto de estudio!!!.

(Si la FRP fuera determinística o no estocástica...no nos harían falta las técnicas de estimación ni de muestreo, ¿por qué?)

Origen o naturaleza del término de error

- Omisión de variables que influencian a Y (porque no son medibles o por parsimonia).
- Errores de medición (errores en los datos!!!).
- Aleatoriedad debida al tipo de fenómenos que estudiamos: el comportamiento humano/social nunca es exacto!!!.

Elementos de la FRP (y de la FRP estocástica)



Modelo de Regresión Lineal Simple

La FRP estocastica $Y_i = \beta_1 + \beta_2 X_i + \mu_i$ define el modelo de regresion lineal simple (MRLS). En la practica, un analisis de regresion lineal simple trata a todos los factores que influyen en Y ademas de X como inobservables, de alli que incluimos μ_i. Si los otros factores (incluidos en μ_i) se mantienen fijos de modo que un cambio en μ_i sea cero, entonces X_i tiene un efecto lineal sobre Y_i , o en otras palabras, la pendiente es una constante

Si:

$$\triangle \mu_i = 0 \Longrightarrow \triangle Y_i = \beta_2 \triangle X_i \Longrightarrow \beta_2 = \frac{\triangle Y_i}{\triangle X_i}$$

Importancia de la pendiente

El coeficiente β_2, la pendiente, mide entonces (bajo el supuesto de que $\triangle \mu_i = 0$) cuánto varía Y_i cuando varía X_i en una unidad. En modelos económicos o de negocios este parámetro es crucial. En nuestro ejemplo, β_2 mide el cambio en el salario por hora con otro año adicional de educación media o superior.

Aplicación: elasticidades

Suponga que Y es el gasto en consumo, X es ingreso. El modelo matemático es:

$$Y = \beta_1 + \beta_2 X$$

Luego, la elasticidad del gasto en consumo con respecto al ingreso está dado por:

$$\epsilon_Y = \frac{dY}{dX}\frac{X}{Y} = \beta_2 \frac{X}{Y}$$

El coeficiente β_2 (si $\triangle \mu_i = 0$) afecta la elasticidad.

Aplicación: elasticidades

Sea el MRLS:

$$lnY = \beta_1 + \beta_2 lnX + \mu$$

$$Y = exp(\beta_1 + \beta_2 lnX + \mu)$$

$$\frac{dY}{dX} = exp(\beta_1 + \beta_2 lnX + \mu)\beta_2 \frac{1}{X}$$

$$\frac{dY}{dX} = \frac{X}{Y}\beta_2 \Longrightarrow \epsilon_Y = \frac{dY}{dX}\frac{X}{Y} = \beta_2$$

Especificación adecuada cuando la elasticidad es constante.

Linealidad

El modelo en logaritmos sigue siendo un modelo de regresión "LINEAL" simple. La razón es porque nos interesa la linealidad con respecto a los coeficientes de la regresión. Metodológicamente, se aplica la misma técnica estimando un modelo lineal del tipo:

$$W = \beta_1 + \beta_2 Z + \mu$$

Donde:

$$W = logY$$

y

$$Z = logX$$

Obviamente, el dibujo en el espacio (X,Y) deja de ser una recta. Hay modelos que no lineales en los parámetros, pero no los estudiaremos aquí!!.

Función de Regresión de la Muestra (FRM)

- Objetivo: Estimar los parámetros o coeficientes de la regresión.
- Si tuviéramos los datos de todos los pares (X,Y) de la distribución (población), calcularíamos $E(Y \mid X = x)$ para cada valor de X y obtendríamos la línea de regresión y luego, la ecuación (FRP).
- Pero en general este no es el caso, y como es usual, disponemos de una muestra.
- Necesitamos, por lo tanto, encontrar un **ESTIMADOR** de la FRP.

Función de Regresión de la Muestra (FRM)

El ESTIMADOR de la FRP se denomina función de regresión de la muestra (FRM):

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

$\hat{\beta}_1$ es el estimador de β_1, $\hat{\beta}_2$ es el estimador de β_2, $\hat{Y}_i$ es el estimador de $E(Y \mid X_i)$

FRM estocástica

La versión estocástica de la FRM es:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\mu}_i$$

Donde $\hat{\mu}_i$ es el estimador de μ_i; se denomina término residual o residuo y se puede definir como:

$$\hat{\mu}_i = Y_i - \hat{Y}_i$$

Objetivo análisis de regresión

Estimar:

$$Y_i = \beta_1 + \beta_2 X_i + \mu_i$$

apartir de:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\mu}_i$$

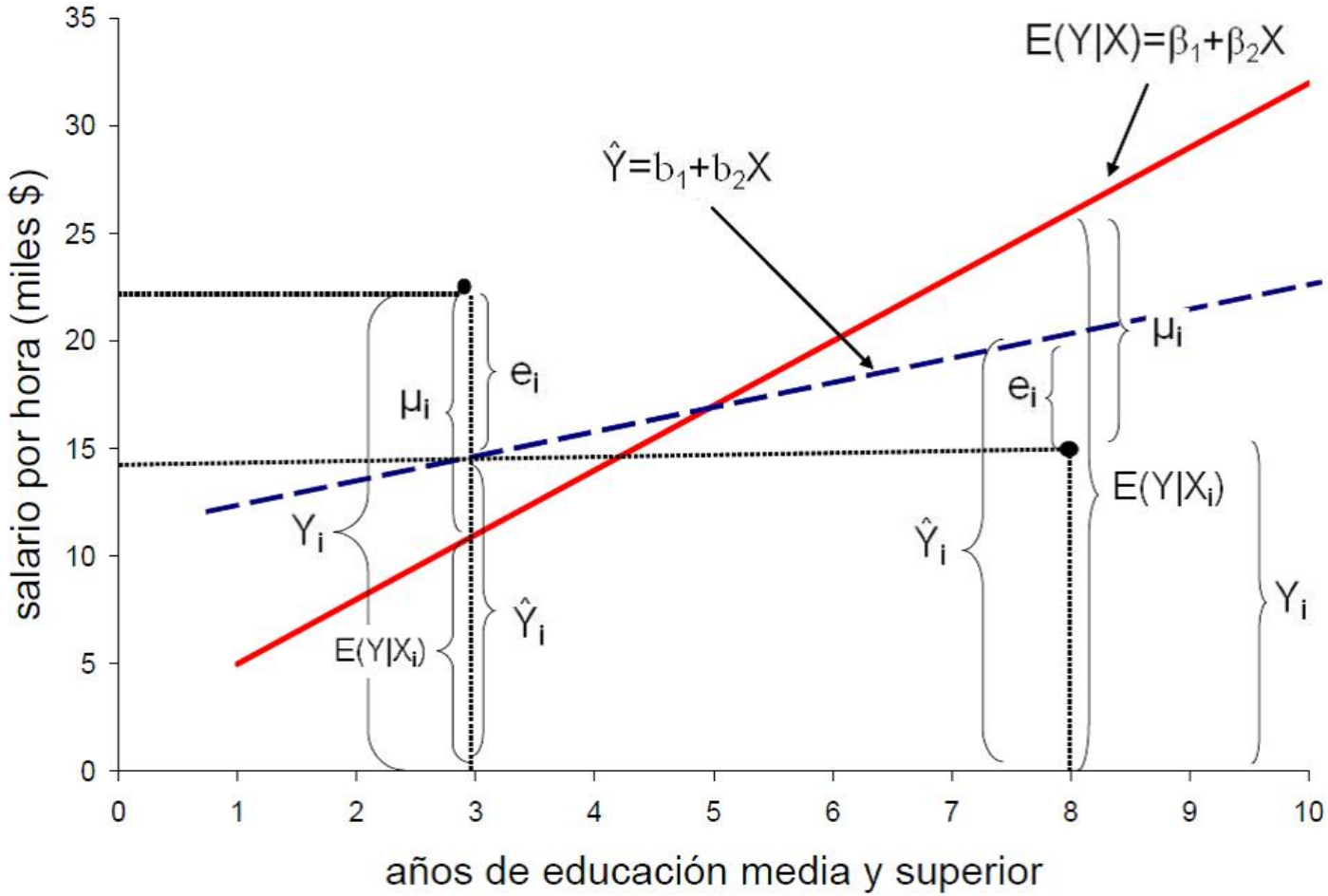
Note que el valor observado Y_i:

$$Y_i = \hat{Y}_i + \hat{\mu}_i$$

$$Y_i = E(Y \mid X_i) + \mu_i$$

Al tomar una muestra particular y obtener valores particulares de $\hat{\beta}_1$ y $\hat{\beta}_2$ obtenemos una estimación de la FRP.

Regresión poblacional y regresión muestral



Sobre la figura anterior...

asuman $\hat{\mu}_i = e_i$. Primer punto (a la izquierda, $X = 3$), $Y_i = \hat{Y}_i + \hat{\mu}_i$ donde $\hat{\mu}_i > 0$, $Y_i = E(Y \mid X = 3) + \mu_i$ donde $\mu_i > 0$. Para el segundo punto (a la derecha, X $= 8$), $\hat{\mu}_i < 0$ al igual que $\mu_i < 0$ IMPORTANTE:

Estos son sólo dos posibles ejemplos. En realidad podríamos tener distintos casos donde dado un par (X_i, Y_i). y los valores correspondientes de $E(Y \mid X)$ y la estimación $\hat{Y}_i$: $\hat{\mu}_i > 0$.

Recuerde...

- La FRP estocástica es una relación teórica entre Y y X que especifica que $E(Y \mid X)$ es una función lineal en parámetros de X y que cada Y_i es igual $E(Y \mid X_i)$ más o menos una cantidad μ_i.
  $Y_i = E(Y \mid X_i) + \mu_i = \beta_1 + \beta_2 X_i + \mu_i$
- No observamos (conocemos) $E(Y \mid X_i)$ ni μ_i (solo sabemos que $\mu_i = Y_i - E(Y \mid Xi)$)..
- La FRM estocástica es el ESTIMADOR de la FRP estocástica.
  $Y_i = \hat{Y}_i + \hat{\mu}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\mu}_i$
- Cuando obtenemos valores particulares de β_1 y β_1, dada una muestra, obtenemos una ESTIMACION de la FRP.
- Al obtener estimaciones de β_1 y β_1, es posible calcular $\hat{\mu}_i = Y_i - \hat{Y}_i$
- Es trivial, pero tenga en cuenta que: $\hat{Y}_i \neq E(Y \mid X)$ y $\hat{\mu}_i \neq \mu_i$!!!
  (es lo mismo que ocurre cuando comparamos la media muestral $\hat{X}$ y la media poblacional μ de una v.a.)