Análisis de regresión

Origen del término *regresión*

- Francis Galton acuñó el término regresión en un ensayo famoso de 1886.

  "La estatura de los niños de padres de determinada estatura tienden a ***regresar*** a la estatura promedio de la población total"

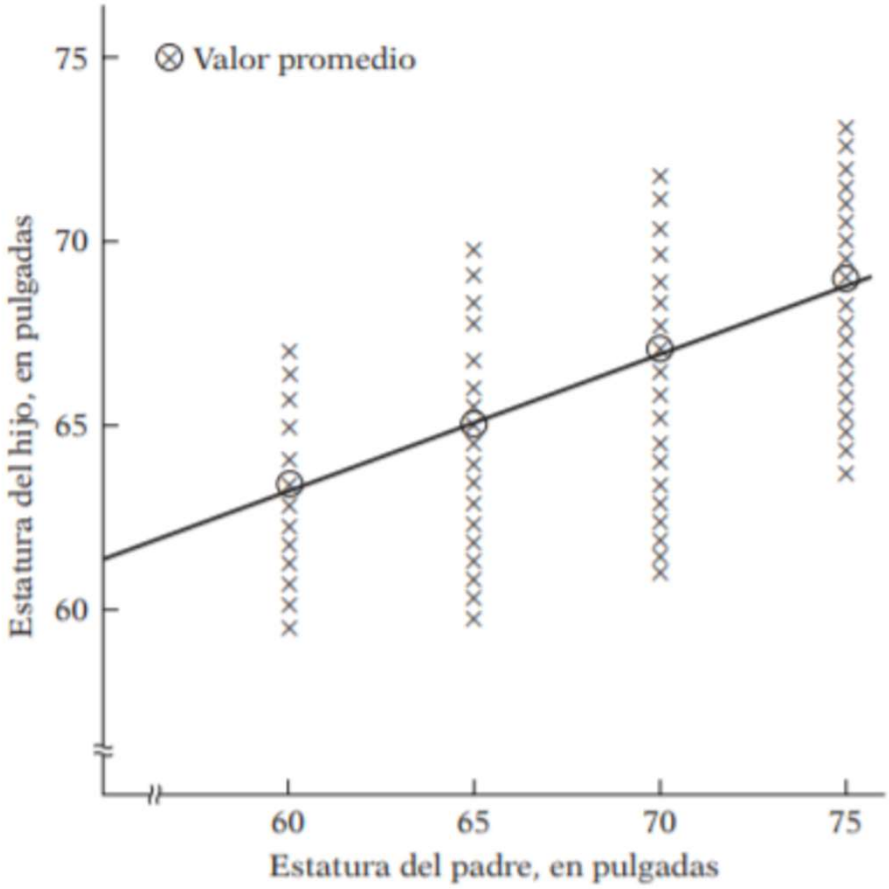
- La ley de regresión universal de Galton fue confirmada por su amigo Karl Pearson, quien reunió más de 1000 registros de estaturas de miembros de grupos familiares, confirmando la teoría de Galton, que llamó "**regresión a la mediocridad**"

Origen del término *regresión*

- La interpretación moderna del término regresión en muy diferente.
- El análisis de regresión trata del estudio de la dependencia de una variable (variable dependiente) respecto de una o más variables (variables explicativas) con el objetivo de estimar o predecir la media o valor promedio poblacional de la variable dependiente en términos de los valores conocidos o fijos (en muestras repetidas) de las variables explicativas.

Enfoque del análisis de regresión

- A Francis Galton le interesaba las razones de estabilidad en la distribución de las estaturas dentro de una población



En el enfoque moderno de regresión, el interés radica en conocer como cambia la estatura promedio de los hijos a partir del conocimiento de la estatura de sus padres, i.e. predecir la estatura de los hijos a partir de la estatura de los padres.

Enfoque del análisis de regresión: ejemplos en economía

- La dependencia del consumo personal respecto del ingreso personal neto disponible (después de impuestos). Con un análisis de este tipo se calcula la propensión marginal a consumir (PMC), i.e. el cambio promedio del consumo ante un cambio de una unidad monetaria en el ingreso real.
- Un monopolista que puede fijar precio o cantidad (pero no ambos factores) y quiera conocer la demanda de un bien con diversos precios. Tal experimento permite estimar la elasticidad precio de la demanda del bien, i.e. la respuesta a variaciones del precio, y permite determinar el precio que maximiza las ganancias.

Enfoque del análisis de regresión: ejemplos en economía

- Cual es la tasa de cambio de los salarios nominales en relación con la tasa de desempleo?. La representación de esta relación es la célebre curva de Phillips, que relaciona los cambios en los salarios nominales con la tasa de desempleo. Un diagrama de dispersión de este tipo permite al economista laboral predecir el cambio promedio en los salarios nominales con una cierta tasa de desempleo. Tal conocimiento sirve para establecer supuestos sobre el proceso inflacionario en una economía, pues es probable que los incrementos en los salarios monetarios se reflejen en incrementos de precios.

- En economía monetaria se sabe que, si se mantienen constantes otros factores, cuanto mayor sea la tasa de inflación π, menor será la proporción k del ingreso que la gente deseará mantener en forma de dinero. Un análisis cuantitativo de esta relación permite predecir la cantidad de dinero, como proporción del ingreso, que la gente deseará mantener con diversas tasas de inflación.

Regresión y causalidad

- El análisis de regresión tiene que ver con la dependencia de una variable respecto de otras variables, pero esto no necesariamente implica causalidad.
- “Una relación estadística, por más fuerte y sugerente que sea, nunca podrá establecer una conexión causal: nuestras ideas de causalidad deben provenir de cuestiones externas y, en último término, de una u otra teoría” (M. G. Kendall y A. Stuart, The Advanced Theory of Statistics)
- **Una relación estadística por sí misma no implica causalidad**: hay que recurrir al sentido común o a cuestiones teóricas.
  - El rendimiento del cultivo depende de la lluvia (sentido común)
  - El consumo depende del ingreso real disponible (teoría económica)

Regresión y correlación

- El análisis de correlación se relaciona con el de regresión, aunque conceptualmente los dos son muy diferentes.
- **En el análisis de correlación**, el objetivo principal es medir la fuerza o el grado de **asociación lineal** entre dos variables. Recordemos que el coeficiente de correlación, mide esta fuerza de asociación (lineal).
- **En el análisis de regresión**, en cambio, se trata de estimar o predecir el valor promedio de una variable con base en los valores fijos en muestras repetidas de otras variables.

Regresión y correlación

- La regresión y la correlación presentan diferencias que vale la pena destacar.
- En el análisis de regresión hay una **asimetría** en el tratamiento a las variables dependientes y explicativas. La variable dependiente es aleatoria o estocástica, i.e., tiene una distribución de probabilidad. Las variables explicativas se asumen que toman valores fijos (en muestras repetidas).
- En el análisis de correlación, se tratan dos variables cualesquiera en forma **simétrica** y se asume que ambas variables son aleatorias.
- La teoría de correlación asume aleatoriedad de las variables.
- Gran parte de la teoría de regresión está condicionada al supuesto de que la variable dependiente es estocástica y que las variables explicativas son fijas o no estocásticas

Terminología en el análisis de regresión

Variable dependiente	**Variable explicativa**
Variable explicada	Variable independiente
Predicha	Predictora
Regresada	**Regresora**
Respuesta	Estímulo
Endógena	Exógena
Resultado	Covariante
Variable controlada	Variable de control

Medida de asociación entre dos variables

- Supongamos que tenemos 2 muestras de tamaño n. Sean $x_1, x_2, \ldots, x_n$ y $y_1, y_2, \ldots y_n$ y $\bar{x}$ y $\bar{y}$ las medias muestrales

- La covarianza muestral se define como

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

- Y el coeficiente de correlación muestral se define como

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\dfrac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1}}{\sqrt{\sum (x_i - \bar{x})^2 / (n-1)} \sqrt{\sum (y_i - \bar{y})^2 / (n-1)}}$$

Medida de asociación entre dos variables

- El coeficiente de correlación toma valores entre -1 y 1.

$$-1 \leq r_{xy} \leq 1$$

- Si el coeficiente de correlación es igual a 1, se tiene una asociación lineal positiva perfecta, intensidad máxima.
- Si el coeficiente de correlación es igual a -1, se tiene una asociación lineal negativa perfecta, intensidad máxima.
- Si el coeficiente de correlación es igual a 0, indica que no hay relación lineal.

Análisis de regresión con dos variables

- El análisis de regresión relaciona la estimación o predicción de la media (de la población) o valor promedio de la variable dependiente, con base en los valores conocidos o fijos de las variables explicativas.

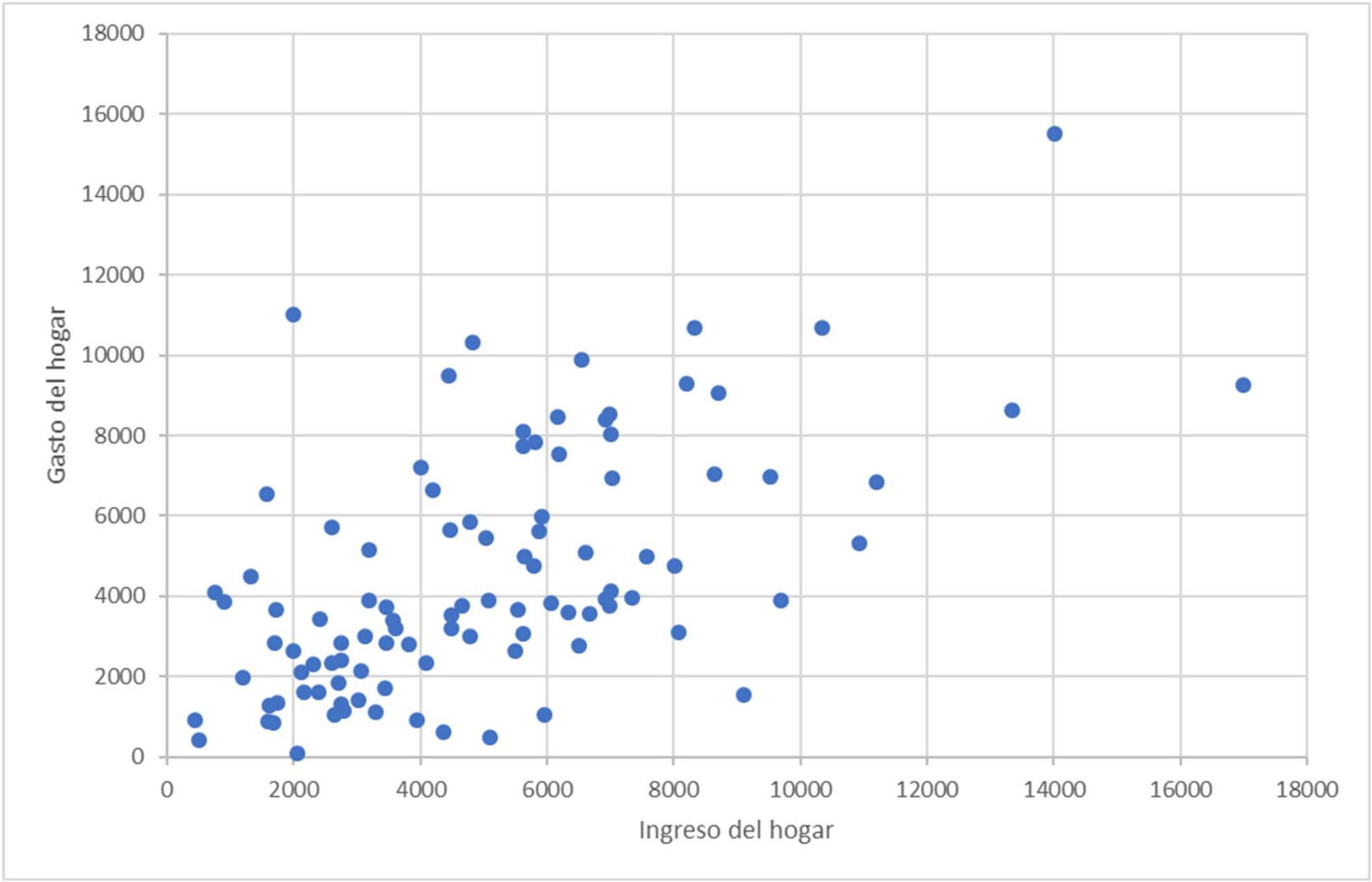
- A modo de ejemplo vamos a trabajar con datos de la última "Encuesta Nacional de Gastos de los Hogares 2012 / 2013" https://www.indec.gob.ar/indec/web/Institucional-Indec-BasesDeDatos-4

- Vamos a considerar la variable **ingreso del hogar**, nuestra variable X (llamada regresor/ variable independiente/ variable explicativa) y el **gasto de hogar**, nuestra variable Y (llamada regresada/ variable independiente/ variable explicada).

Análisis de regresión con dos variables

- Vamos a comenzar tomando una muestra aleatoria de tamaño n=100 de la base de datos.
- Supongamos por el momento que la muestra se trata de la población.
- Dejemos para más adelante como evaluar el impacto de diferentes muestras en nuestras estimaciones.

Diagrama de puntos: Ingreso del hogar vs. gasto del hogar



Esperanza condicional vs. esperanza incondicional

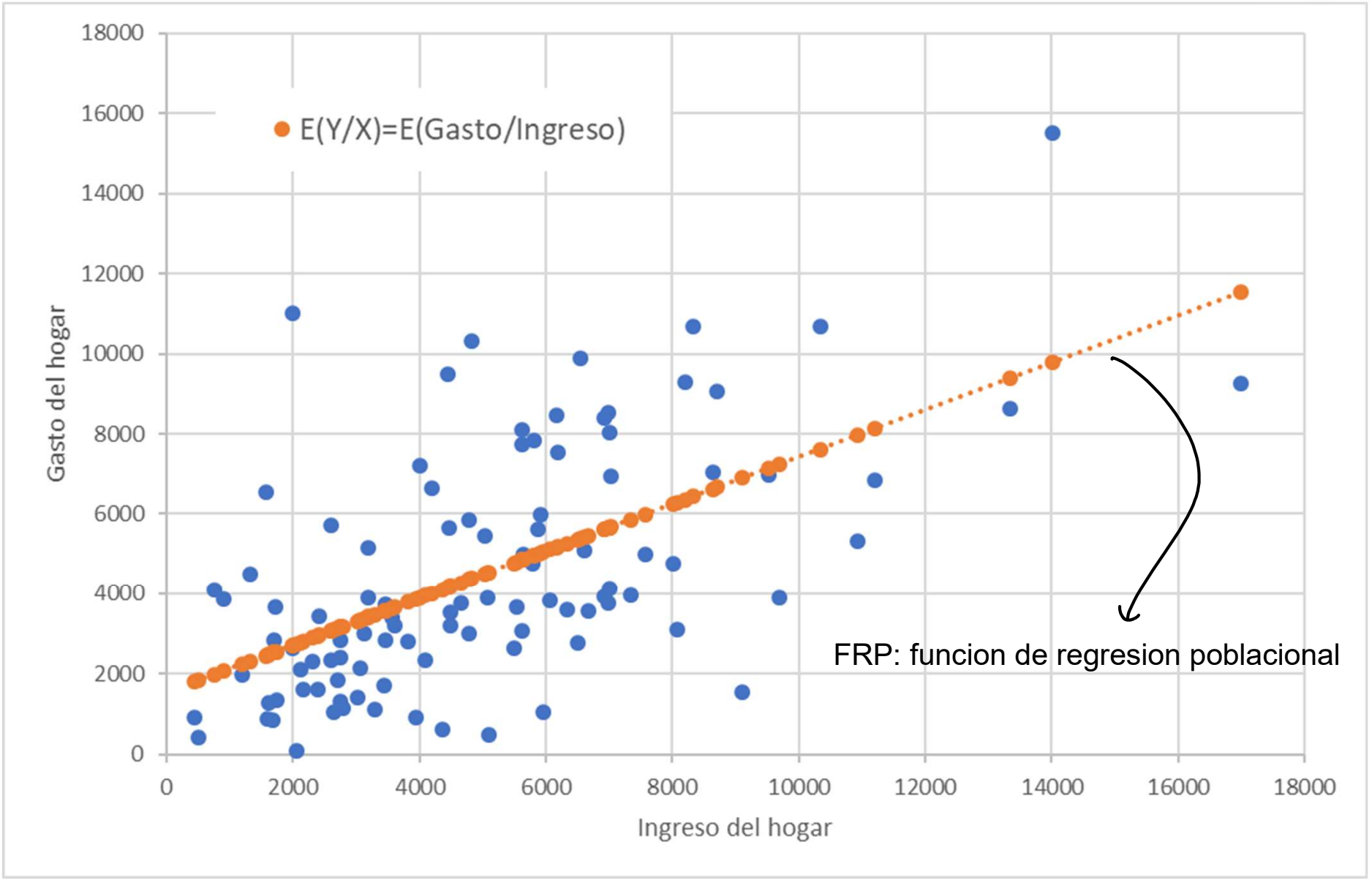
	Ingresos de los hogares						
	< a $2.000	$2.000 - $4.000	$4.000 - $6.000	$6.000 - $8.000	$8.000 - $10.000	> a $10.000	Total muestral
Ingreso promedio	$ 1.291	$ 2.871	$ 5.069	$ 6.756	$ 8.701	$ 12.801	$ 5.018
Gasto promedio	$ 2.553	$ 2.717	$ 4.879	$ 5.728	$ 6.255	$ 9.377	$ 4.488
Cantidad de hogares	13	29	26	17	9	6	100

- ¿Cual es el valor esperado del gasto mensual de un hogar?
- ¿Cual es el valor esperado del gasto mensual de un hogar cuyos ingresos promedios rondan los $5.000?

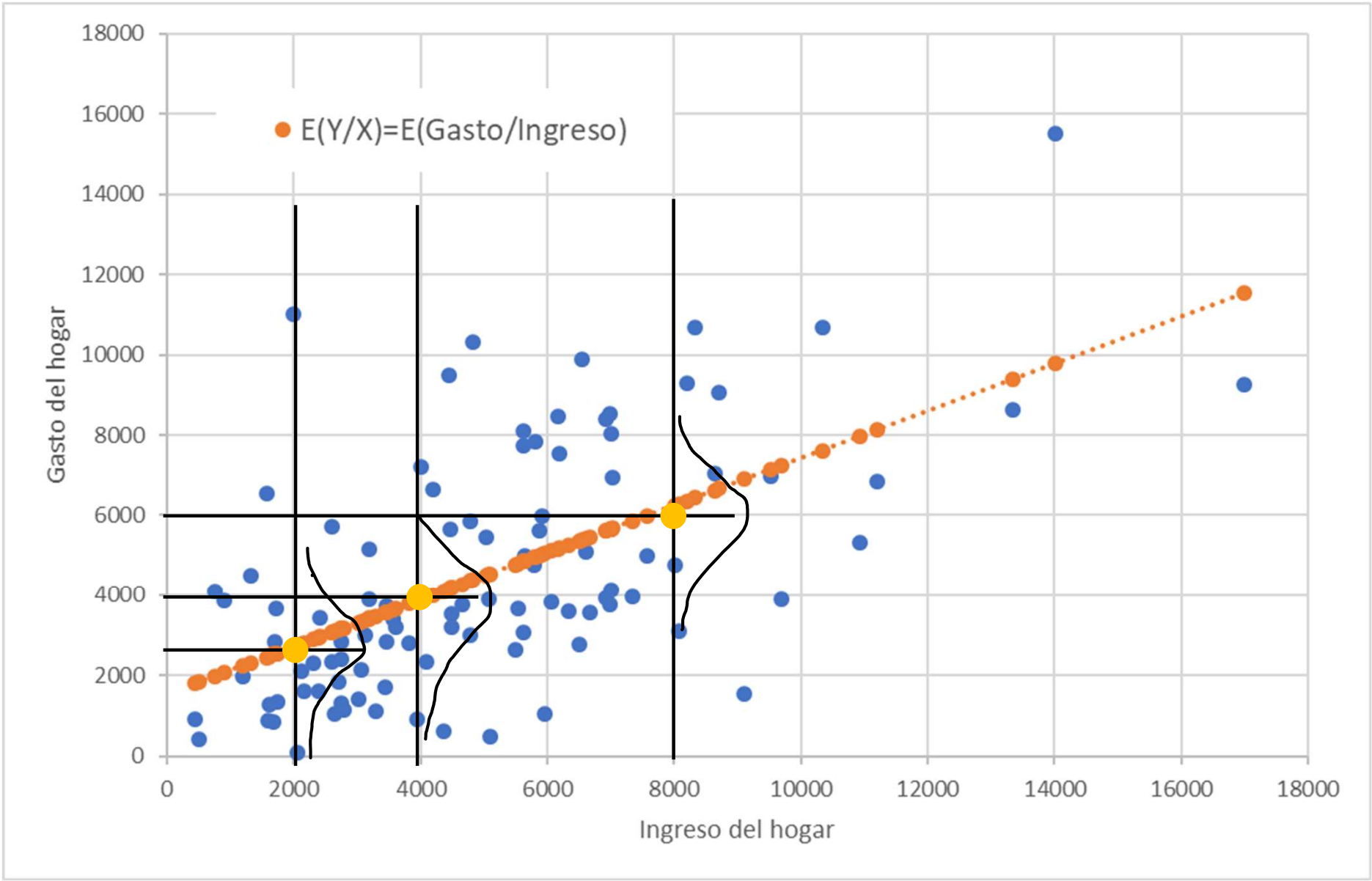
$$E(Y) = E(Gasto) = \ ?$$

$$E(Y/X) = E(Gasto/Ingreso) = \ ?$$

Esperanza condicional vs. esperanza incondicional



Esperanza condicional o media condicional



Función de regresión poblacional (FRP)

- Lo que acabamos de ver nos dice que la esperanza condicional de Y dado X_i es una función de X_i. Matemáticamente se tiene

$$E(Y/X_i) = f(X_i)$$

- donde $f(X_i)$ es una función de X_i
- En el ejemplo que acabamos de ver $E(Y/X_i) = f(X_i)$ es una funcion lineal de X_i .
- A $E(Y/X_i) = f(X_i)$ se la conoce como **función de esperanza condicional** (FEC), **función de regresión poblacional** (FRP) o regresión poblacional (RP).

Función de regresión poblacional (FRP)

- ¿Qué forma adopta la función $f(X_i)$?
- Esta pregunta es importante porque rara vez disponemos de toda la población para efectuar el análisis.
- La forma funcional de la FRP es por consiguiente una pregunta empírica, aunque la teoría puede ayudar.
- En nuestro caso la teoría económica plantea que el consumo tiene una relación lineal con el ingreso, con lo cual podríamos aproximar a la esperanza condicional con

$$E(Y/X_i) = f(X_i) = \beta_1 + \beta_2 X_i$$

- donde β_1 (intercepto) y β_2 (pendiente) son parámetros **no conocidos pero fijos** que se denominan coeficientes de regresión

Función de regresión poblacional (FRP) lineal

- $E(Y/X_i) = f(X_i) = \beta_1 + \beta_2 X_i$ se conoce como **función de regresión poblacional lineal,** modelo de regresión poblacional lineal o simplemente **modelo de regresión lineal.**

- ¿Cual es el objetivo entonces? Estimar los valores desconocidos de β_1 y β_2 en base a los valores observados de Y y X.

- Así como la **media de una v.a. Y es una característica desconocida de una distribución poblacional, la pendiente de la recta que relaciona X e Y es una característica desconocida de la distribución poblacional conjunta de Y y X.**

- El problema estadístico consiste en **estimar esta pendiente**, i. e. estimar el efecto sobre Y de una unidad de cambio en X, usando los datos muestrales de ambas variables.

Modelo de regresión lineal

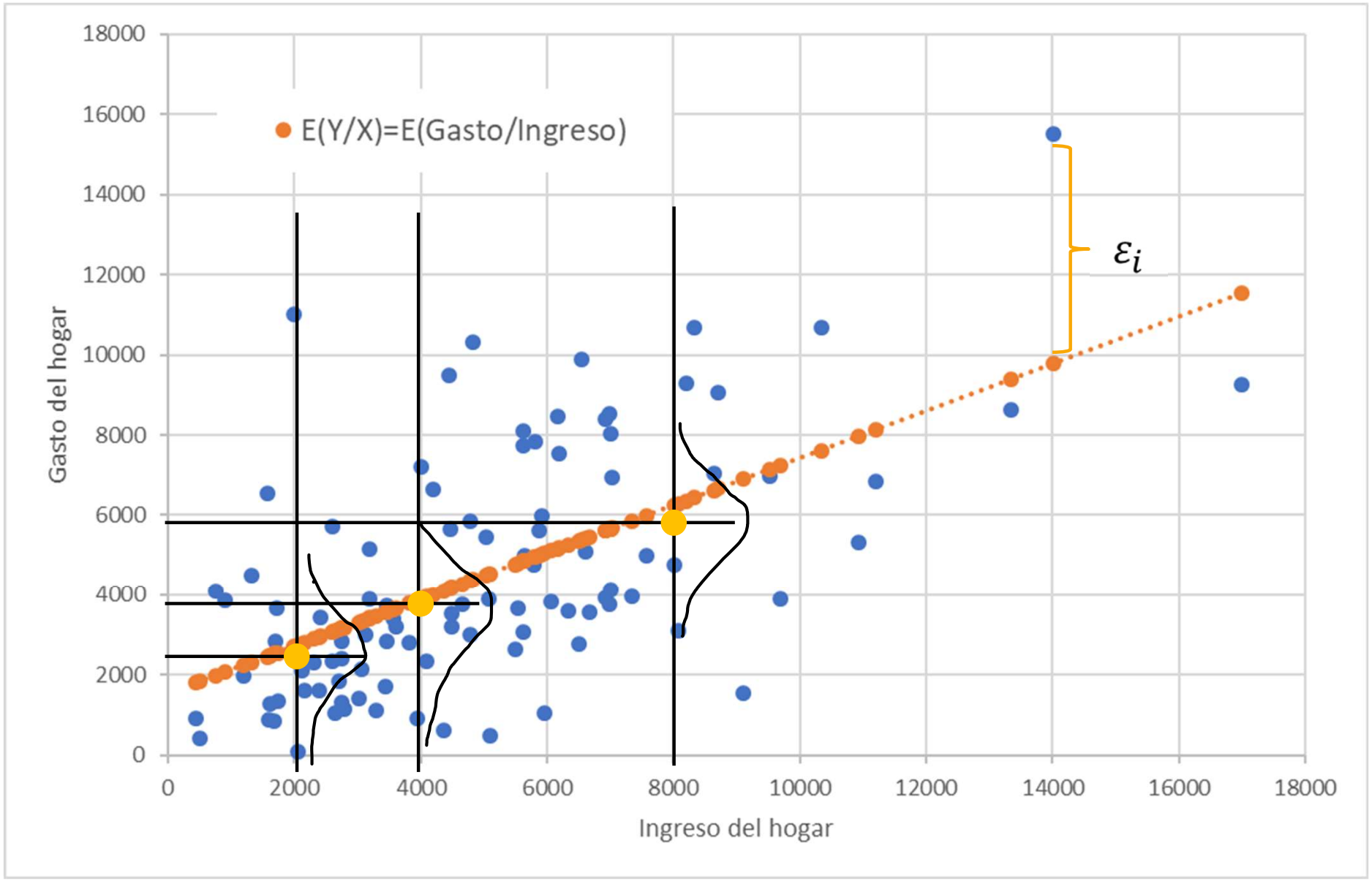
- **Lineal en la variable X**
  - ✓ $E(Y/X_i) = \beta_1 + \beta_2 X_i$ lineal en X
  - ✓ $E(Y/X_i) = \beta_1 + \beta_2 X_i^2$ no es lineal en X

- **Lineal en los parámetros** que es a lo que se refiere el modelo de regresión lineal
  - ✓ $E(Y/X_i) = \beta_1 + \beta_2 X_i$ lineal en los β
  - ✓ $E(Y/X_i) = \beta_1 + \beta_2 X_i^2$ lineal en los β
  - ✓ $E(Y/X_i) = \beta_1 + \beta_2^2 X_i$ no lineal en los β

Función de regresión poblacional (FRP) estocástica

- Es claro que, a medida que aumenta el ingreso familiar, el consumo familiar, en promedio, también aumenta (ver grafico).
- ¿Qué sucede con el consumo de una familia en particular en relación con su nivel de ingreso (fijo)? No siempre aumenta cuando aumenta el ingreso
- Sin embargo, **el consumo promedio** de las familias con ingreso de $6.000 es mayor que **el consumo promedio** de las familias con un ingreso $4.000.
- ¿Qué se puede decir entonces sobre la relación entre el consumo de una familia y un nivel determinado de ingresos?
- Para un nivel de ingresos dados X_i el consumo tiende a agruparse alrededor de su esperanza condicional.

Esperanza condicional o media condicional



Función de regresión poblacional (FRP) estocástica

$$Y_i = E(Y/X_i) + \varepsilon_i = \beta_1 + \beta_2 X_i + \varepsilon_i$$

$$\text{ó} \quad \varepsilon_i = Y_i - E(Y/X_i)$$

- ¿Qué representa ε_i ?
- Recordemos que estamos tratando de predecir el consumo promedio a partir del nivel de ingreso promedio. Pueden existir otros factores que también determinen el nivel de consumo, además del ingreso. Consideremos a todos estos factores juntos y llamémoslo ε_i.
- Técnicamente, ε_i se conoce como **perturbación estocástica** o **término de error estocástico**.
- ε_i es una variable aleatoria **no observable** que toma valores positivos o negativos.

Función de regresión poblacional (FRP) estocástica

- ¿Cómo se interpreta la ecuación $Y_i = E(Y/X_i) + \varepsilon_i$?

- Se puede decir que el gasto de una familia en particular, según su nivel de ingreso, se expresa como la suma de dos componentes:

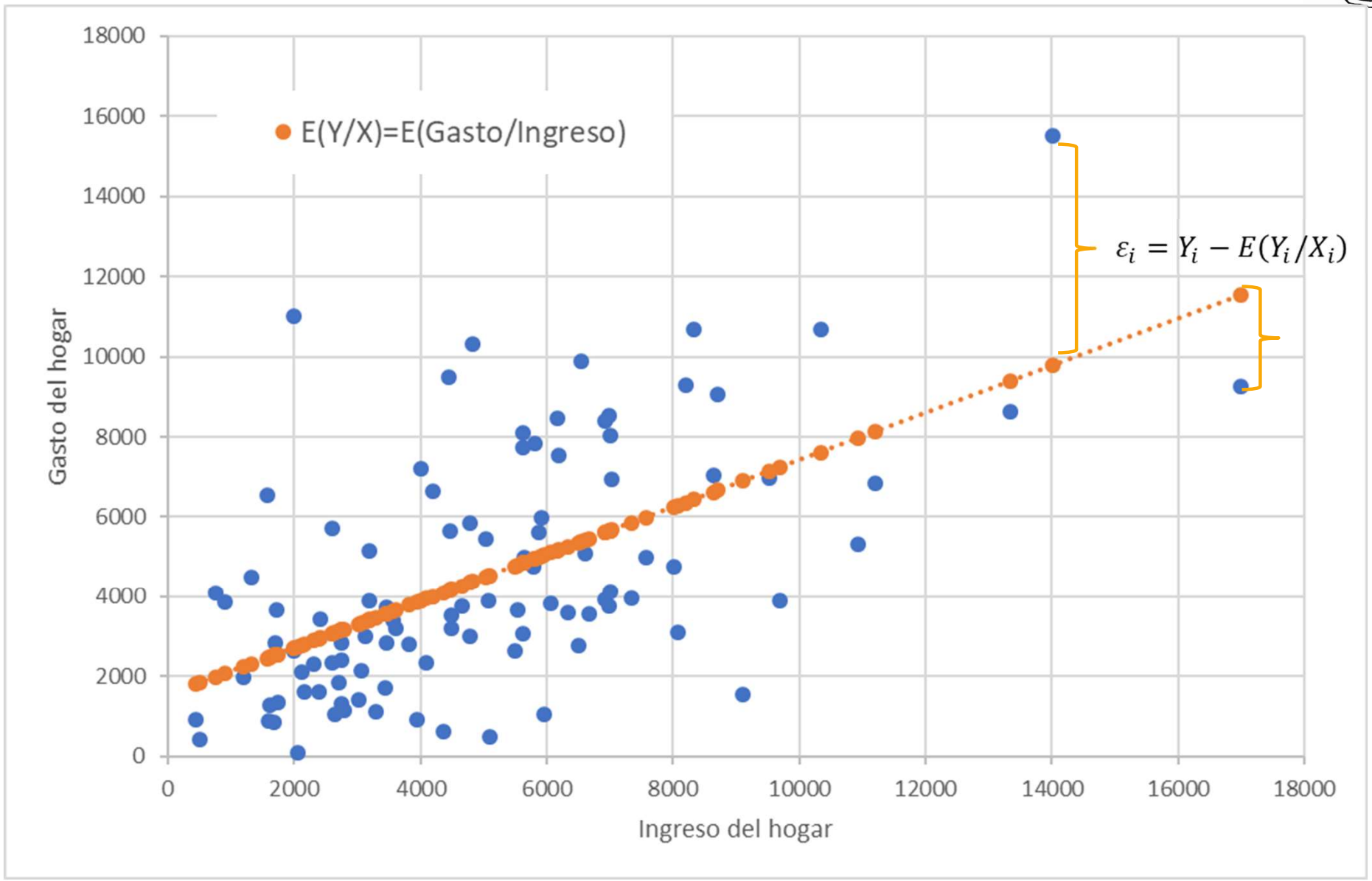
  1. $E(Y/X_i)$, que es simplemente el consumo promedio de todas las familias con el mismo nivel de ingreso. Este componente se conoce como **componente sistemático**.

  2. ε_i que es el componente aleatorio, o no sistemático.

- Si además suponemos que la $E(Y/X_i)$ es lineal en X_i, entonces se tiene

$$Y_i = E(Y/X_i) + \varepsilon_i$$
$$= \beta_1 + \beta_2 X_i + \varepsilon_i$$

La perturbación estocástica o el término de error



La perturbación estocástica o el término de error

- El término de error ε_i resume también a todas las variables que se omiten en el modelo, pero que, en conjunto, afectan a Y.
- La pregunta obvia es: ¿por qué no se introducen explícitamente estas variables en el modelo?
- O de otra forma, ¿por qué no se crea un modelo de regresión múltiple con tantas variables como sea posible?
- Se podría intentar tener un modelo más sofisticado pero no es sencillo.

La perturbación estocástica o el término de error

1. Vaguedad de la teoría
2. Falta de disponibilidad de datos
3. Variables centrales y variables periféricas
4. Aleatoriedad intrínseca en el comportamiento humano
5. Variables relevantes o proxy inadecuadas
6. Principio de parsimonia (principio de la navaja de Occam)
7. Forma funcional incorrecta.

Por todas estas razones, las perturbaciones estocásticas ε_i asumen un rol importante en el análisis de regresión.

Modelo de regresión lineal con una variable: resumen

- El modelo de regresión lineal es $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$ donde
- $Y_i = es\ la\ variable\ dependiente\ o\ explicada\ o\ variable\ del\ lado\ izquierdo$
- $X_i = es\ la\ variable\ independiente\ o\ explicativa\ o\ variable\ del\ lado\ derecho$
- $\beta_1 + \beta_2 X_i = es\ la\ recta\ de\ regresión\ poblacional\ o\ FRP$
- $\beta_1 = es\ el\ intercepto\ de\ la\ recta\ de\ regresión\ poblacional$
- $\beta_2 = es\ la\ pendiente\ de\ la\ recta\ de\ regresión\ poblacional$
- $\varepsilon_i = es\ la\ perturbación\ estocástica\ o\ término\ de\ error$

Modelo de regresión lineal con una variable: interpretación de los parámetros

- La pendiente β_2 mide el cambio marginal en Y asociado a una unidad de cambio en X.

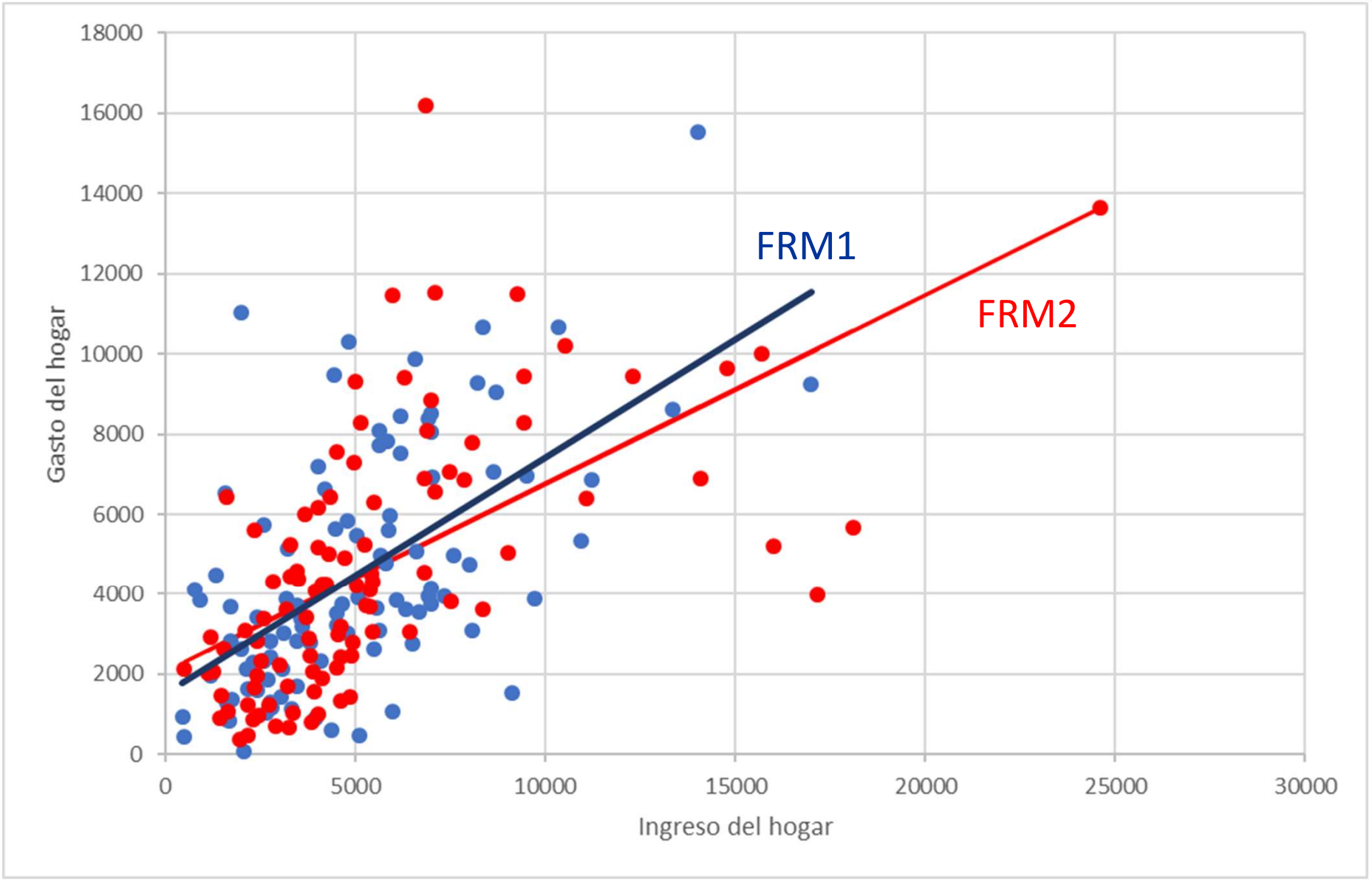
- El intercepto, β_1 , es el valor de la recta de regresión cuando $X = 0$. En algunas regresiones el intercepto tiene una interpretación útil. En otras (como en este caso) no tiene ningún significado económico. En estos casos se lo piensa desde la matemática, como el coeficiente que determina el nivel de la recta de regresión.

- El otro factor es el ε_i, o término de error. En este caso incorporó todos los otros factores que afectan los gastos y que no dependen del ingreso. Se trata de un término **no observable**.

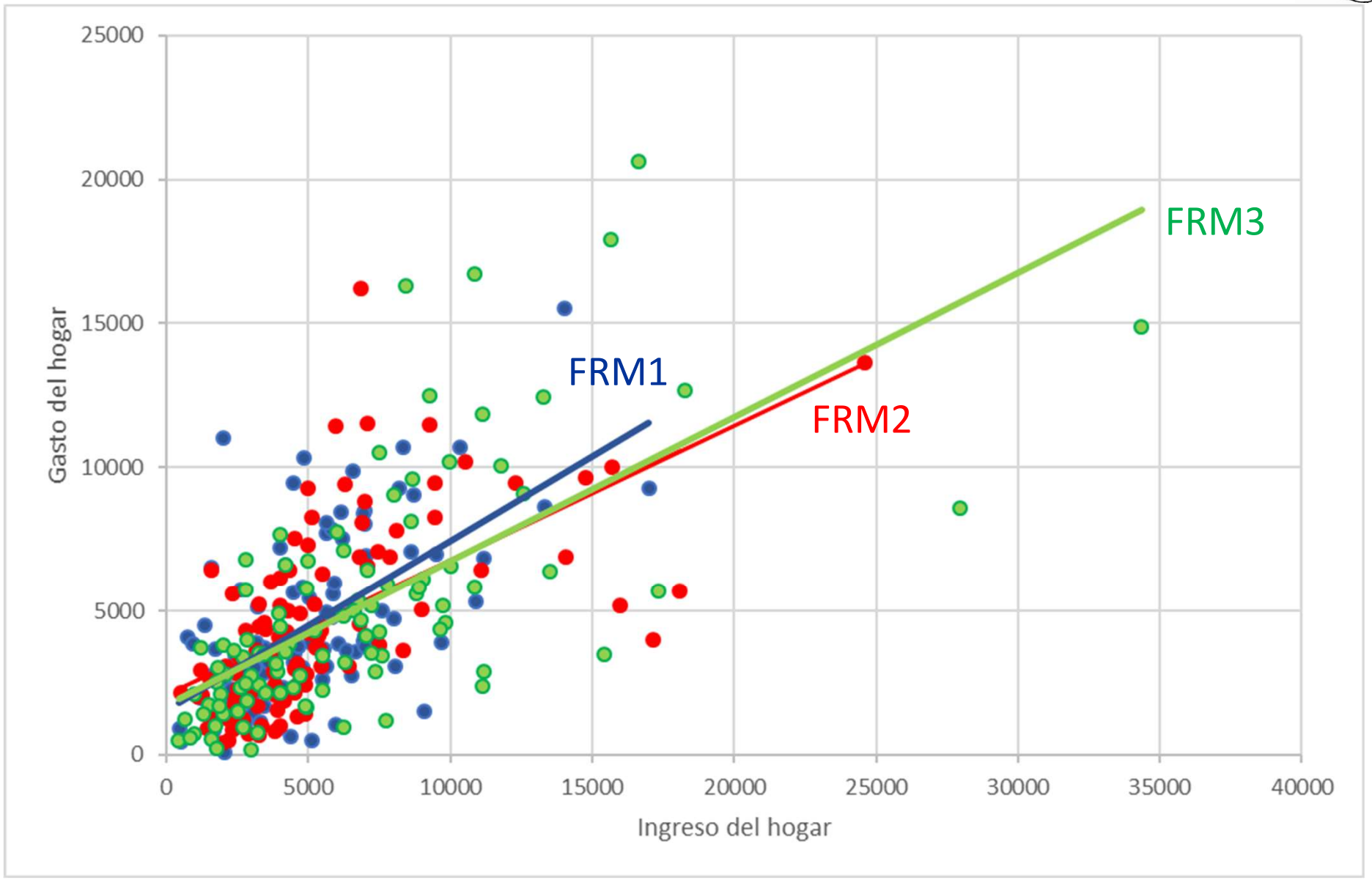
Función de regresión muestral

- Hasta ahora estuvimos trabajando con una muestra tomada al azar, pero asumimos que se trataba de la población.
- La realidad es que la población no la conocemos. Y lo que estamos tratando de estimar es la función (recta) de regresión poblacional (FRP) que desconocemos.
- Lo mejor que podemos hacer es estimar la FRP mediante una muestra, a la que llamaremos función (recta) de regresión muestral (FRM).
- ¿Se puede estimar la FRP a partir de los datos de la muestra?
- ¿Será precisa esta estimación de la FRP? Incertidumbre muestral.
- Tomemos otra muestra aleatoria de n=100 y veamos que ocurre.

Función de regresión muestral



Función de regresión muestral



Función de regresión muestral

- Esta $Y = E(Y/X_i) = \beta_1 + \beta_2 X_i$ es la FRP.
- La contraparte muestral puede escribirse como $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$
- donde $\hat{Y}_i = estimador\ de\ E(Y/X_i)$

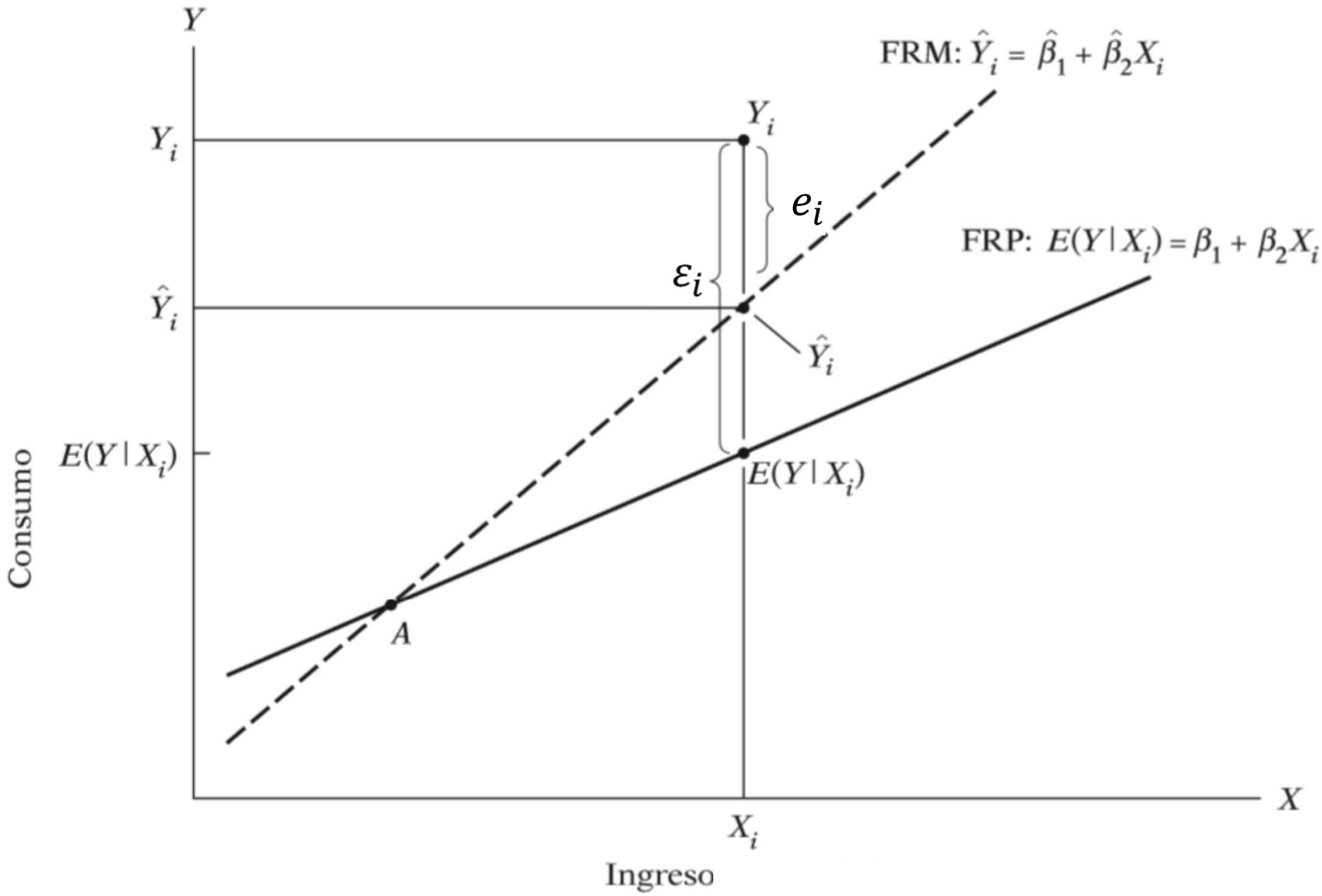
  $\hat{\beta}_1 = estimador\ de\ \beta_1$

  $\hat{\beta}_2 = estimador\ de\ \beta_2$
- Recordemos que un estimador no es otra cosa que una función de la muestra.
- Y si calculamos la diferencia entre el valor observado del gasto Y_i y el valor predicho por la FRM, $\hat{Y}_i$, se tiene el error o residuo que llamaremos $e_i = Y_i - \hat{Y}_i = Y_i - (\hat{\beta}_1 + \hat{\beta}_2 X_i)$

Función de regresión muestral

- Aclaración sobre el error o residuo e_i
- El residuo e_i, no es la perturbación estocástica o término de error del modelo, ε_i, sino que se trata de una **medida combinada** de error del modelo y de los errores que se deben a que $\hat{\beta}_1$ y $\hat{\beta}_2$ son estimaciones muestrales y, por ende están sujetas a la variación debida al azar, lo cual a su vez le otorga variación al valor predicho $\hat{Y}_i$.

Función de regresión muestral



Resultados de la regresión lineal

- La regresión lineal brinda dos resultados importantes:

1. Los valores predichos $\hat{Y}_i$ de la variable dependiente en función de la variable independiente X_i.

2. El cambio marginal de la variable dependiente (gasto) que reporta $\hat{\beta}_2$ ante un cambio unitario de la variable independiente (ingreso).

Estimación: el método de cuadrados mínimos (MCO)

- Dos son los métodos de estimación mas frecuentes:
  1. El método de mínimos cuadrados ordinarios (MCO) o “ordinary least squares” (OLS).
  2. El método de máxima verosimilitud o “maximun likelihood”.
- Bajo determinadas condiciones ambos métodos arrojan los mismos estimadores para $\hat{\beta}_1$ y $\hat{\beta}_2$
- Vamos a utilizar el MCO para estimar $\hat{\beta}_1$ y $\hat{\beta}_2$. Este método fue desarrollado por Gauss (S. XVIII) y es el método de estimación mas utilizado en la práctica. Además es un método que tiene propiedades deseables desde el punto de vista estadístico (consistencia e insesgadez de $\hat{\beta}_1$ y $\hat{\beta}_2$).

Estimación: el método de cuadrados mínimos (MCO)

- Recordemos que la FRP es $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$.
- Sin embargo la FRP no es observable y debemos estimarla mediante una muestra con la FRM,

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$$

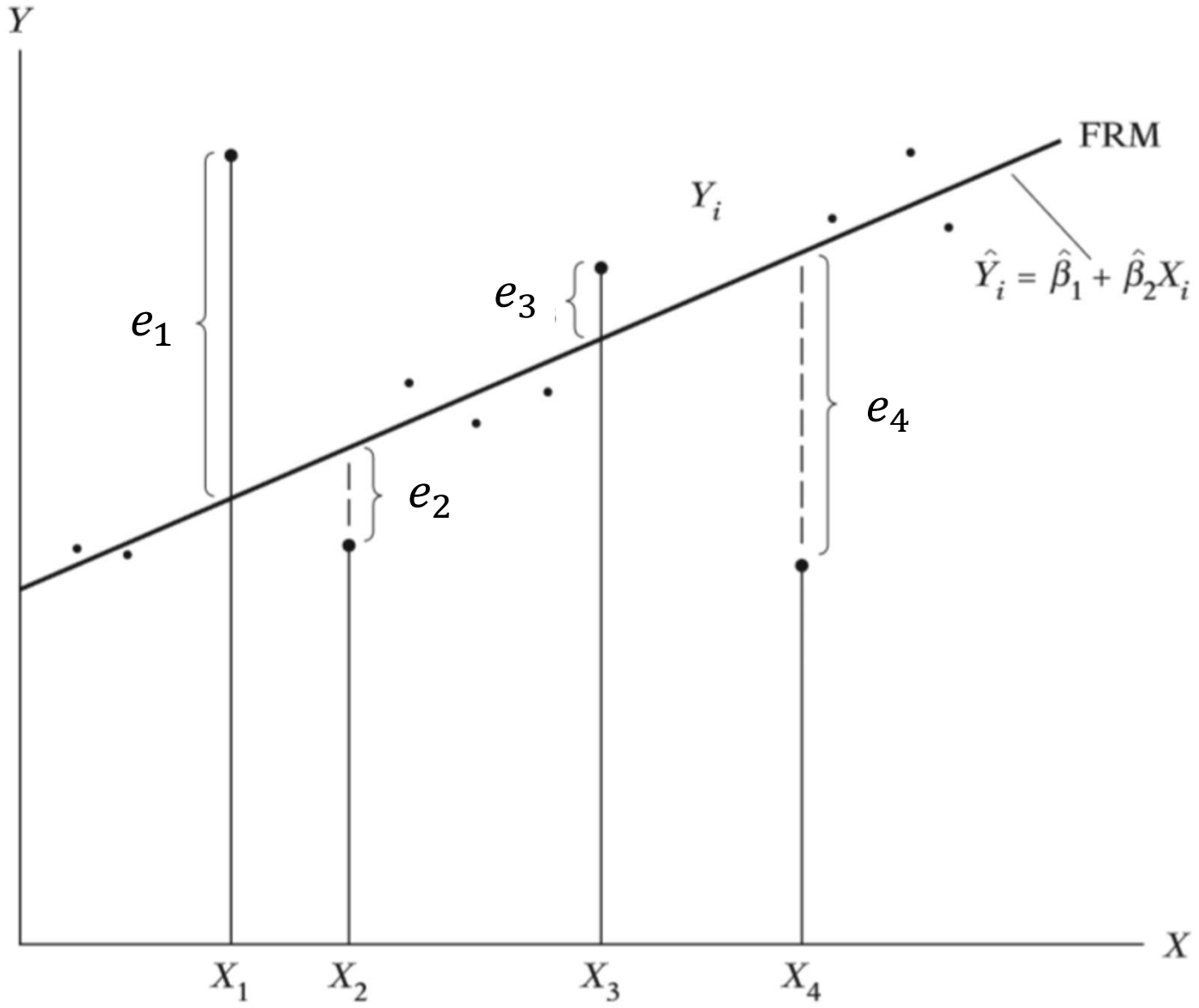
$$Y_i = \hat{Y}_i + e_i$$

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

$$e_i = Y_i - \left(\hat{\beta}_1 + \hat{\beta}_2 X_i\right)$$

$$e_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$$

Función de regresión muestral



Estimación: el método de cuadrados mínimos (MCO)

$$\text{minimizar } \sum_{i=1}^{n} e_i^2 = \text{minimizar} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

$$\text{minimizar } \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)^2$$

- Se deriva respecto de los valores desconocidos $\hat{\beta}_1$ y $\hat{\beta}_2$ y se iguala a cero

$$\frac{\partial\left(\sum_{i=1}^{n} e_i^2\right)}{\partial \hat{\beta}_1} = -2 \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right) = 0$$

$$\frac{\partial\left(\sum_{i=1}^{n} e_i^2\right)}{\partial \hat{\beta}_2} = -2 \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right) X_i = 0$$

Estimación: el método de cuadrados mínimos (MCO)

- Con un poco de álgebra se llega a que la solución del sistema de ecuaciones normales es:

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{S_{XY}}{S_X^2}$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$$

- Además podemos reescribir $\hat{\beta}_2$ en término del coeficiente de correlación

$$\hat{\beta}_2 = \frac{S_{XY}}{S_X^2} = \frac{S_{XY} S_Y}{S_X^2 S_Y} = \frac{S_{XY} S_Y}{S_X S_Y S_X} = r_{XY} \frac{S_Y}{S_X}$$

Estimación: el método de cuadrados mínimos (MCO)

- Notar que la recta de regresión siempre pasa por $(\bar{X}, \bar{Y})$
- Recordemos que $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$
- Sustituyendo $\hat{\beta}_1$ en la recta de regresión $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ se tiene

$$\hat{Y}_i = \bar{Y} - \hat{\beta}_2 \bar{X} + \hat{\beta}_2 X_i$$

$$\hat{Y}_i = \bar{Y} + \hat{\beta}_2 (X_i - \bar{X})$$

$$\hat{Y}_i - \bar{Y} = \hat{\beta}_2 (X_i - \bar{X})$$

- Por lo tanto cuando $X_i = \bar{X}$ resulta que $\hat{Y}_i = \bar{Y}$, y por ende la ecuación de regresión siempre pasa $(\bar{X}, \bar{Y})$

Función de regresión muestral: ejemplo muestra 1

Dependent Variable: GASTOS_M1
Method: Least Squares
Date: 10/10/19 Time: 19:36
Sample: 1 100
Included observations: 100

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1534.152	470.3554	3.261686	0.0015
INGRESO_M1	0.588552	0.080329	7.326786	0.0000

R-squared	0.353911	Mean dependent var	4487.571
Adjusted R-squared	0.347318	S.D. dependent var	3000.105
S.E. of regression	2423.748	Akaike info criterion	18.44381
Sum squared resid	5.76E+08	Schwarz criterion	18.49592
Log likelihood	-920.1907	Hannan-Quinn criter.	18.46490
F-statistic	53.68180	Durbin-Watson stat	2.273001
Prob(F-statistic)	0.000000		

Función de regresión muestral: ejemplo muestra 1

Dependent Variable: GASTOS_M1
Method: Least Squares
Date: 10/10/19 Time: 19:36
Sample: 1 100
Included observations: 100

$\hat{\beta}_1$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1534.152	470.3554	3.261686	0.0015
INGRESO_M1	0.588552	0.080329	7.326786	0.0000

R-squared	0.353911	Mean dependent var	4487.571
Adjusted R-squared	0.347318	S.D. dependent var	3000.105
S.E. of regression	2423.748	Akaike info criterion	18.44381
Sum squared resid	5.76E+08	Schwarz criterion	18.49592
Log likelihood	-920.1907	Hannan-Quinn criter.	18.46490
F-statistic	53.68180	Durbin-Watson stat	2.273001
Prob(F-statistic)	0.000000		

$\hat{\beta}_2$

- ¿Cómo se interpreta $\hat{\beta}_2$?
- ¿Cómo se obtiene $\hat{Y}_i$?

Función de regresión muestral: ejemplo muestra 2

Dependent Variable: GASTOS_M2
Method: Least Squares
Date: 10/08/19 Time: 19:17
Sample: 1 100
Included observations: 100

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	2062.348	434.3770	4.747828	0.0000
INGRESO_M2	0.470420	0.063727	7.381856	0.0000

R-squared	0.357342	Mean dependent var	4644.695
Adjusted R-squared	0.350785	S.D. dependent var	3195.839
S.E. of regression	2575.012	Akaike info criterion	18.56489
Sum squared resid	6.50E+08	Schwarz criterion	18.61700
Log likelihood	-926.2447	Hannan-Quinn criter.	18.58598
F-statistic	54.49179	Durbin-Watson stat	2.086219
Prob(F-statistic)	0.000000		

Función de regresión muestral: ejemplo muestra 3

Dependent Variable: GASTOS_M3
Method: Least Squares
Date: 10/08/19 Time: 19:20
Sample: 1 100
Included observations: 100

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1736.649	484.1497	3.587007	0.0005
INGRESO_M3	0.501499	0.057249	8.759908	0.0000

R-squared	0.439154	Mean dependent var	5013.232
Adjusted R-squared	0.433431	S.D. dependent var	4083.877
S.E. of regression	3073.966	Akaike info criterion	18.91912
Sum squared resid	9.26E+08	Schwarz criterion	18.97123
Log likelihood	-943.9561	Hannan-Quinn criter.	18.94021
F-statistic	76.73599	Durbin-Watson stat	2.228648
Prob(F-statistic)	0.000000		

Fundamentos del modelo de regresión

- Notar que la estimación puntual de $\hat{\beta}_1$ y $\hat{\beta}_2$ que brinda el método de MCO, fue solo desarrollos matemáticos. La estadística no intervino en la obtención de $\hat{\beta}_1$ y $\hat{\beta}_2$.

- Pero si el objetivo es algo más que obtener estimadores puntuales para $\hat{\beta}_1$ y $\hat{\beta}_2$, y lo que se desea es inferir sobre los parámetros poblaciones desconocidos β_1 y β_2 a partir de sus contrapartes muestrales, entonces vamos a tener que recurrir a mas información respecto de la forma en que se generan los Y_i.

- Pero $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$, entonces depende de X_i y ε_i .

- Por lo tanto hay que hacer supuestos sobre la forma en que se generan X_i y ε_i.

Fundamentos del modelo de regresión

- Se trata de los supuestos del modelo de Gauss Markov.
- Estos supuestos parecen bastante abstractos. Pero tratar de entenderlos es escencial para comprender cuando MCO arroja estimaciones de los coeficientes de regresión de utilidad.

Fundamentos del modelo de regresión

- Supuesto #1: La distribución condicional de ε_i dado X_i tiene media cero, i.e. $E(\varepsilon_i/X_i) = 0$ y $E(\varepsilon_i) = 0$ si las X_i son no estocásticas.

- Supuesto #2: (X_i, Y_i) es una muestra i.i.d. para i=1,…,n. extraída de la distribución conjunta de X_i e Y_i

- Supuesto #3: X_i e ε_i tienen momentos finitos de orden 4, i.e. $0 < E(X_i^4) < \infty$ y $0 < E(\varepsilon_i^4) < \infty$. Este supuesto limita la probabilidad de tener valores extremos de (X_i, Y_i).

- Supuesto #4: Los ε_i son variables aleatorias que tienen media 0 y varianza constante. $E(\varepsilon_i) = 0$ y $Var(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$

El poder explicativo del modelo de regresión: el R^2

- ¿Cuan bien la regresión estimada describe los datos?
- La variable independiente o regresor, ¿da cuenta de mucha o poca variabilidad en la variable dependiente?
- Las observaciones muestrales ¿están agrupadas alrededor de la recta de regresión, o por el contrario están todas dispersas?
- El R^2 y el **error estándar de la regresión** (o la varianza de la regresión, o la varianza del error del modelo) son una indicación de cuan bien la recta de regresión ajusta a los datos observados.

El poder explicativo del modelo de regresión: el R^2

- El R^2 es una medida que varía entre 0 y 1, y da cuenta de la fracción de la variabilidad de Y_i que es explicada por X_i.
- Las definiciones de valor predicho y residuo nos permiten expresar a Y_i como la suma de (ver en el gráfico)

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

- Una de las propiedades de MCO es que el promedio muestral de los residuos e_i es 0.
- Además el promedio de los $\hat{Y}_i$ es igual a $\bar{Y}$
- El punto $(\bar{X}, \bar{Y})$ está siempre sobre la recta de regresión estimada.

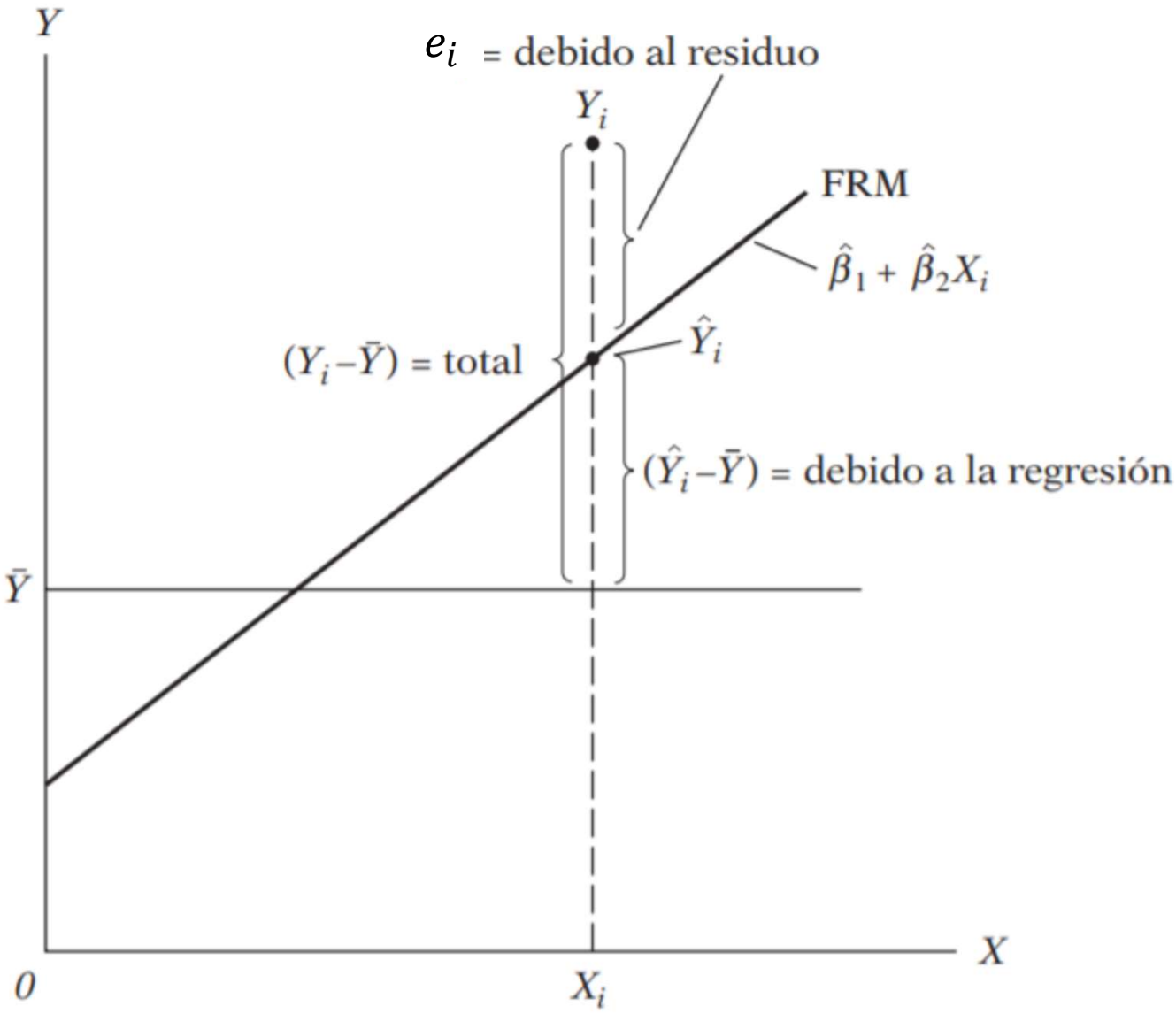
El poder explicativo del modelo de regresión: el R^2

- Definamos tres sumas al cuadrado que son claves en la definición del R^2 .
- Ellas son la suma de cuadrados total SCT, la suma de cuadrados explicada SCE y la suma de cuadrados de residuos SCR.
- Notar que $\mathrm{Y_i} - \bar{Y} = \left(\hat{Y}_\mathrm{i} - \bar{Y}\right) + \left(Y_i - \hat{Y}_i\right)$
- $\mathrm{SCT} = \sum(\mathrm{Y_i} - \bar{Y})^2 \quad \mathrm{SCE} = \sum\left(\hat{Y}_\mathrm{i} - \bar{Y}\right)^2 \quad \mathrm{SCR} = \sum\left(Y_i - \hat{Y}_i\right)^2$
- La cantidad de variabilidad explicada por el modelo de regresión lineal es la SCE porque
- $\mathrm{SCE} = \sum\left(\hat{Y}_\mathrm{i} - \bar{Y}\right)^2 = \sum \hat{\beta}_2 (X_i - \bar{X})^2 = \hat{\beta}_2 \sum (X_i - \bar{X})^2$

El poder explicativo del modelo de regresión: el R^2

- $\mathrm{SCT} = \sum(\mathrm{Y_i} - \bar{Y})^2 \quad \mathrm{SCE} = \sum(\hat{Y}_\mathrm{i} - \bar{Y})^2 \quad \mathrm{SCR} = \sum(Y_i - \hat{Y}_i)^2$
- La SCT es una medida de la suma de cuadrados total de los desvíos de Y_i en torno a su valor medio $\bar{Y}$, que es explicado por la suma de los cuadrados de los desvíos debidos a la regresión más la suma de los cuadrados de los residuos e_i o la parte no explicada por el modelo de regresión (ver gráfico).
- Con un poco de álgebra se prueba que $\mathrm{SCT} = \mathrm{SCE} + \mathrm{SCR}$
- $\frac{\mathrm{SCT}}{\mathrm{SCT}} = \frac{\mathrm{SCE}}{\mathrm{SCT}} + \frac{\mathrm{SCR}}{\mathrm{SCT}}$
- $1 = \frac{\mathrm{SCE}}{\mathrm{SCT}} + \frac{\mathrm{SCR}}{\mathrm{SCT}} \quad \rightarrow \quad \boxed{R^2 = \frac{\mathrm{SCE}}{\mathrm{SCT}} = 1 - \frac{\mathrm{SCR}}{\mathrm{SCT}}}$

El poder explicativo del modelo de regresión: el R^2



Correlación y el R^2

- El R^2 de una regresión de Y **sobre un único regresor** X, es igual al cuadrado del coeficiente de regresión entre X e Y.

- $R^2 = r_{X,Y}^2 \rightarrow r_{X,Y} = \sqrt{R^2}$

El poder explicativo del modelo de regresión: el R^2

- $0 \leq R^2 \leq 1$
- Si $\hat{\beta}_2 = 0$ entonces X no explica nada de la variabilidad de Y, los valores predichos de Y basados en la regresión es exactamente $\bar{Y}$. Por lo tanto la $\text{SCE} = 0$ y el $R^2 = 0$.
- Por el contrario si X explica toda la variabilidad de la Y, el $R^2 = 1$.
- Por lo general, la interpretación del R^2 suele multiplicarse por 100, y se lee como **el porcentaje de la variación muestral en** Y **explicada por el modelo de regresión, i.e. la variación en** X.
- Cuidado con la interpretación del $\boxed{R^2 = \frac{\text{SCE}}{\text{SCT}} = 1 - \frac{\text{SCR}}{\text{SCT}}}$

El error estándar de la regresión

- La suma de los cuadrados de los errores se utiliza para obtener una estimación de la varianza del error del modelo ε_i, que a su vez esta varianza nos va a servir para realizar la inferencia estadística del modelo de regresión.
- De acuerdo al supuesto #4 $E(\varepsilon_i) = 0$ y $Var(\varepsilon_i) = \sigma^2$
- Una estimación de la varianza del error del modelo esta dada por

$$\hat{\sigma}^2 = S_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SCR}{n-2}$$

- Se divide por n-2 porque se estimaron dos parámetros $\hat{\beta}_1$ y $\hat{\beta}_2$
- Este estimador de la varianza del error del modelo es **la base para la inferencia estadística en el modelo de regresión**.

El poder explicativo del modelo de regresión: el R^2

Dependent Variable: GASTOS_M1
Method: Least Squares
Date: 10/10/19 Time: 19:36
Sample: 1 100
Included observations: 100

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1534.152	470.3554	3.261686	0.0015
INGRESO_M1	0.588552	0.080329	7.326786	0.0000

R-squared	0.353911	Mean dependent var	4487.571	$\bar{Y}$
Adjusted R-squared	0.347318	S.D. dependent var	3000.105	
S.E. of regression $\hat{\sigma}$	2423.748	Akaike info criterion	18.44381	
SCR Sum squared resid	5.76E+08	Schwarz criterion	18.49592	
Log likelihood	-920.1907	Hannan-Quinn criter.	18.46490	
F-statistic	53.68180	Durbin-Watson stat	2.273001	
Prob(F-statistic)	0.000000			

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

Inferencia estadística

- Hasta el momento estudiamos la estimación de los parámetros de modelo de regresión lineal con dos variables.
- Mediante el método MCO, obtuvimos estimaciones β_1, β_2 y σ^2 que llamamos $\hat{\beta}_1$, $\hat{\beta}_2$ y $\hat{\sigma}^2 = S_e^2$. Dado que son estimadores estos valores cambian de muestra en muestra.
- La estimación es la mitad de trabajo que debemos realizar, la otra mitad es la inferencia o test de hipótesis.
- Se debe tener presente que, en el análisis de regresión, el objetivo no solo consiste en estimar la FRM, sino utilizar la FRM para realizar inferencia sobre la FRP.
- Por lo tanto querríamos saber cuan cerca esta $\hat{\beta}_2$ del verdadero β_2 o $\hat{\sigma}^2$ del verdadero σ^2.

Inferencia estadística

- Entonces como $\hat{\beta}_1$, $\hat{\beta}_2$ y $\hat{\sigma}^2$ son variables aleatorias es necesario conocer sus valores medios, sus varianzas y sus distribuciones de probabilidad para poder realizar la inferencia.

- Para hallar las distribuciones de estos estimadores se deben recurrir a los supuestos de Gauss Markov, que enumeramos anteriormente.

- Si se cumple el supuesto #4 (los ε_i son variables aleatorias que tienen media 0 y varianza constante. $E(\varepsilon_i) = 0$ y $Var(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$) y además si se asume que los ε_i **son normales** y las X_i son fijas, entonces Y_i también tiene distribución normal con la misma varianza de los ε_i.

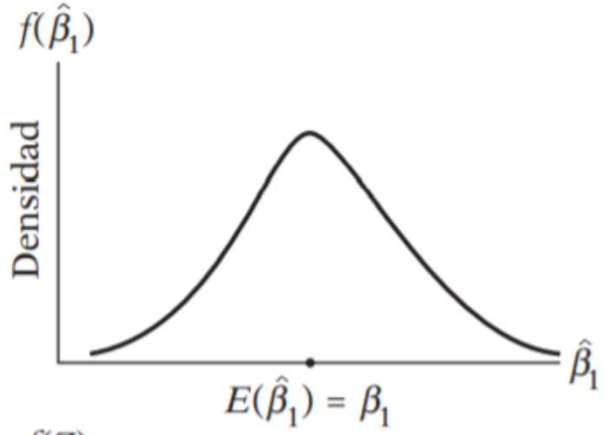
Inferencia estadística

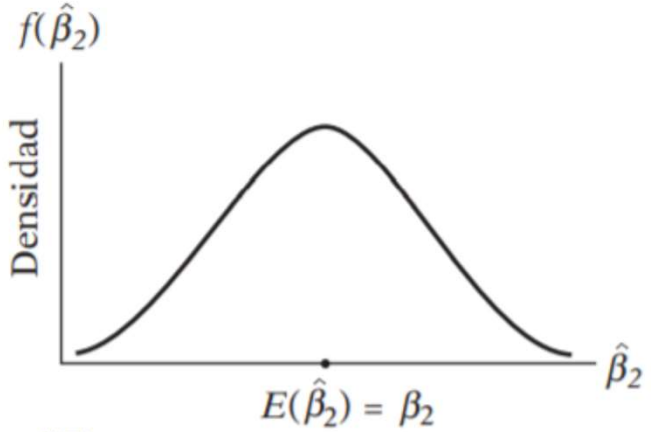
- Entonces a partir de los supuestos de Gauss Markov sumado a la normalidad de los ε_i se pueden obtener las propiedades de los estimadores
  - Son insesgados
  - Son de mínima varianza en la clase de estimadores insesgados, i.e. son eficientes
  - Son consistentes (a medida que el tamaño de muestra crece a ∞, convergen a los valores poblacionales.
- Y como los ε_i son normales se tiene que

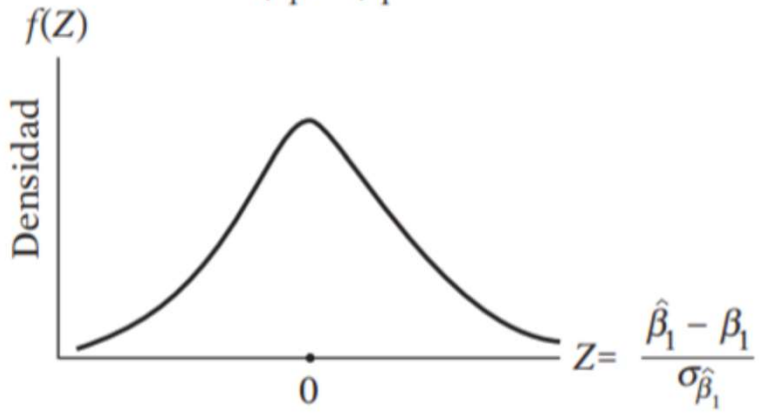
Distribuciones de probabilidad de $\hat{\beta}_1$ y $\hat{\beta}_2$

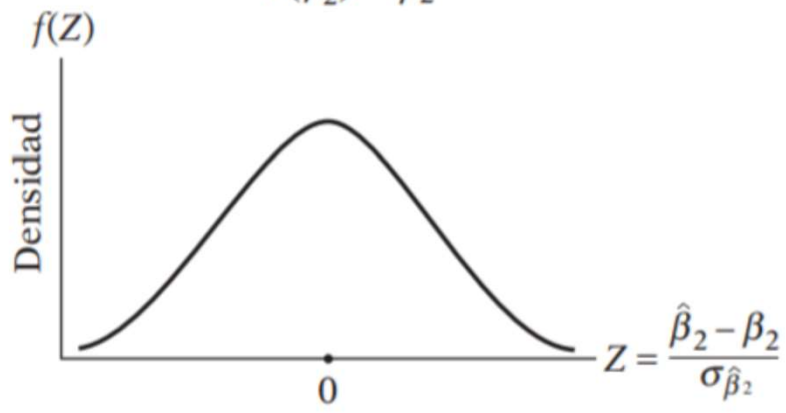
- $E(\hat{\beta}_1) = \beta_1 \quad Var(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1} = \frac{\sum X_i}{n \sum (X_i - \bar{X})^2} \sigma^2$
- O en forma compacta $\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2_{\hat{\beta}_1}\right)$
- $E(\hat{\beta}_2) = \beta_2 \quad Var(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}$
- O en forma compacta $\hat{\beta}_2 \sim N\left(\beta_2, \sigma^2_{\hat{\beta}_2}\right)$
- Por lo tanto se puede afirmar que
- $Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}}$ y $Z = \frac{\hat{\beta}_2 - \beta_2}{\sigma_{\hat{\beta}_2}}$ tienen distribución $N(0,1)$

Distribuciones de probabilidad de $\hat{\beta}_1$ y $\hat{\beta}_2$









Intervalo de confianza para $\hat{\beta}_1$ y $\hat{\beta}_2$

- $Z = \frac{\hat{\beta}_2 - \beta_2}{\sigma_{\hat{\beta}_2}} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\frac{\sigma^2}{\sum(X_i - \bar{X})^2}}} = \frac{(\hat{\beta}_2 - \beta_2)\sqrt{\sum(X_i - \bar{X})^2}}{\sigma}$

- Pero no conocemos σ . Pero $\hat{\sigma}^2 = S_e^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{SCR}{n-2}$ es un estimador insesgado de σ^2

- Por lo tanto $\hat{\sigma}^2_{\hat{\beta}_2} = \frac{S_e^2}{\sum(X_i - \bar{X})^2} = \frac{S_e^2}{(n-1)S_X^2}$

- Y sabemos que cuando reemplazamos $\sigma_{\hat{\beta}_2}$ por $\hat{\sigma}_{\hat{\beta}_2}$ puede escribirse como

Intervalo de confianza para $\hat{\beta}_1$ y $\hat{\beta}_2$

- $t = \frac{\hat{\beta}_2 - \beta_2}{\hat{\sigma}_{\hat{\beta}_2}} = \frac{estimador - parámetro}{Error\ estandar\ del\ estimador} = \frac{\hat{\beta}_2 - \beta_2}{\sqrt{\frac{\hat{\sigma}^2}{\sum(X_i - \bar{X})^2}}} =$

  $= \frac{(\hat{\beta}_2 - \beta_2)\sqrt{\sum(X_i - \bar{X})^2}}{\hat{\sigma}}$

- El estadístico t tiene una distribución t de student con $(n-2)$ grados de libertad.
- El intervalo de confianza para β_2 al $100(1-\alpha)\%$:
- $\hat{\beta}_2 \pm t_{\alpha/2} * \hat{\sigma}_{\hat{\beta}_2}$
- Notar que la varianza de $\hat{\beta}_2$ depende de S_e^2 y S_X^2

Intervalo de confianza para $\hat{\beta}_1$ y $\hat{\beta}_2$

- Idem para β_1 al $100(1-\alpha)\%$: $\hat{\beta}_1 \pm t_{\alpha/2} * \hat{\sigma}_{\hat{\beta}_1}$

- Donde $\hat{\sigma}_{\hat{\beta}_1} = \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right) S_e^2 = \left(\frac{1}{n} + \frac{\bar{X}^2}{(n-1)S_X^2}\right) S_e^2$

- Recordar la interpretación del intervalo de confianza

- $P\left(\hat{\beta}_2 - t_{\alpha/2} * \hat{\sigma}_{\hat{\beta}_2} \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2} * \hat{\sigma}_{\hat{\beta}_2}\right) = 1 - \alpha$

Intervalo de confianza para σ^2

- Para la varianza del modelo σ^2, se tiene que $(n-2)\frac{\hat{\sigma}^2}{\sigma^2}$ se distribuye según una chi-cuadrado con $(n-2)$ grados de libertad, siempre que los ε_i sean normales.

- $$P\left((n-2)\frac{\hat{\sigma}^2}{\chi^2_{\alpha/2}} \leq \sigma^2 \leq (n-2)\frac{\hat{\sigma}^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha$$

Test de hipótesis para $\hat{\beta}_1$ y $\hat{\beta}_2$

- Recordemos que $t = \frac{\hat{\beta}_2 - \beta_2}{\hat{\sigma}_{\hat{\beta}_2}} \sim t_{n-2}$
- $H_0: \beta_2 = \beta_2^* \quad vs. \quad H_A: \beta_2 \neq \beta_2^*$
- Todos los paquetes estadísticos evalúan la siguiente hipótesis $H_0: \beta_2 = 0 \quad vs. \quad H_A: \beta_2 \neq 0$
- Pero uno puede decidir que evaluar ya que conoce $\hat{\beta}_2$ y $\hat{\sigma}_{\hat{\beta}_2}$ para construir el test de hipótesis que desee, tanto bilateral como unilateral a derecha o izquierda.
- Idem para β_1 y σ^2

La significatividad estadística del $\hat{\beta}_2$

$$H_0: \beta_2 = 0 \quad vs. \quad H_A: \beta_2 \neq 0$$

$$t = \frac{\hat{\beta}_2}{\hat{\sigma}_{\hat{\beta}_2}} \sim t_{n-2}$$

Dependent Variable: GASTOS_M1
Method: Least Squares
Date: 10/10/19 Time: 19:36
Sample: 1 100
Included observations: 100

$\hat{\beta}_2$ $\hat{\sigma}_{\hat{\beta}_2}$

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1534.152	470.3554	3.261686	0.0015
INGRESO_M1	0.588552	0.080329	7.326786	0.0000

R-squared	0.353911	Mean dependent var	4487.571
Adjusted R-squared	0.347318	S.D. dependent var	3000.105
S.E. of regression	2423.748	Akaike info criterion	18.44381
Sum squared resid	5.76E+08	Schwarz criterion	18.49592
Log likelihood	-920.1907	Hannan-Quinn criter.	18.46490
F-statistic	53.68180	Durbin-Watson stat	2.273001
Prob(F-statistic)	0.000000		

Valor-p asociado al estadístico t

Idem para $\hat{\beta}_1$

Bibliografía para Análisis de Regresión

- Newbold, Paul (2008) . Sexta Edición. Estadística para los negocios y la economía. Pearson. Prentice Hall. Cápitulo 11

- Gujarati, Damodar y D. C. Porter (2009). Quinta Edición. Econometría. Mc Graw Hill. Capítulos 1 a 5.

- Stock, James y M.W. Watson (2003). Introduction to Econometrics. Addison Wesley. Cápitulos 1 a 4.