



Modelos de regresión uniecuaionales

- En estos modelos se expresa la variable dependiente (Y) como una función lineal de una o más variables (explicativas).
- Las relaciones causales van en una sola dirección: las variables explicativas causan a la variable dependiente.

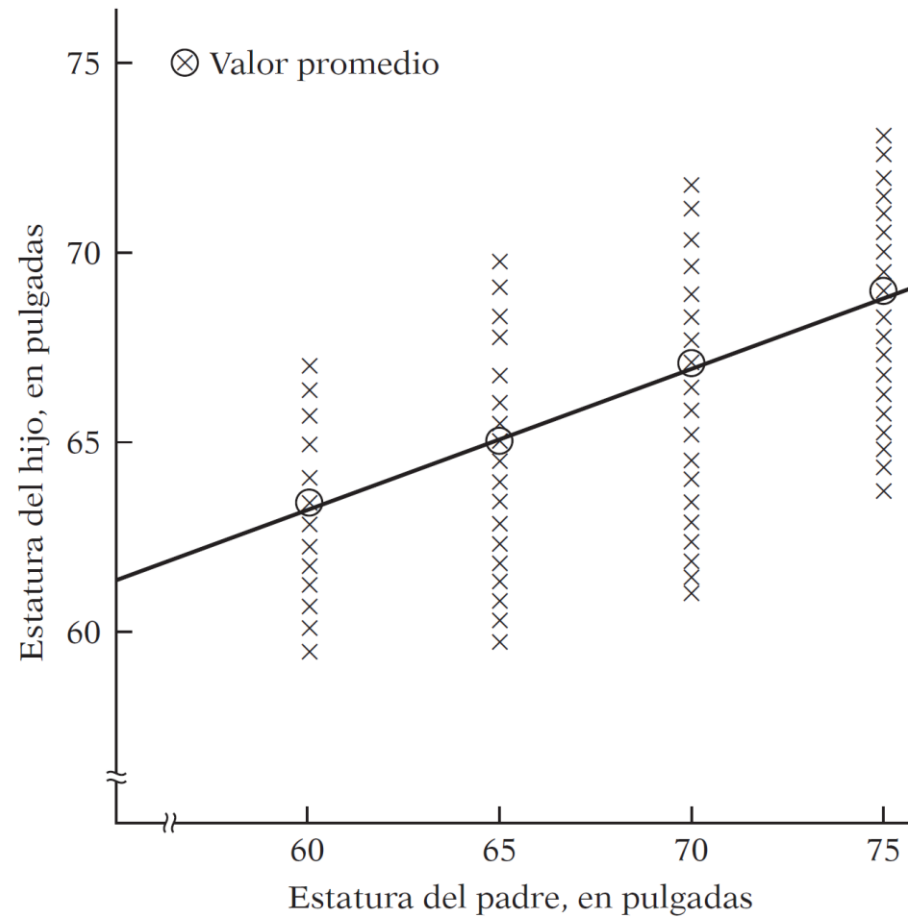
El origen del término “regresión”



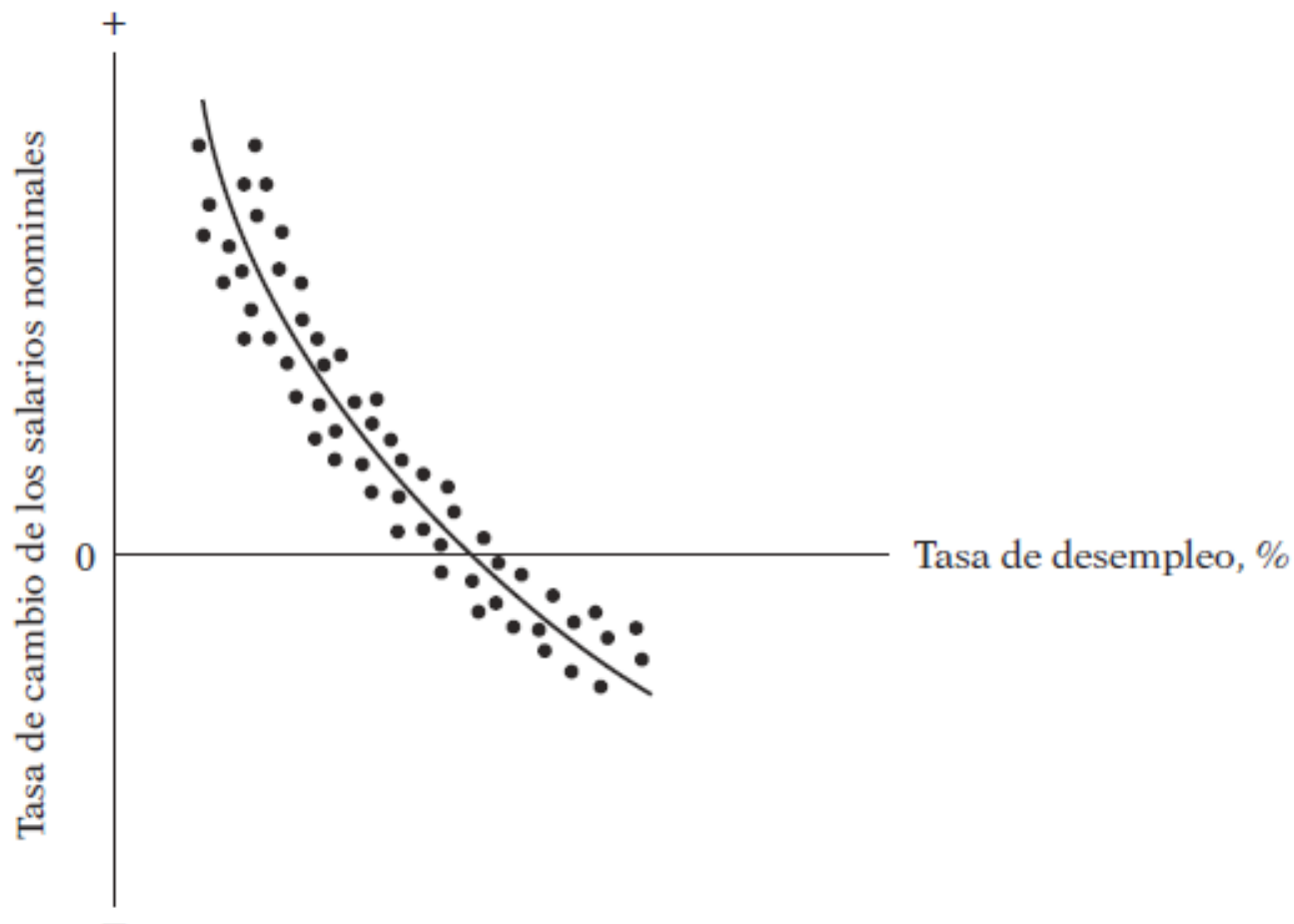
*(Galton, 1886)

El análisis de regresión (en tiempos modernos)

- Trata del estudio de la dependencia de una variable (*var. dependiente*) respecto de una o más variables (*explicativas*) para predecir el valor promedio de la var. dependiente.



Curva de Phillips



Regresión y causalidad

“Una relación estadística, por más fuerte y sugerente que sea, **nunca** podrá establecer una conexión causal: nuestras ideas de causalidad deben provenir de estadísticas externas y, en último término, de una u otra teoría” (Kendall & Stuart, 1961)

- **Una relación estadística por sí misma no puede implicar causalidad.**
 - Se debe acudir a consideraciones *a priori* o teóricas.
- La causalidad puede determinarse también por sentido común.

Regresión y correlación

ITESO



Universidad Jesuita
de Guadalajara

- El análisis de correlación está estrechamente relacionado al de regresión.
- Correlación
 - Medir el grado de asociación lineal entre dos variables.
- Regresión
 - Estimar o predecir el valor promedio de una var. con base en valores fijos de otra.

Terminología

ITESO



Universidad Jesuita
de Guadalajara

**Variable
dependiente**

Explicada

Predicha

Regresada

Respuesta

Endógena

Resultado

Controlada

**Variable
independiente**

Explicativa

Predictora

Regresora

Estímulo

Exógena

Covariante

De Control

- Análisis de regresión simple
 - Estudio de la dependencia de una variable respecto de una sola variable explicativa.
- Análisis de regresión múltiple
 - Estudio de la dependencia de una variable respecto de más de una variable explicativa.

Análisis de regresión con dos variables

La regresión bivariable

- Existe, hipotéticamente, un país de 60 familias. A esta población se le ha medido su ingreso (X) y su gasto de consumo semanal (Y).

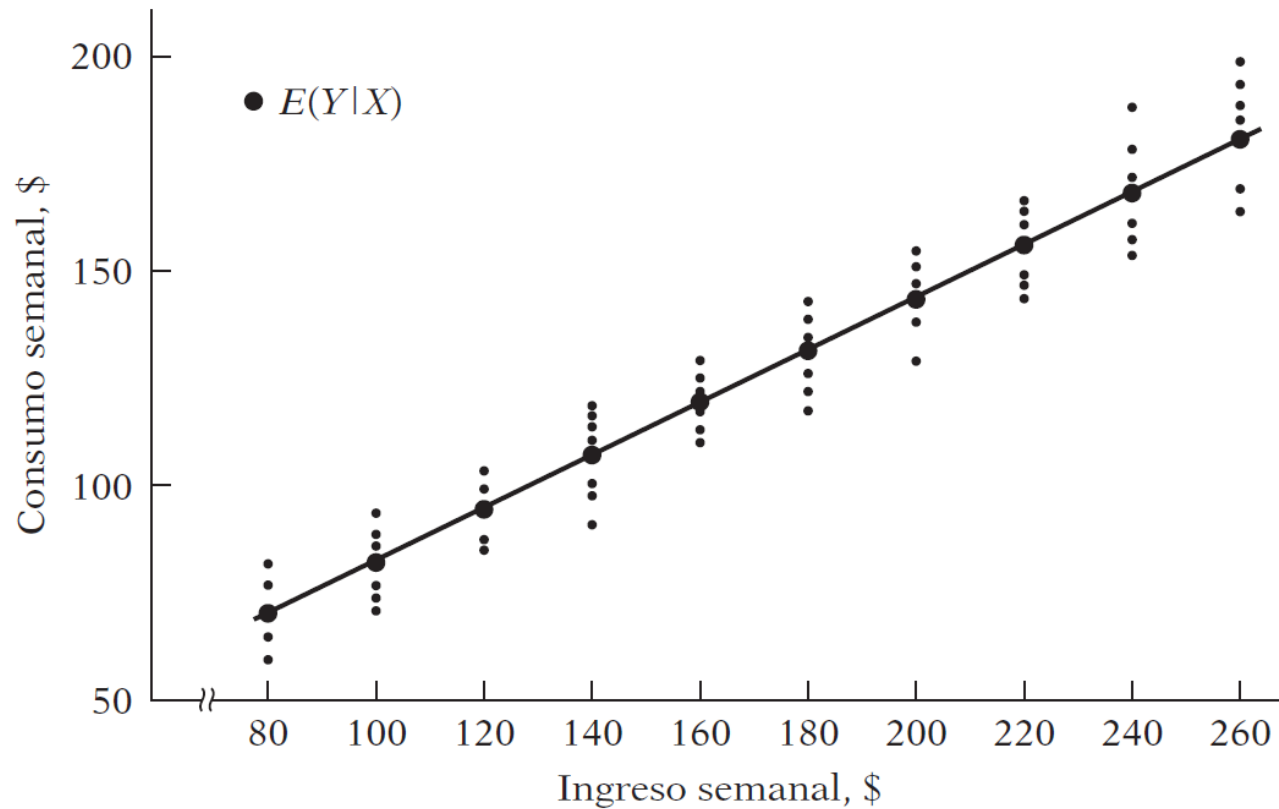
$Y \downarrow \quad X \rightarrow$	80	100	120	140	160	180	200	220	240	260
Consumo familiar semanal Y , \$	55 60 65 70 75 – –	65 70 74 80 85 88 –	79 84 90 94 98 – –	80 93 95 103 108 113 115	102 107 110 116 118 125 –	110 115 120 130 135 140 –	120 136 140 144 145 – –	135 137 140 152 157 160 162	137 145 155 165 175 189 –	150 152 175 178 180 185 191
Total	325	462	445	707	678	750	685	1 043	966	1 211
Media condicional de Y , $E(Y X)$	65	77	89	101	113	125	137	149	161	173

Curva de regresión poblacional

ITESO

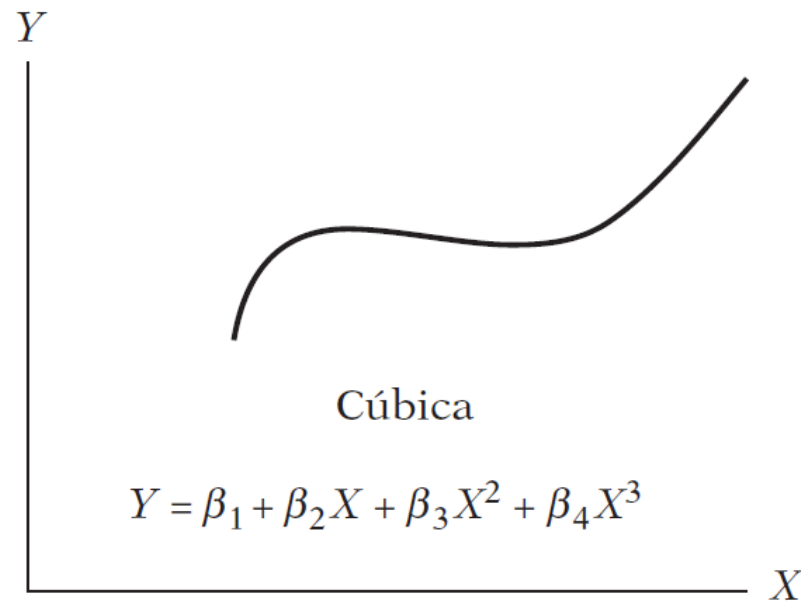
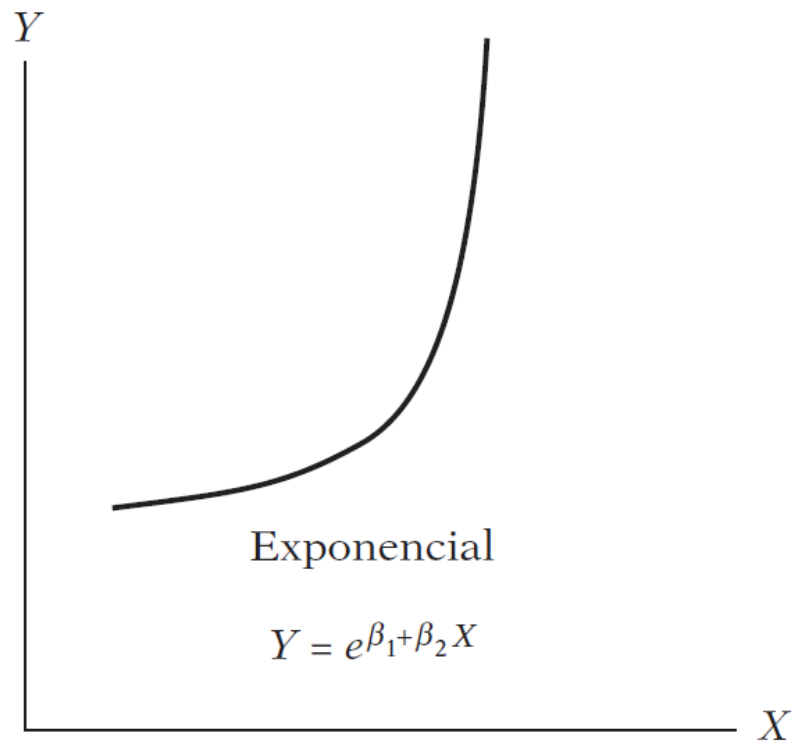
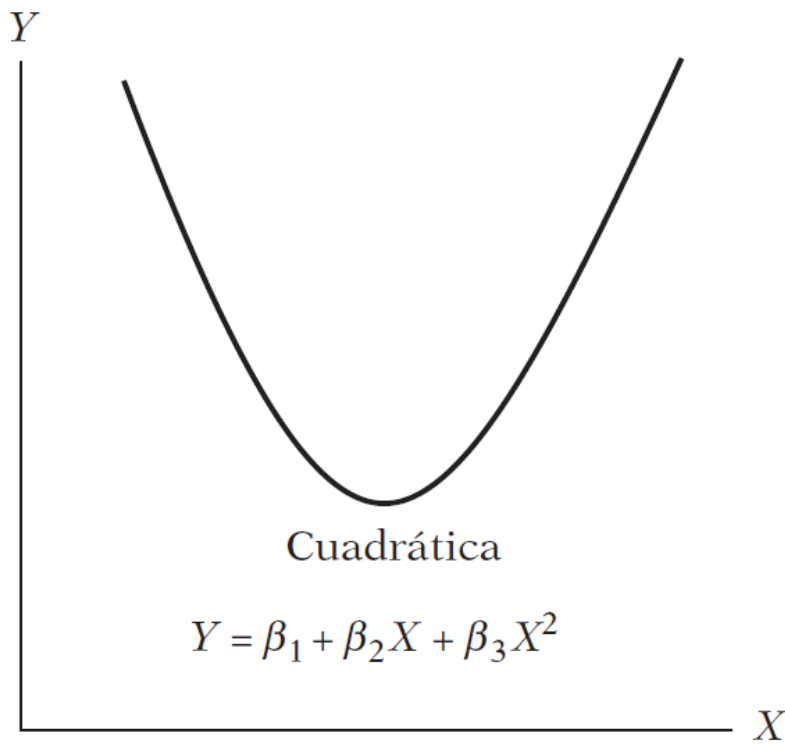


Universidad Jesuita
de Guadalajara



¿Qué significa que sea un modelo
lineal?



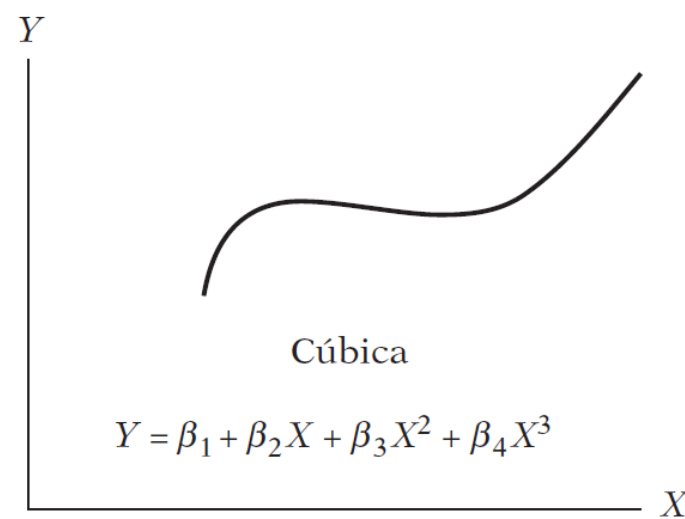
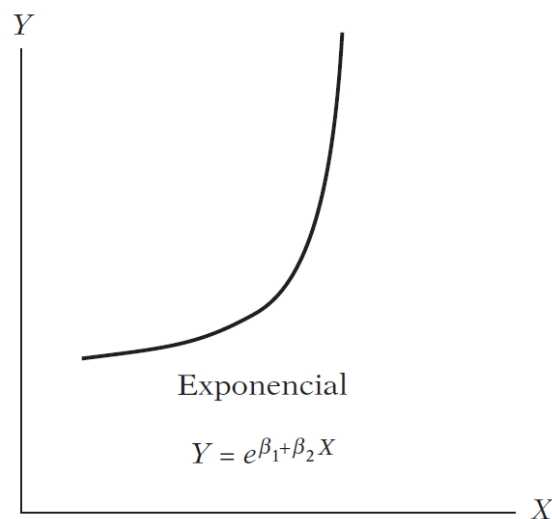
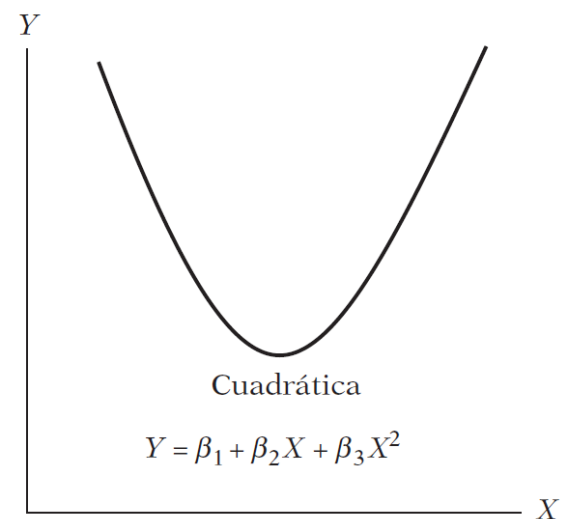


- Linealidad en las variables
 - La esperanza condicional de Y es una función lineal de X_i . $E(Y|X_i) = \beta_1 + \beta_2 X_i$
 - $E(Y|X_i) = \beta_1 + \beta_2 X_i^2$

- Linealidad en las variables
 - La esperanza condicional de Y es una función lineal de X_i . $E(Y|X_i) = \beta_1 + \beta_2 X_i$
 - ~~$E(Y|X_i) = \beta_1 + \beta_2 X_i^2$~~

- Linealidad en las variables
 - La esperanza condicional de Y es una función lineal de X_i . $E(Y|X_i) = \beta_1 + \beta_2 X_i$
 - ~~$E(Y|X_i) = \beta_1 + \beta_2 X_i^2$~~
- Linealidad en los parámetros, β

- Linealidad en las variables
 - La esperanza condicional de Y es una función lineal de X_i . $E(Y|X_i) = \beta_1 + \beta_2 X_i$
 - ~~$E(Y|X_i) = \beta_1 + \beta_2 X_i^2$~~
- Linealidad en los parámetros, β
 - $E(Y|X_i) = \beta_1 + \beta_2 X_i^2$ es lineal en los parámetros.



Modelo de regresión lineal

- Significa que los parámetros, β , de la ecuación son lineales.
- Las variables explicativas pueden o no ser lineales.

$$- E(Y|X_i) = \beta_1 + \beta_2 X_i$$

$$- E(Y|X_i) = \beta_1 + \beta_2 X_i^2$$

Especificación estocástica de la función de regresión poblacional

- Los datos se agrupan alrededor de la esperanza condicional, por lo que existe una desviación de Y_i , ε_i .

$$\varepsilon_i = Y_i - E(Y|X_i)$$

$$Y_i = E(Y|X_i) + \varepsilon_i$$

- ε_i representa a todas las variables omitidas o ignoradas que pudieran afectar a Y .

$Y \downarrow \quad X \rightarrow$	80	100	120	140	160	180	200	220	240	260
Consumo familiar semanal Y , \$	55 60 65 70 75 – –	65 70 74 80 85 88 –	79 84 90 94 98 – –	80 93 95 103 108 113 115	102 107 110 116 118 125 –	110 115 120 130 135 140 –	120 136 140 144 145 – –	135 137 140 152 157 160 162	137 145 155 165 175 189 –	150 152 175 178 180 185 191
Total	325	462	445	707	678	750	685	1 043	966	1 211
Media condicional de Y , $E(Y X)$	65	77	89	101	113	125	137	149	161	173



- ¿Cómo sería el consumo con $X=80$ si asumimos que es una función lineal?

$Y \downarrow \quad X \rightarrow$	80	100	120	140	160	180	200	220	240	260
Consumo familiar semanal Y , \$	55 60 65 70 75 – –	65 70 74 80 85 88 –	79 84 90 94 98 – –	80 93 95 103 108 113 115	102 107 110 116 118 125 –	110 115 120 130 135 140 –	120 136 140 144 145 – –	135 137 140 152 157 160 162	137 145 155 165 175 189 –	150 152 175 178 180 185 191
Total	325	462	445	707	678	750	685	1 043	966	1 211
Media condicional de Y , $E(Y X)$	65	77	89	101	113	125	137	149	161	173



- ¿Cómo sería el consumo con $X=80$ si asumimos que es una función lineal?
- $Y_1 = 55 = \beta_1 + \beta_2(80) + \varepsilon_1$

$Y \downarrow \backslash X \rightarrow$	80	100	120	140	160	180	200	220	240	260
Consumo familiar semanal Y , \$	55	65	79	80	102	110	120	135	137	150
	60	70	84	93	107	115	136	137	145	152
	65	74	90	95	110	120	140	140	155	175
	70	80	94	103	116	130	144	152	165	178
	75	85	98	108	118	135	145	157	175	180
	–	88	–	113	125	140	–	160	189	185
	–	–	–	115	–	–	–	162	–	191
Total	325	462	445	707	678	750	685	1 043	966	1 211
Media condicional de Y , $E(Y X)$	65	77	89	101	113	125	137	149	161	173



- ¿Cómo sería el consumo con $X=80$ si asumimos que es una función lineal?
- $Y_1 = 55 = \beta_1 + \beta_2(80) + \varepsilon_1$
- $Y_2 = 60 = \beta_1 + \beta_2(80) + \varepsilon_2$
- $Y_3 = 65 = \beta_1 + \beta_2(80) + \varepsilon_3$
- $Y_4 = 70 = \beta_1 + \beta_2(80) + \varepsilon_4$
- $Y_5 = 55 = \beta_1 + \beta_2(80) + \varepsilon_5$

- Si tomamos la esperanza de $Y_i = E(Y|X_i) + \varepsilon_i$

$$E(Y_i|X_i) = E[E(Y|X_i)] + E(\varepsilon_i|X_i)$$

- Si tomamos la esperanza de $Y_i = E(Y|X_i) + \varepsilon_i$

$$E(Y_i|X_i) = E[E(Y|X_i)] + E(\varepsilon_i|X_i)$$

$$= E(Y|X_i) + E(\varepsilon_i|X_i)$$

- Si tomamos la esperanza de $Y_i = E(Y|X_i) + \varepsilon_i$

$$E(Y_i|X_i) = E[E(Y|X_i)] + E(\varepsilon_i|X_i)$$

$$= E(Y|X_i) + E(\varepsilon_i|X_i)$$

Dado que $E(Y_i|X_i) = E(Y|X_i)$, esto implica que

$$E(\varepsilon_i|X_i) = 0$$

¿Por qué es importante el término de error estocástico, ε_i ?

- Si ε_i incluye a todas las demás variables que pudieran afectar a Y ,
- ¿Por qué no simplemente se incluyen explícitamente en el modelo?
- ¿Por qué no intentamos agregar las más posibles? (modelo de regresión múltiple)

1. Teoría incompleta

- No se tiene la seguridad de todas las variables que afectan a Y .

2. Disponibilidad de los datos

3. Variables centrales y periféricas

- Variables que pudieran afectar a Y pero no tan significativamente que valga la pena el costo de introducirlas.

4. Aleatoriedad intrínseca en el comportamiento humano

5. Variables proxy inadecuadas

- Errores de medición.

6. Principio de parsimonia

- Si dos o tres variables explican suficientemente bien a Y , con eso basta.

7. Forma funcional incorrecta

Función de regresión muestral

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

- \hat{Y}_i : *estimador de $E(Y|X_i)$*
- $\hat{\beta}_1$: *estimador de β_1*
- $\hat{\beta}_2$: *estimador de β_2*
- Estimación:
 - Un valor numérico particular obtenido por un estimador en un análisis.

- En su forma estocástica

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

$\widehat{\varepsilon}_i$: *residual*