Elementos de Probabilidad y Estadística

Diplomatura en Ciencias de Datos

2024

Entrega 4

Dr. Matías Hisgen – Lic. Celine Cabás – Lic. Fernando Álvarez

FACENA - UNNE

Valor Esperado o Esperanza

- Anteriormente definimos a μ_y como la *media* poblacional de la variable aleatoria *y*. Dicha *media* puede ser vista como el *Valor Esperado* o *Esperanza* de *y:*

$$E(y) = \mu_y$$

Así, es posible escribir *y* como:

$$y = \mu_y + u,$$

en donde $u = (y - \mu_y)$ son las desviaciones respecto de la media.

Esperanza Condicional

- Si dos variables (y, x) están relacionadas positivamente, los valores de y tienden a aumentar a medida que x aumenta.
- Generalizando, la ***media*** de una variable (y) puede ***cambiar*** su valor a medida que otra variable (x) ***cambia***. Así es posible considerar a $E(y) = \mu_y$ como una *función* de x. Tal función se conoce como la *esperanza condicional:*

$$E(y|x) = \mu_{y|x}$$

Modelo de Regresión Lineal Simple

- Si la "esperanza de y condicional a x" , $E(y|x)$, es modelada como una *función lineal* de x, surge el modelo de Regresión Lineal Simple:

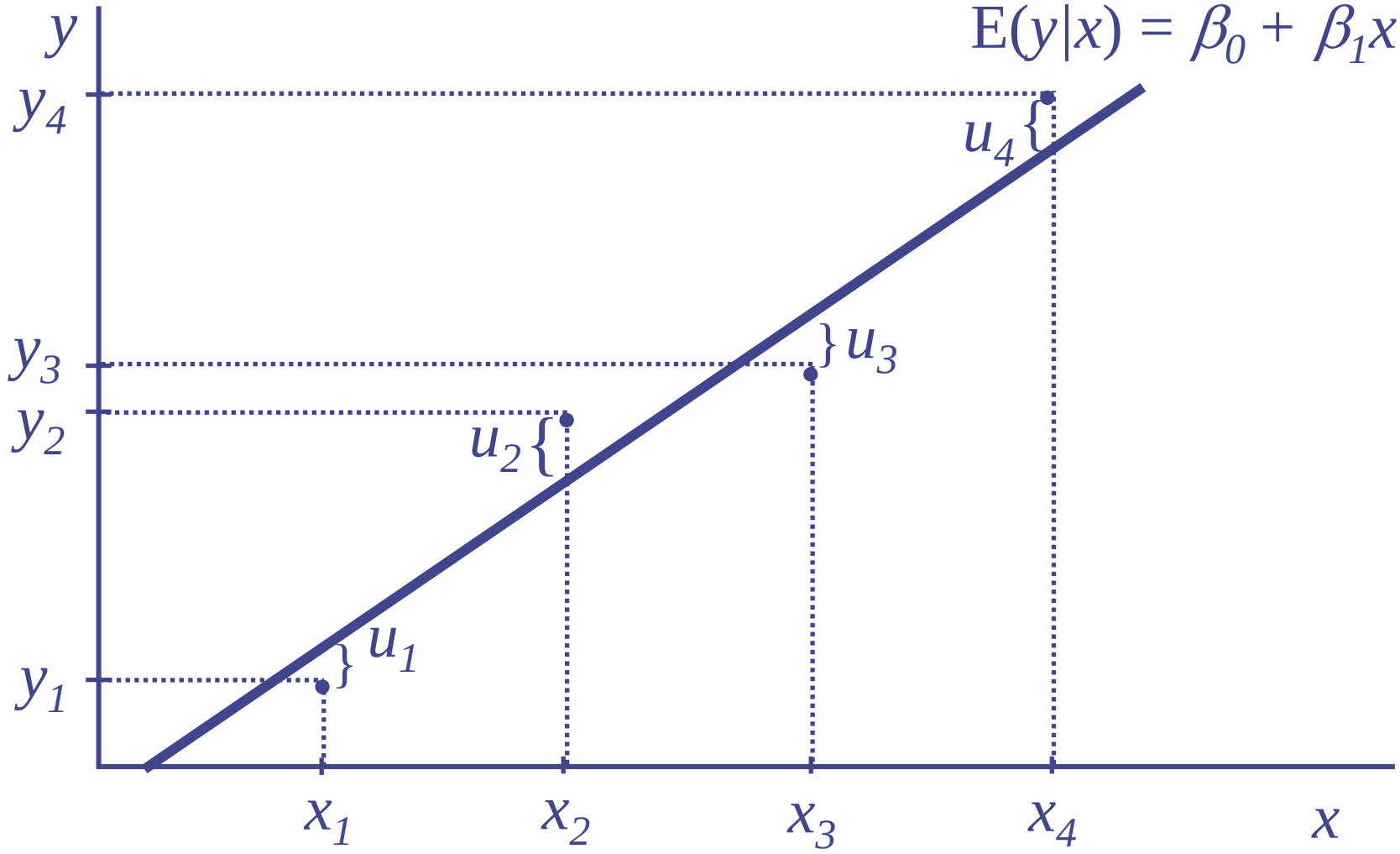
$$E(y|x) = \mu_{y|x} = \beta_0 + \beta_1 x$$

Y como antes, es posible escribir

$$y = E(y|x) + u = \beta_0 + \beta_1 x + u$$

donde u son las desviaciones respecto de la esperanza condicional (o media condicional).

Línea de regresión poblacional siendo *x* una variable continua.



Ejemplos

- $rendimiento = \beta_0 + \beta_1\ fertilizante + u$
- $salario = \beta_0 + \beta_1\ años\_educación + u$
- $años\_educación = \beta_0 + \beta_1\ \ sexo + u$

- La linealidad de estas ecuaciones implica que todo cambio de *x* en una unidad tiene *siempre* el mismo efecto sobre *y* (que es igual a β_1 en este caso), sin importar el valor inicial de *x*.

El término de “error aleatorio” (*u*)

- El componente aleatorio del modelo es *u*, dentro del cual se encuentran todos los demás factores que afectan la variable dependiente (*y*) y que no se han *incluido* como variables *independientes* (o *regresores*) en el modelo.

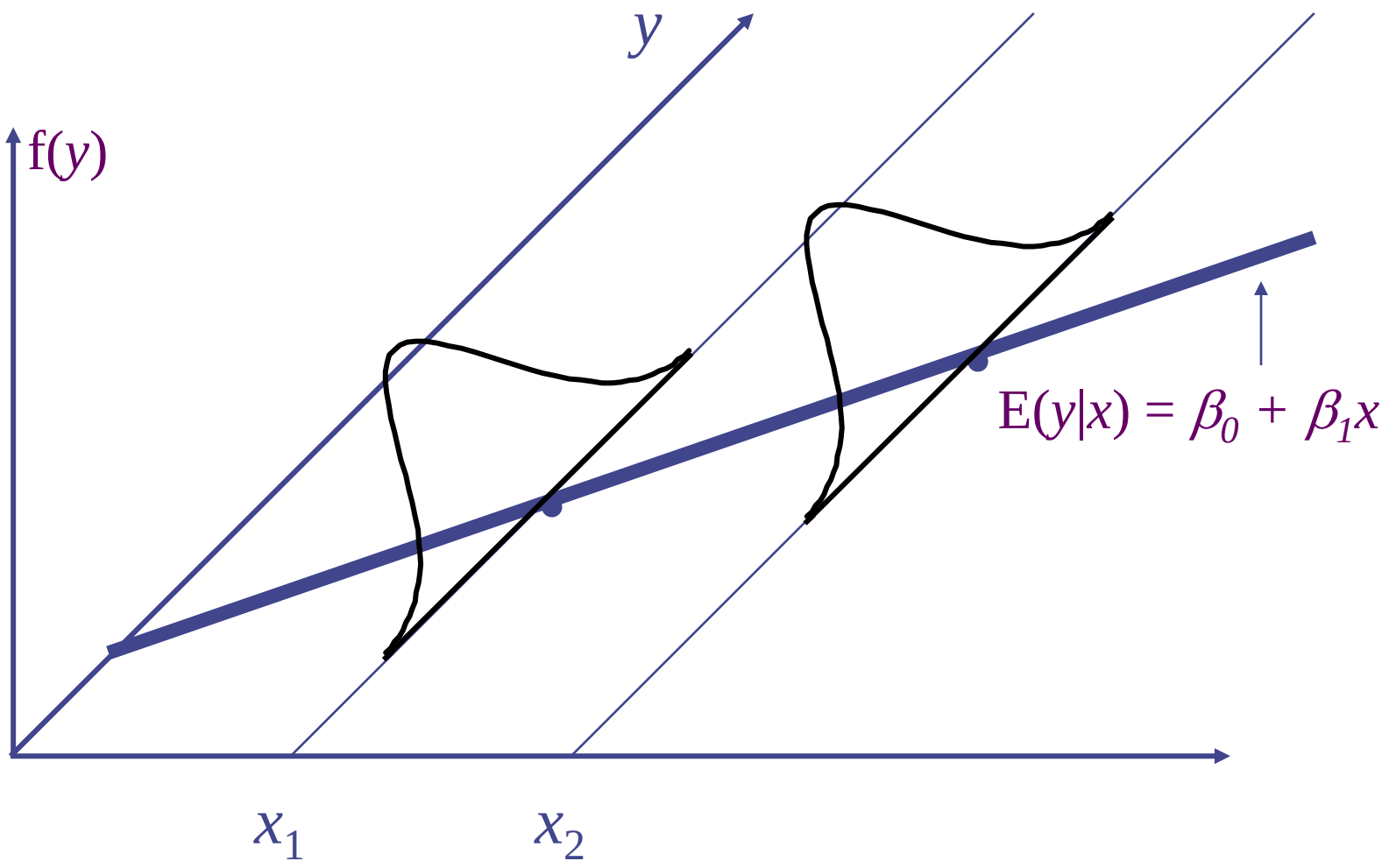
Esperanza Condicional Cero

- Explicitamos un supuesto crucial acerca de cómo u y x están relacionadas:

  $E(u|x) = E(u) = 0$, lo que implica que
  $E(y|x) = \beta_0 + \beta_1 x$, como ya vimos antes.

- Más adelante se entenderá porqué este supuesto es importante para interpretar el modelo.

E($y|x$) como una función lineal de x, donde para cada valor de x, la distribución de y está centrada en E($y|x$)



Mínimos Cuadrados Ordinarios

- Dada una muestra aleatoria de tamaño n de la población $\{(x_i,y_i): i=1, \ldots,n\}$, podemos escribir cada observación de la muestra como

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

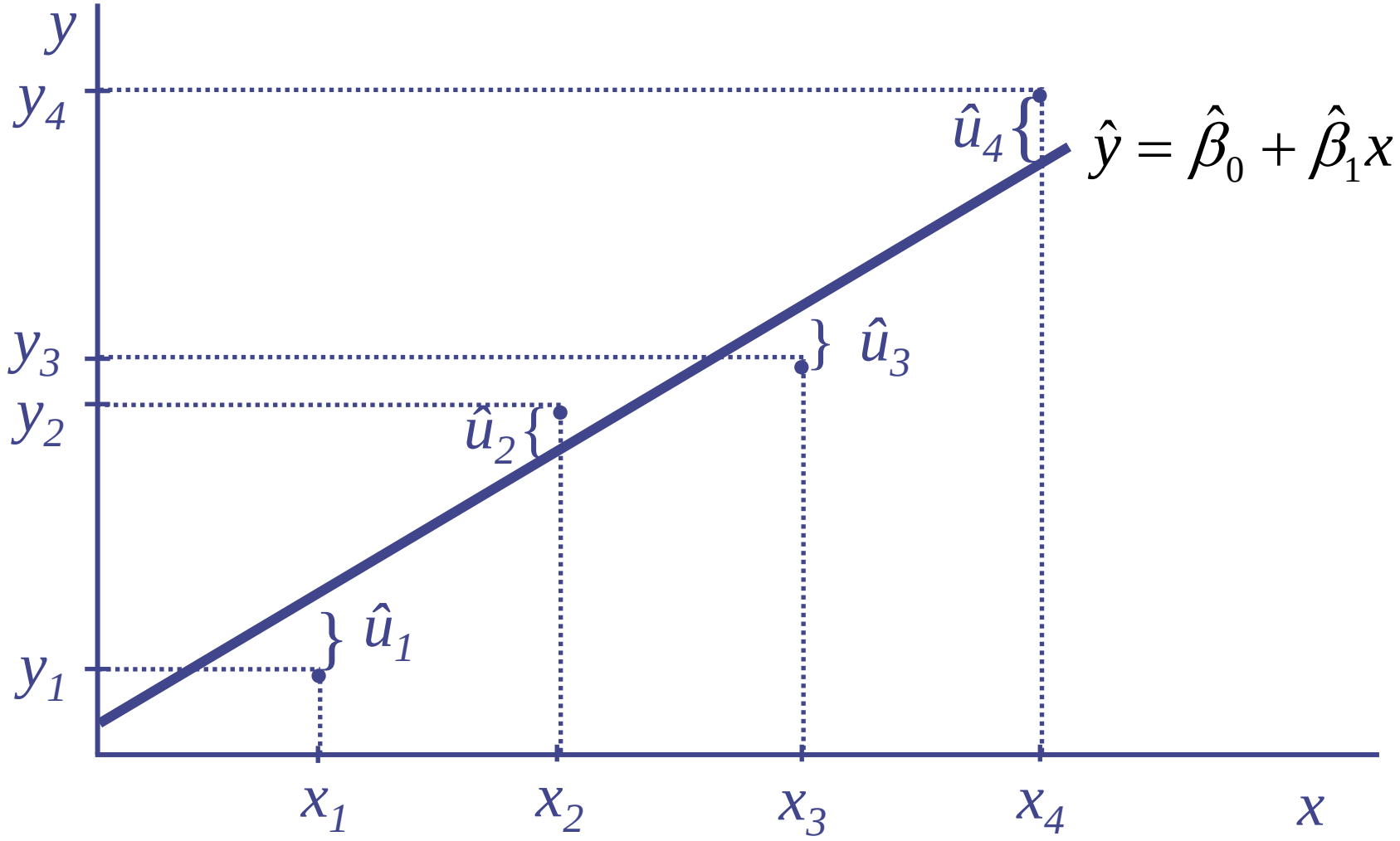
La idea básica de la regresión es estimar los parámetros poblacionales (β_0 y β_1) usando la muestra, para obtener

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$$

Mínimos Cuadrados Ordinarios

- El residuo $\hat{u}_i$ es un estimador del término de error u_i y es la diferencia entre la línea ajustada y el *i-esimo* punto de la muestra.
- Intuitivamente, MCO consiste en ajustar una línea a través de los n puntos muestrales (x_i,y_i) de tal forma que la suma de los residuos $(\hat{u}_i)$ elevados al cuadrado sea tan pequeña como fuese posible, de allí el término "mínimos cuadrados"

Línea de regresión muestral ajustada, puntos de datos muestrales y los correspondientes resuiduos



El problema de minimización

- Dada la idea intuitiva de ajustar una línea, podemos establecer ahora un problema formal de minimización
- Esto es, queremos elegir los parámetros de tal forma que se minimice la siguiente expresión:

$$\sum_{i=1}^{n}(\hat{u}_i)^2 = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

El problema de minimización

- Resolviendo el problema de minimización para los dos parámetros, obtenemos las condiciones de primer orden siguientes,

$$\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)=0$$

$$\sum_{i=1}^{n} x_i\left(y_i-\hat{\beta}_0-\hat{\beta}_1 x_i\right)=0$$

Derivación de estimadores MCO

- Dada la definición de media muestral, y las propiedades de la sumatoria, podemos reescribir la primera condición para obtener el estimador de la *ordenada al origen* o *intercepto*

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x},$$

o

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

La pendiente estimada por MCO

Despejando la pendiente

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

siendo $\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$

Resumen de la estimación de la pendiente

- El estimador MCO de la pendiente es igual a la *covarianza muestral* entre *y* y *x* dividida por la varianza muestral de *x*.
- Si *x* y *y* están correlacionadas positivamente, la pendiente será positiva.
- Si *x* y *y* están correlacionadas negativamente, la pendiente será negativa.
- Notar que es necesario que *x* tenga variabilidad en la muestra.

Descomposición de la varianza

Podemos ver a cada observación y_i como compuesta de una parte explicada $\hat{y}_i$, y otra parte no explicada $\hat{u}_i$, $y_i = \hat{y}_i + \hat{u}_i$. Luego definimos lo siguiente:

$\sum (y_i - \bar{y})^2$: suma total de cuadrados (STC)

$\sum (\hat{y}_i - \bar{y})^2$: suma explicada de cuadrados (SEC)

$\sum \hat{u}_i^2$: suma de residuos al cuadrado (SRC)

Luego tenemos que $\text{STC} = \text{SEC} + \text{SRC}$

Bondad del ajuste

- Cómo podemos medir cuán bien se ajusta a los datos la línea de regresión estimada?
- Podemos computar la proporción de la *suma de cuadrados totales* (STC) que es explicada por el modelo (es decir, SEC/STC), a esta medida la llamamos la R-cuadrada de la regresión o "coeficiente de determinación":

$$R^2 = SEC/STC = 1 - SRC/STC$$

Propiedades estadísticas de los estimadores MCO

- Supuestos de Gauss-Markov (G-M)

1. El modelo poblacional es lineal en los parámetros:
$$y = \beta_0 + \beta_1 x + u$$
2. Tenemos a disposición una muestra aleatoria de tamaño n, $\{(x_i, y_i): i=1, 2, \ldots, n\}$, extraída de la población. Por lo que podemos escribir el modelo para cada observación muestral como
$$y_i = \beta_0 + \beta_1 x_i + u_i$$
3. Suponemos $E(u|x) = 0$ y por lo tanto $E(u_i|x_i) = 0$
4. Suponemos que hay variación muestral en las x_i

Insesgamiento

Bajo los 4 supuestos de G-M anteriores, el estimador MCO es insesgado en *muestras repetidas*:

$$E\left(\hat{\beta}_0\right) = \beta_0 \qquad E\left(\hat{\beta}_1\right) = \beta_1$$

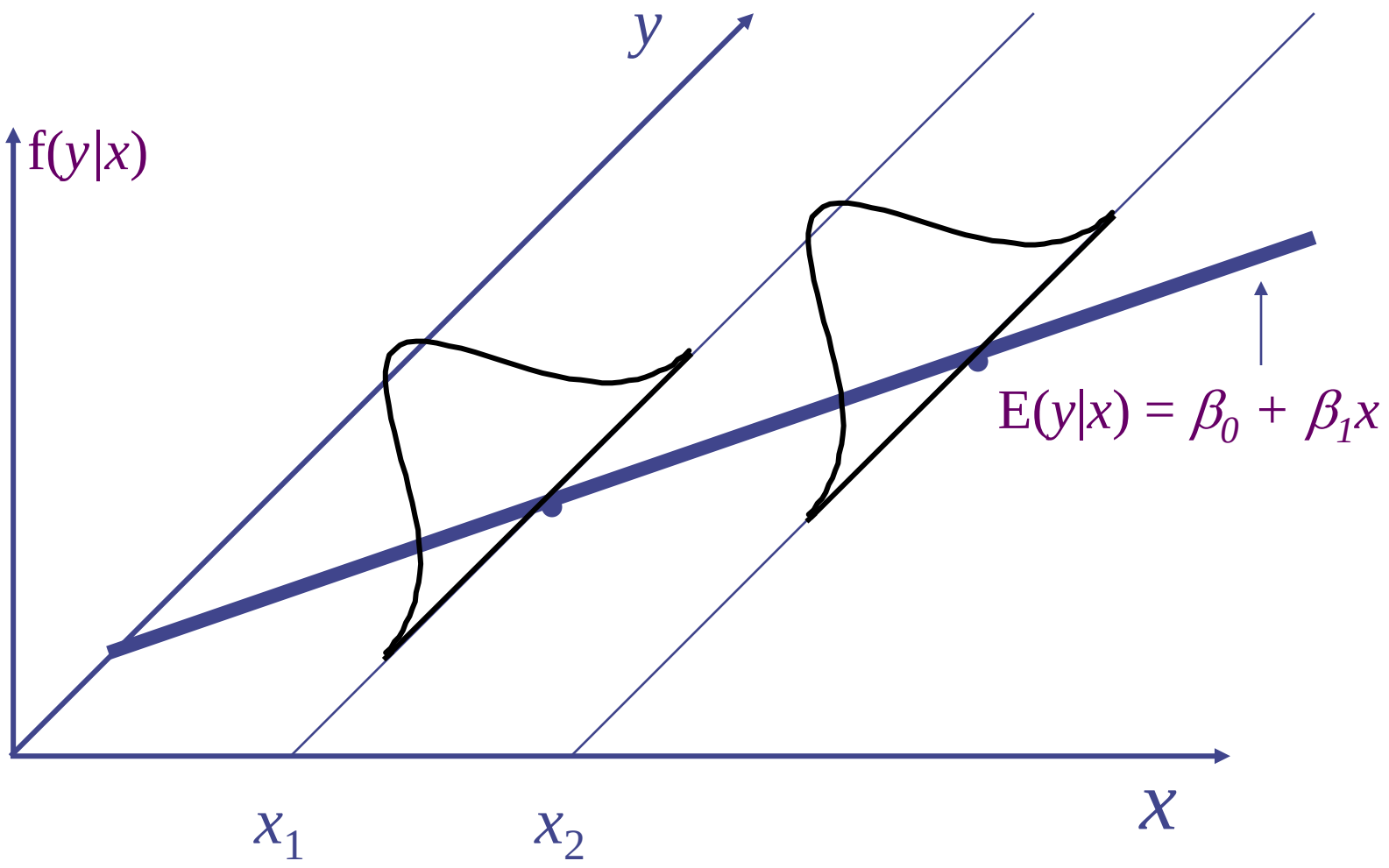
Recordar que insesgamiento es una propiedad del **estimador** – en una muestra dada podemos estar "cerca" o "lejos" del verdadero valor del parámetro.

Varianza de los estimadores MCO

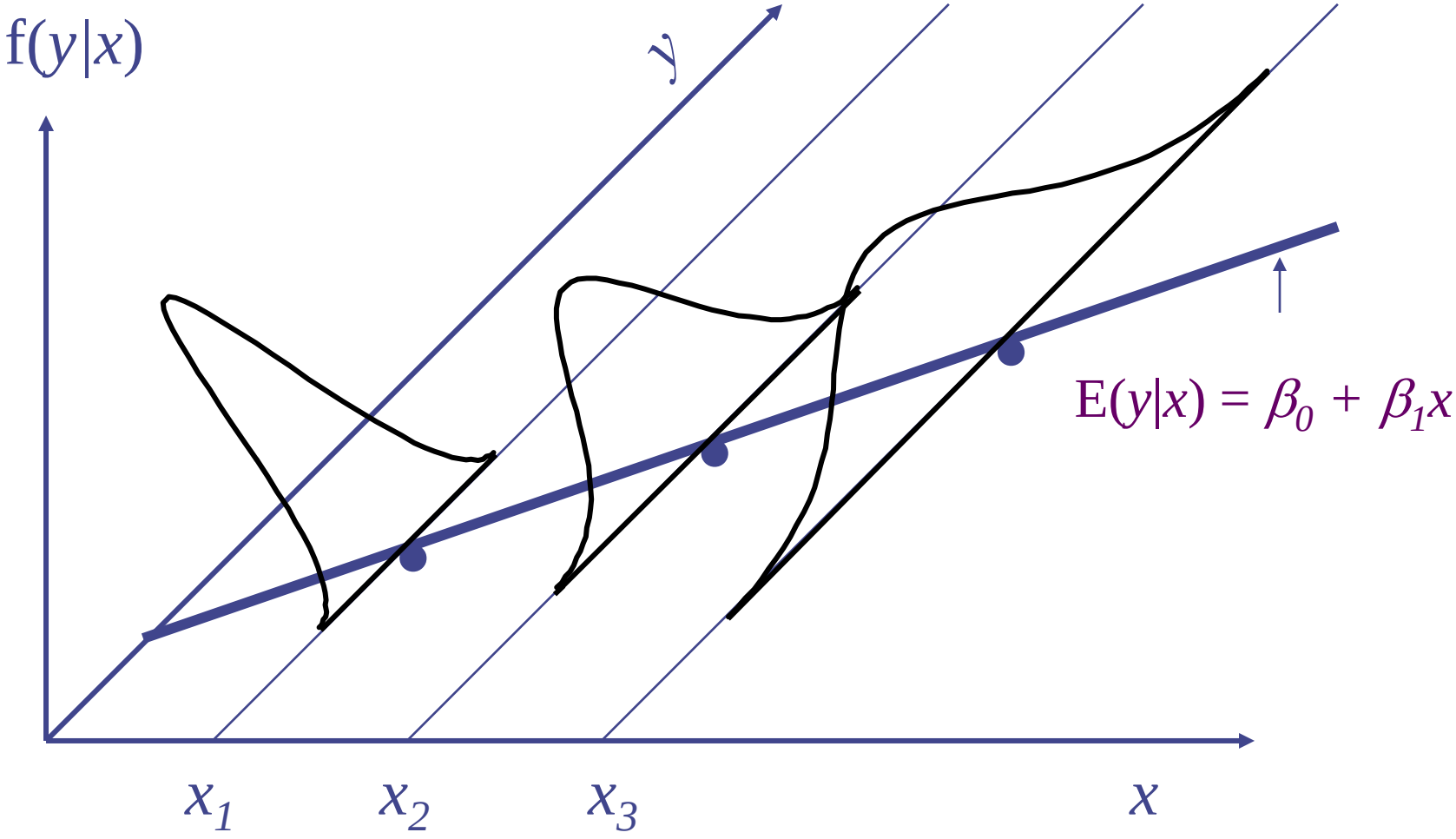
- Hasta ahora lo que sabemos es que la distribución muestral (en *muestras repetidas*) del estimador está centrada alrededor del verdadero parámetro (por insesgamiento).
- Pero queremos saber cuán dispersa es esta distribución.
- Es mas fácil analizar esta varianza si establecemos un supuesto adicional

  $\mathrm{Var}(u|x) = \mathrm{E}(u^2|x) = \sigma^2$ (Homocedasticidad).

El caso Homocedástico



El caso Heterocedástico



Varianza de MCO

El estimador de la varianza del error es:

$$\hat{\sigma}^2 = \frac{1}{(n-2)} \sum \hat{u}_i^2 = SRC/(n-2)$$

Luego, el estimador de la varianza MCO es

$$\text{Var}(\hat{\beta}_1) = \sigma^2 / \left(\sum (x_i - \bar{x})^2 \right)^{1/2}$$

Tomando raíz cuadrada, se tiene el error estándar

$\text{ee}(\hat{\beta}_1) = \hat{\sigma} / (\sum (x_i - \bar{x})^2)^{1/2}$, donde

Varianza de MCO (resumen)

- A mayor varianza del error, σ^2, mayor varianza del estimador de la pendiente
- A mayor variablilidad en las x_i, menor la varianza del estimador de la pendiente
- Un mayor tamaño de la muestra hace disminuir la varianza del estimador de la pendiente

Regresión Lineal Múltiple

- El modelo poblacional viene dado por:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$$

- β_0 sigue siendo el intercepto (ordenada)
- β_1 a β_k son los parámetros de las pendientes
- u es el término de error (o perturbación)

Interpretación

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$, entonces

$\Delta\hat{y} = \hat{\beta}_1 \Delta x_1 + \hat{\beta}_2 \Delta x_2 + \ldots + \hat{\beta}_k \Delta x_k$,

así mantner $x_2, \ldots, x_k$ constantes, es decir, $\Delta x_2 = \Delta x_3 = \ldots = \Delta x_k = 0$, implica que $\Delta\hat{y} = \hat{\beta}_1 \Delta x_1$,

es decir, β_1 tiene una interpretación *causal* (*en términos de mantener costantes* las demás x´s)

Regresión Simple vs. Múltiple

Comparar la regresión simple $\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$

con la regresión múltiple $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$

Generalmente, $\tilde{\beta}_1 \neq \hat{\beta}_1$ a menos que :

$\hat{\beta}_2 = 0$ (no hay efecto parcial de x_2)

o bien

x_1 y x_2 no estén correlaciondas en la muestra

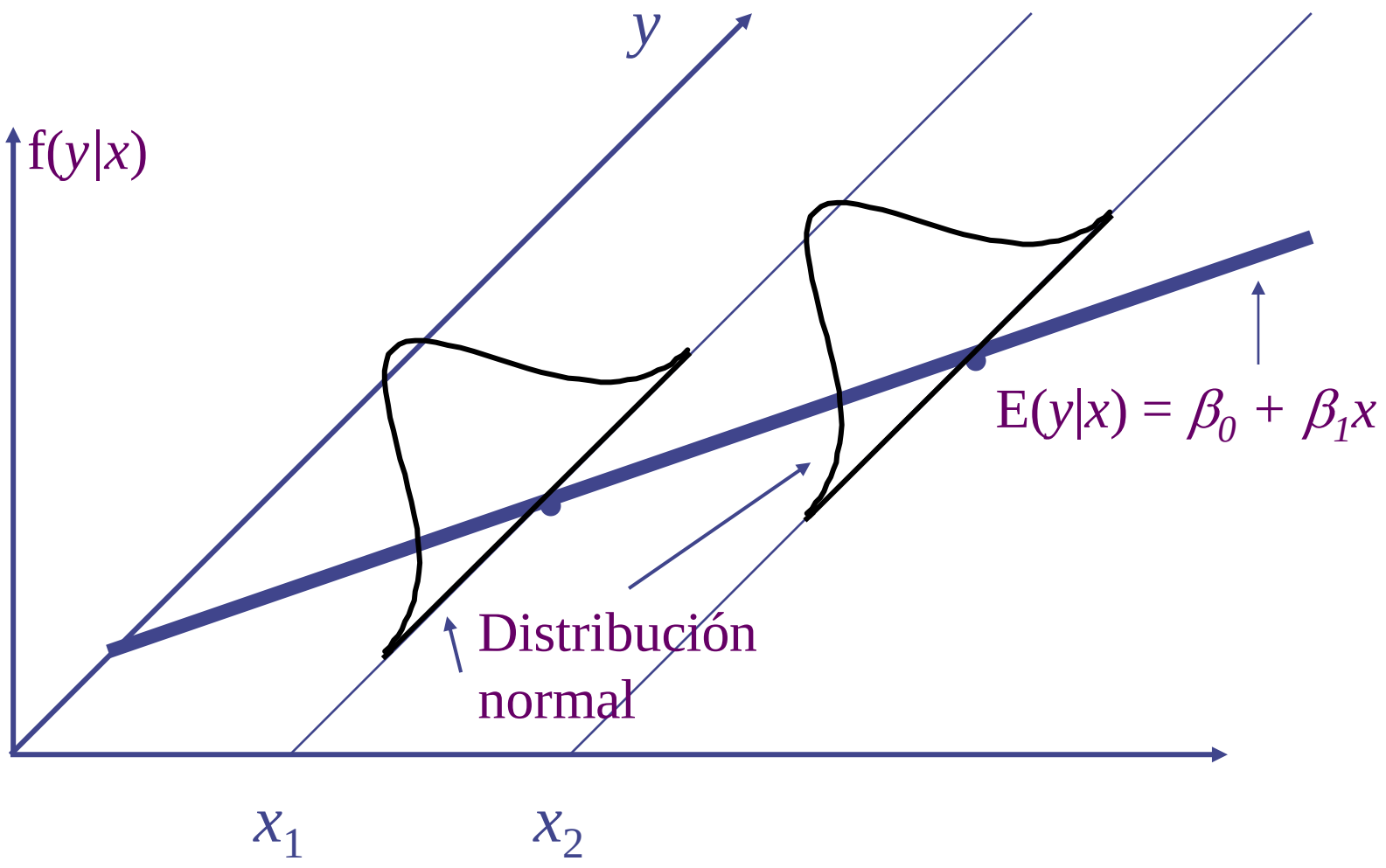
Inferencia en el modelo de regresión

- Ahora pasemos al modelo de regresión lineal

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

- Con el fin de hacer pruebas de hipótesis, vamos a introducir otro supuesto adicional.
- Suponemos que u se distribuye normal con media cero y varianza σ^2: $u \sim \text{Normal}(0, \sigma^2)$
- Este supuesto no es esencial, puede ser obviado siempre que se disponga de un tamaño muestral (n) lo suficientemente grande.

Distribución normal homocedástica con una sola variable explicativa



El estadístico de prueba *t*

Bajo este nuevo supuesto se tiene

$$t = \frac{\left(\hat{\beta}_j - \beta_j\right)}{ee\left(\hat{\beta}_j\right)} \sim t_{n-k-1}$$

es decir, el estimador del coeficiente de x_j posee una distribución $t - Student$ (n-k-1)

Notar que $n - k - 1$ son los grados de libertad

Prueba de Significatividad Individual

- Conocer la distribución muestral de *t* nos permite llevar a cabo pruebas de hipótesis.
- La prueba mas utilizada en el análisis de regresión es la de "*significatividad individual del regresor* x_j"
- Por ejemplo, $H_0: \beta_j = 0$ vs. , $H_1: \beta_j \neq 0$
- Si aceptamos H_0, entonces aceptamos que x_j *no afecta* a *y*, controlando por las otras *x*'s

Prueba de Significativ. Individual II

Para efectuar la prueba primero hay que computar el estdístico t para $\hat{\beta}_j$ suponiendo que $H_0 : \hat{\beta}_j = 0$ es verdadera :

$$t_{\hat{\beta}_j} = \frac{\hat{\beta}_j}{ee\left(\hat{\beta}_j\right)}$$

Luego usamos el estadístico t, junto con su correspondiente p - valor, para determinar si rechazamos o no la hipótesis nula, H_0

Resumen: $H_0: \beta_j = 0$ vs. $H_0: \beta_j \neq 0$

- Normalmente se prueba la significatividad individual mediante una prueba bilateral.
- Si se rechaza H_0, decimos que "x_j es estadísticamente significativa"
- Si no es rechazada H_0, decimos que "x_j NO es estadísticamente significativa"
- El Stata reporta los p-valores de cada coeficiente estimado en la tabla de resultados.

Pruebas de otras hipótesis

- Una forma más general del estadístico *t* reconoce la posibilidad de que se quiera testear una hipótesis como H_0: $\beta_j = a_j$
- En este caso, el estadístico *t* apropiado es

$$t = \left(\hat{\beta}_j - a_j\right) \Big/ ee\left(\hat{\beta}_j\right), \text{ donde}$$

$a_j = 0$ para la prueba vista anteriormente

Intervalos de Confianza (IC)

- Al igual que para el caso de la media, es posible computar un intervalo de confianza (IC) para un coeficiente de regresión.
- Un IC al Nivel de Confianza del (1 - α)% es

$$\hat{\beta}_j \pm t_c \bullet ee\left(\hat{\beta}_j\right), \text{ donde } t_c \text{ es el percentil } \left(1-\frac{\alpha}{2}\right)$$

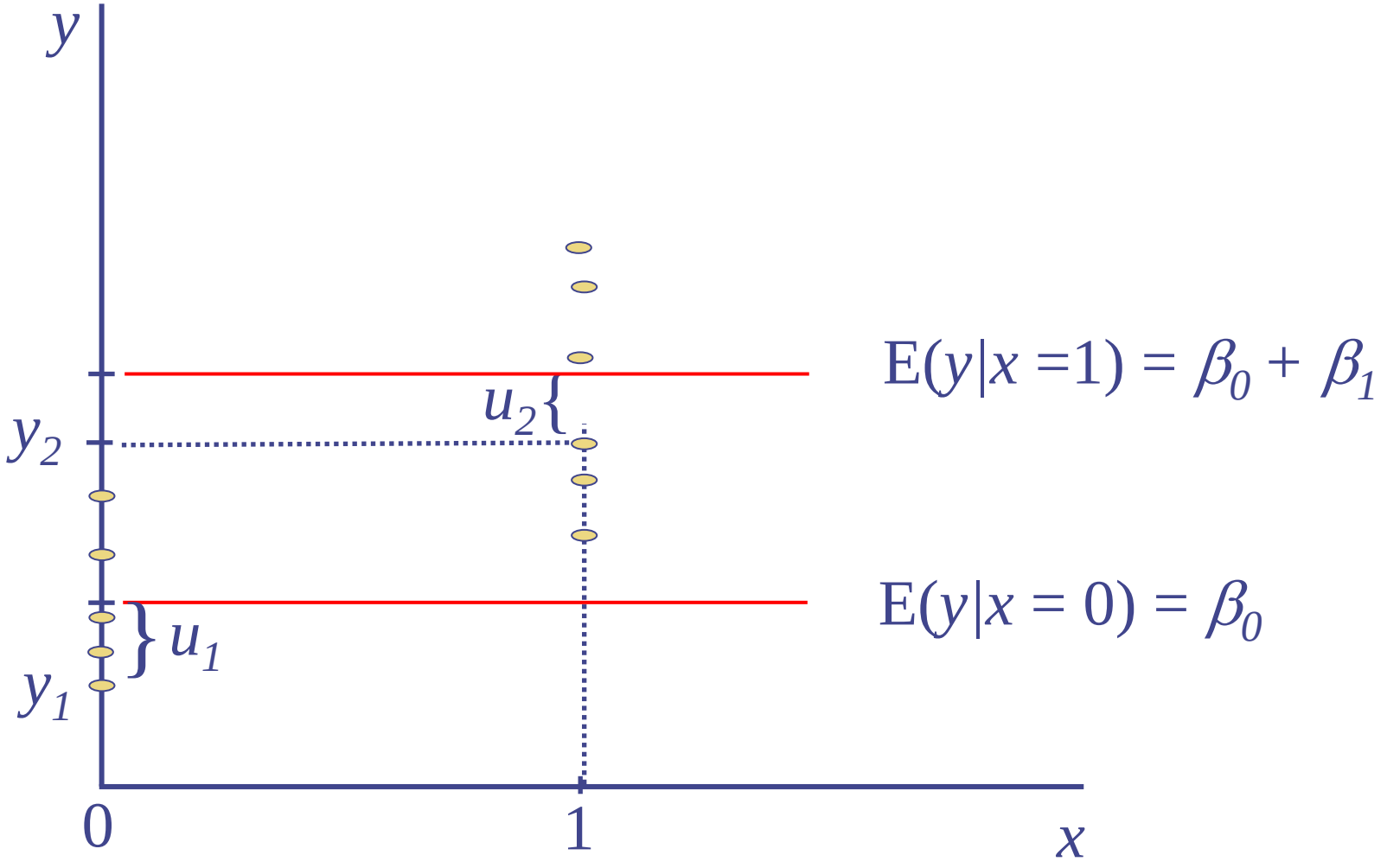
de una distribución t_{n-k-1}.

- Stata reporta automáticamente el IC del 95% para cada uno de los coeficientes estimados.

Variables Cualitativas y Binarias

- Las variables *cualitativas* o *categóricas* nos indican la presencia o ausencia de ciertas características en los objetos bajo estudio
- Dichas variables se modelan a través del uso de variables denominadas *binarias* o *ficticias*
- Una variable binaria es una variable que puede tomar valor 1 o 0
- Ejemplo: *hombre* (= 1 si es hombre, 0 si no lo es), *sur* (= 1 si es del sur, 0 si no lo es), etc.

Línea de regresión poblacional, siendo x una variable binaria $x = \{0, 1\}$



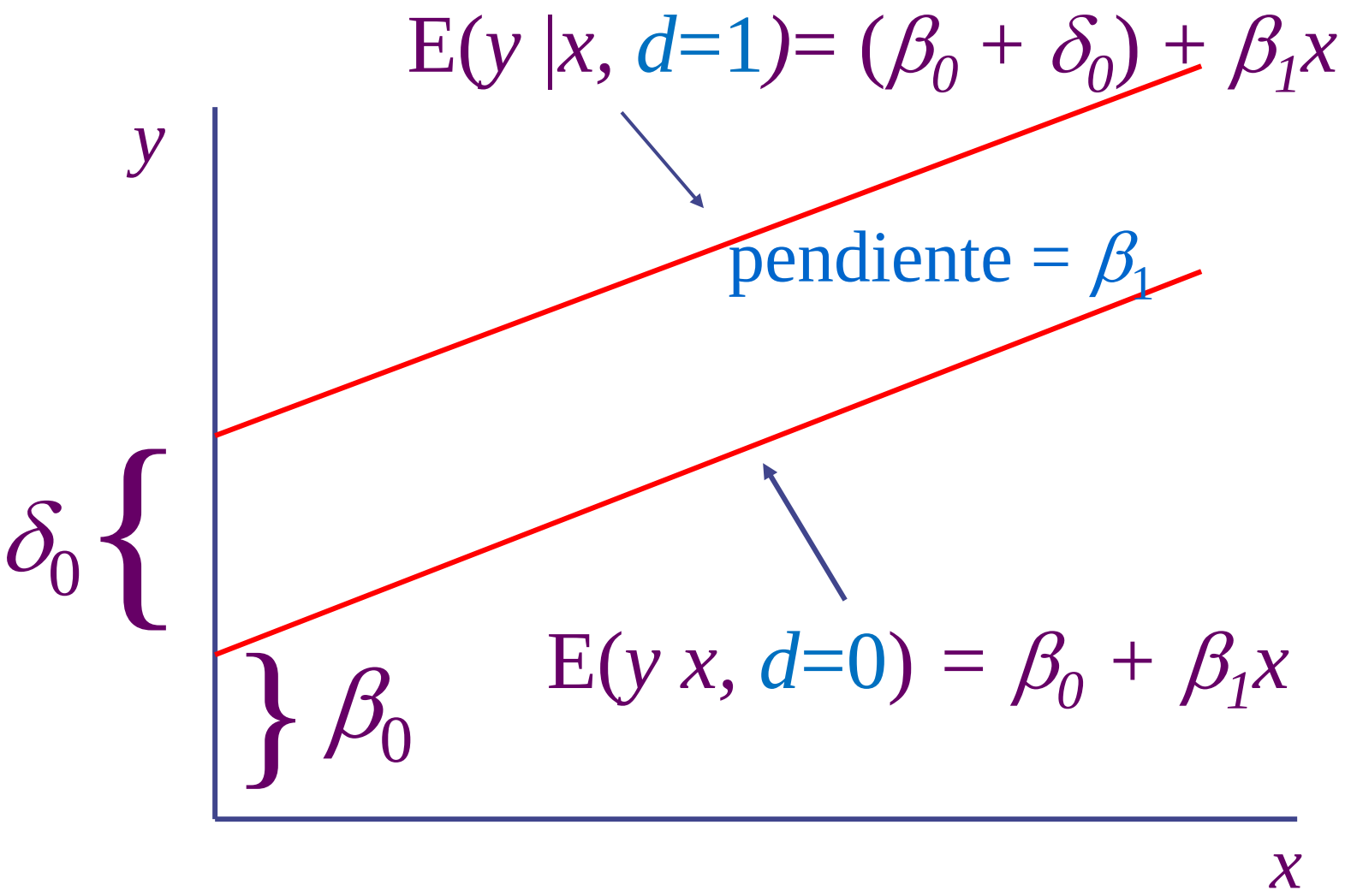
Un regresor binario y otro continuo

- Consideremos un modelo con una variable continua (x) y con una variable binaria (d)

$$y = b_0 + \delta_0 d + b_1 x + u$$

- Esto puede ser interpretado como un cambio de *intercepción* de la función de regresión
- Si $d = 0$, tenemos $y = b_0 + b_1 x + u$
- Si $d = 1$, tenemos $y = (b_0 + \delta_0) + b_1 x + u$
- El caso de $d = 0$ es el llamado *grupo base*

Ejemplo con $\delta_0 > 0$



Caso de Múltiples Categorías

- Ejemplo: dada la variable cualitativa

  educ = {primaria, secundaria, universitaria}

- Por ejemplo, para comparar la educación *secundaria* y la *universitaria* respecto de la primaria, usamos 2 variables binarias:

  secun = 1 si tiene secundaria, 0 si no tiene; y *univ* = 1 si tiene universitaria, 0 si no tiene.

  La categoría contra la cual se compara (en este caso primaria) es el llamado *grupo base*.

Múltiples Categorías (continuación)

- Cualquier variable categórica puede ser modelada mediante un conjunto de binarias
- Dado que el grupo base es representado por el intercepto, si hay n categorías deben utilizarse $n - 1$ variables binarias
- Si hay demasiadas categorías, puede ser útil agruparlas en algún grado.
- Ejemplo: ranking del 1 al 20, agrupo en top 5, luego de 6 a 10, luego de 11 a 15, etc.

Interacción entre Binaria y Continua

- Es posible también la interacción entre una binaria, *d*, y una variable continua, *x*:

$$y = b_0 + \delta_0 d + b_1 x + \delta_1 d*x + u$$

- Si $d = 0$, tenemos $y = b_0 + b_1 x + u$
- Si $d = 1$, queda $y = (b_0 + \delta_0) + (b_1 + \delta_1)\, x + u$
- Esto se interpreta como un cambio en la pendiente de la función de regresión (además del cambio en la ordenada).

Ejemplo con $\delta_0 > 0$ y $\delta_1 < 0$