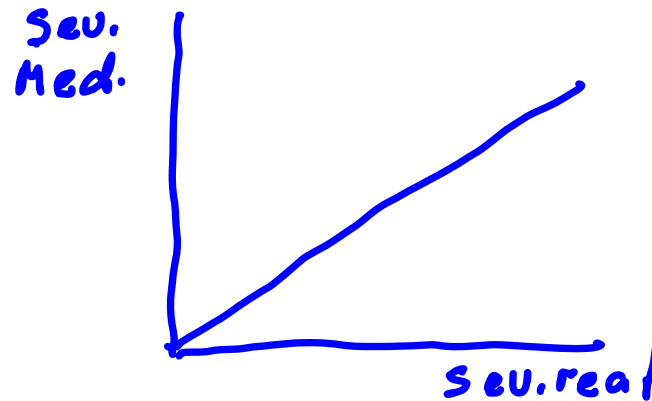
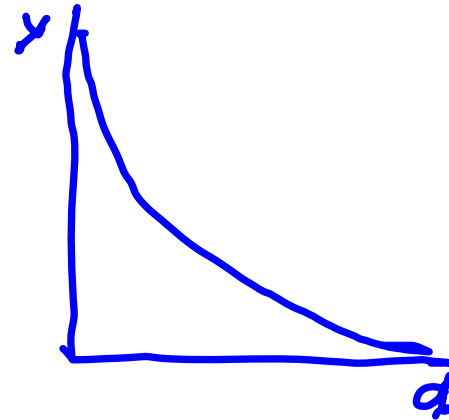
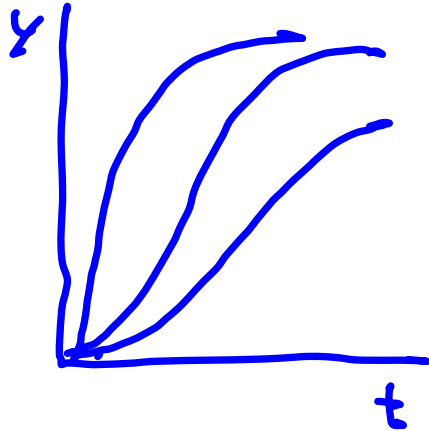


La necesidad de los modelos para las relaciones en epidemiología

- Las epidemias involucran cambio de intensidades sobre el tiempo y el espacio en una población de plantas
- Dado que no se mide u observa cada uno de los individuos en una población, se deben usar **modelos** para representar o resumir el comportamiento de la población
- Más allá, en la epidemiología nosotros estamos particularmente interesados en relaciones (a nivel de poblaciones)
 - Los modelos son necesarios para resumir, en un sentido cuantitativo, las relaciones de interés

La necesidad de los modelos para las relaciones en epidemiología



Modelos:

- **Son una simplificación de la realidad**
- Un modelo es una abstracción de un fenómeno real o proceso que enfatiza aquellos aspectos relevantes para los objetivos impuestos de estudio
- Son usados para **describir, entender, predecir, comparar, comunicar** y hacer **inferencias** acerca de un fenómeno
 - Un modelo dado puede ser usado para uno o más de esas necesidades

Nota: el concepto general de un modelo es mucho más amplio que una **ECUACIÓN**.

De hecho, los modelos no necesitan ser una ecuación

Modelos

- **Modelos mentales**

Es una imagen de cómo es imaginada/o algo que existe

- **Modelos tangibles**

Es una imagen mental transformada en una forma explícita

Mapas, diagramas de flujos, modelos físicos, ecuaciones

- Usados para comunicarse y pueden ser tanto **Físicos** como **Abstractos**

- **Modelos Físicos**

Reproducen la forma o función de un objeto o sujeto real, escalado o simplificado (avión, ADN)

- **Modelos Abstractos** (símbolos en sentido amplio)

- **Cualitativos**

(diagramas de flujos, gráficos,.....)

- **Cuantitativos** (poseen reglas y anotaciones matemáticas)

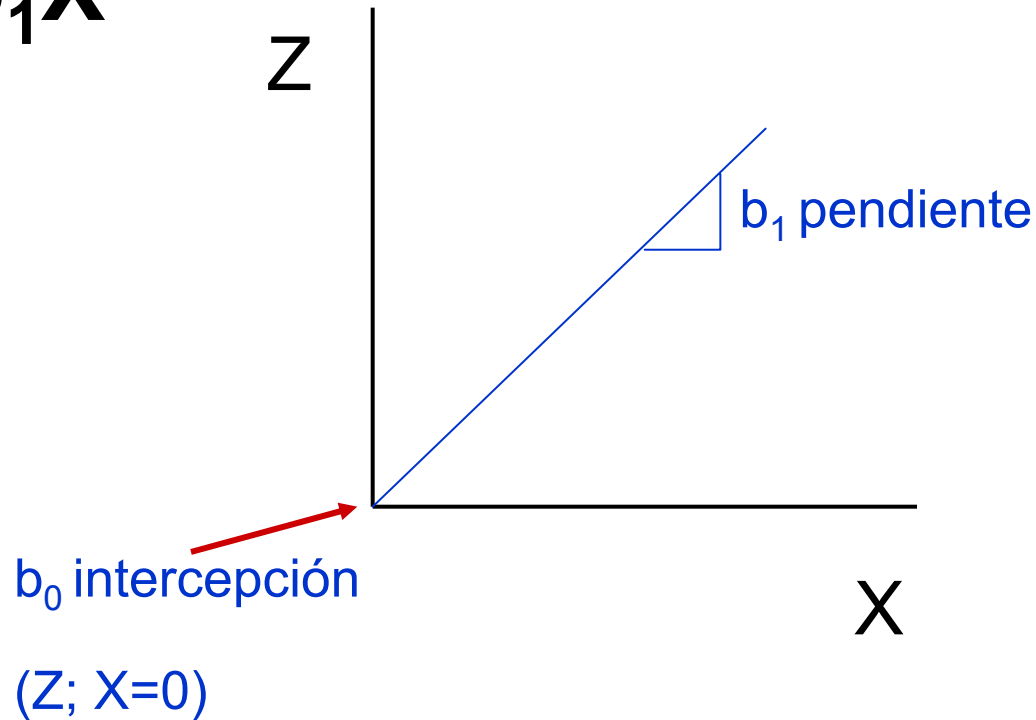
- **Matemáticos**

- **Estadísticos**

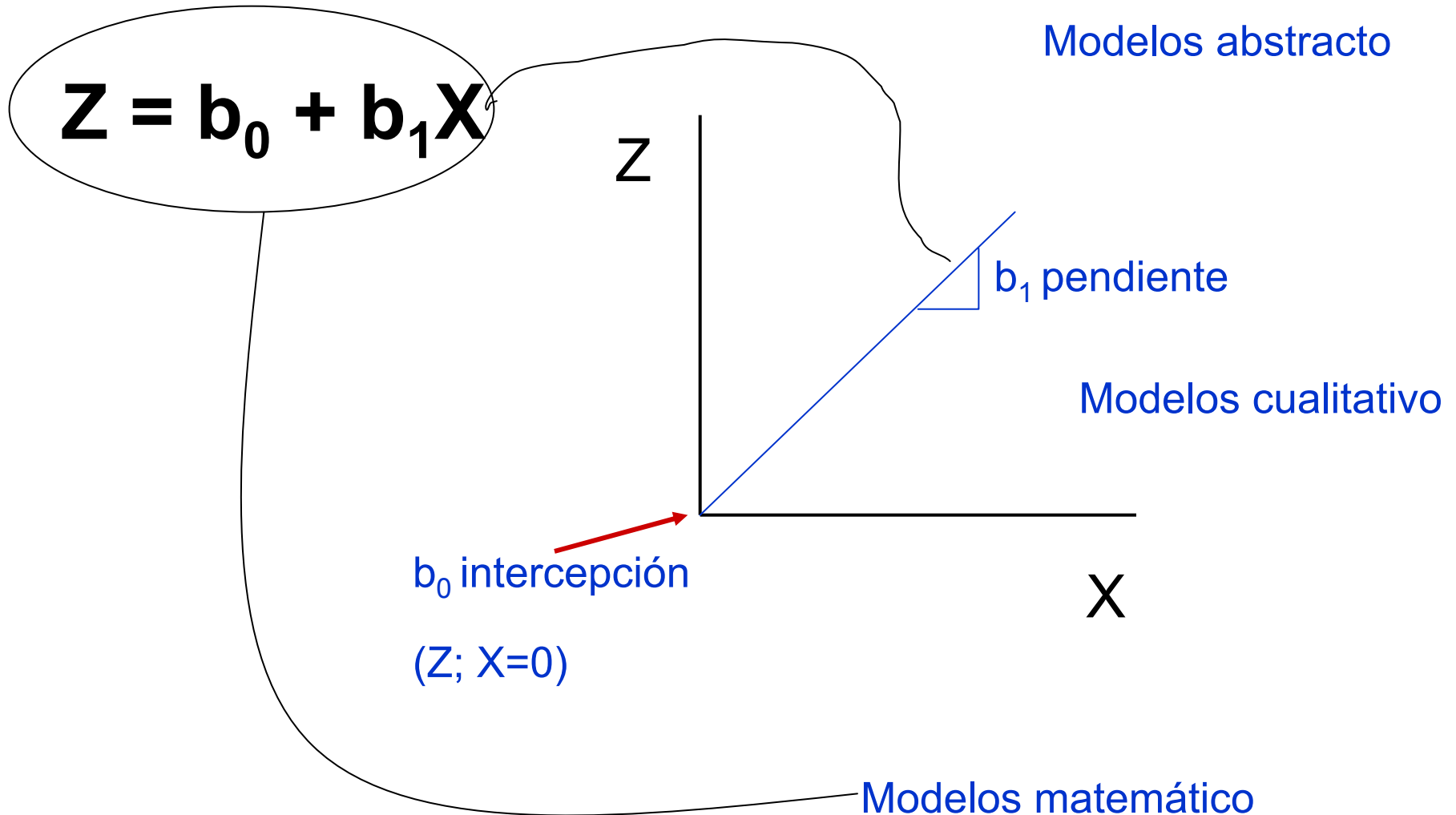
(Son un tipo especial de modelos donde explícitamente unas la aleatoriedad y la variación)

Modelo: esporas/lesiones (Z) en relación al área e la lesión (X)

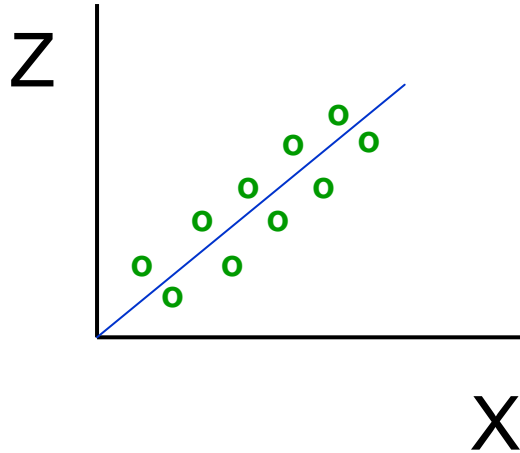
$$Z = b_0 + b_1 X$$



Modelo: esporas/lesiones (Z) en relación al área e la lesión (X)



Modelo: esporas/lesiones (Z) en relación al área e la lesión (X)

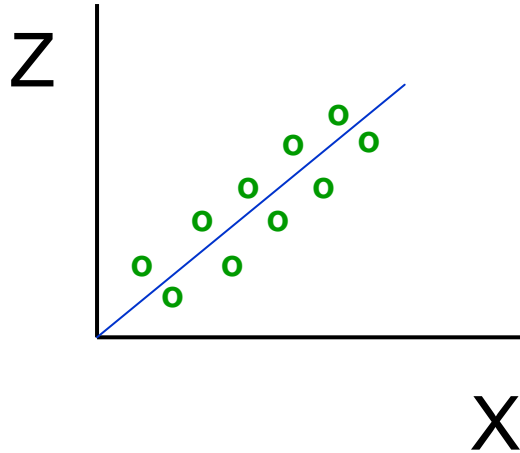


$$Z = b_0 + b_1X + e$$

e = “término error” que representa la variabilidad no explicada **“RUIDO”**

- Hay diferencia con el gráfico anterior
- Antes solo existía una línea sola
 - esto significa que por cada lesión de un tamaño dado, tiene un valor de esporulación igual y único
- Ahora, este gráfico muestra que para cada tamaño de lesión, hay un rango de valores posibles de esporulación
 - El **“valor central”** de esporulación se incrementa linealmente

Modelo: esporas/lesiones (Z) en relación al área e la lesión (X)



$$Z = b_0 + b_1X + e$$

e = “término error” que representa la variabilidad no explicada **“RUIDO”**

Modelo estadístico: consideración explícita de la variabilidad o aleatoriedad

- Sin e , Z es siempre la misma para cualquier valor dado de
 - Si $X=10$, $b_0=0$, $b_1=100$
 - $Z = 0 + 100 \cdot 10 = 1000$ esporas por cada lesión (no es algo probable)
- Se asume, estadísticamente, que e es una **variable aleatoria**
- **A menudo, se asume que e es normal con media 0 y varianza constante**

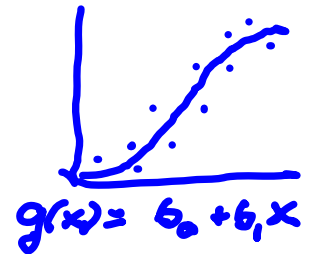
$$Z = b_0 + b_1X + e$$

- **e** representa a **todas** las variables (factores) que afectan a Z además de X
- La teoría estadística puede (a menudo) justificar el supuesto de normalidad (porque muchas variables afectan Z, en general)
- Algunas veces (para algunas lesiones), Z será más **grande** que $b_0 + b_1X$, y otras será **menor** que $b_0 + b_1X$
 - Pero, en promedio, Z será igual a $b_0 + b_1X$
 - Esto es en consecuencia de la media igual 0 de e
- Por lo tanto el promedio, esperado, o media de Z es igual a $b_0 + b_1X$
 - Escrito como $E(Z) = b_0 + b_1X$
 - La esperanza es la media de la población
- **La línea está dada por $b_0 + b_1X$, (valores esperados) desviaciones de la línea dada por e**

$$Z = b_0 + b_1X + e$$

$$E(Z) = b_0 + b_1X$$

- Estas son expresiones equivalentes para muchos modelos comunes estadísticos
- Este modelo es para una línea recta
- En el uso regular de modelos parecidos a éste (ej. En predicciones), el “error” no es usado directamente
 - Por lo tanto, se está prediciendo la media de Z a un valor dado de X
- Muchas relaciones no son descritas por una línea recta
- Una forma más general de la expresión es:
 - **$Z = g(X) + e$, ó $E(Z) = g(X)$**
 - $g(x)$ es una anotación par alguna función de X
 - Esta expresión enfatiza que un componente aleatorio está en el modelo



Primero vimos $Z = b_0 + b_1X$

Conocido como un modelo determinístico, porque Z es completamente determinado por X (o sea que no existe aleatoriedad o variabilidad de Z a un valor dado de X)

Luego vimos $Z = b_0 + b_1X + e$

Componente
determinístico
del modelo estadístico

Componente estocástico (o aleatorio)
del modelo estadístico

$Z = b_0 + b_1 X + e$

Variable respuesta o dependiente

Constantes (= **parámetros**), el cuál se estima a partir de los datos

Predictor o variable independiente

Término del error (variable aleatoria), con media = 0. A menudo se asume con distribución normal, todas las observaciones son independientes, y la varianza de esta variable es constante

Una definición más refinada del modelo estadístico:

Modelo con componentes estocásticos (y pueden ser con otros componentes) conteniendo constantes desconocidas (parámetros) que son motivo de estimación

$$\varepsilon \sim \text{NID}(0, \sigma^2)$$

$$Z = b_0 + b_1X + e$$

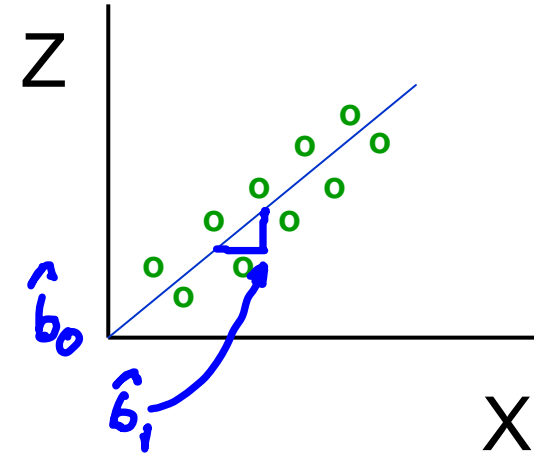
- Se necesita estimar los parámetros (b 's)

- Se colocan “sombrosos” a los datos estimados

- Aquí estamos interesados acerca de estimar aquellos basados en los datos disponibles (observaciones)

- La forma más común es a través de la regresión por **mínimos cuadrados**

- Para datos con distribución normal, esto es lo mismo que el **método de máxima verosimilitud**

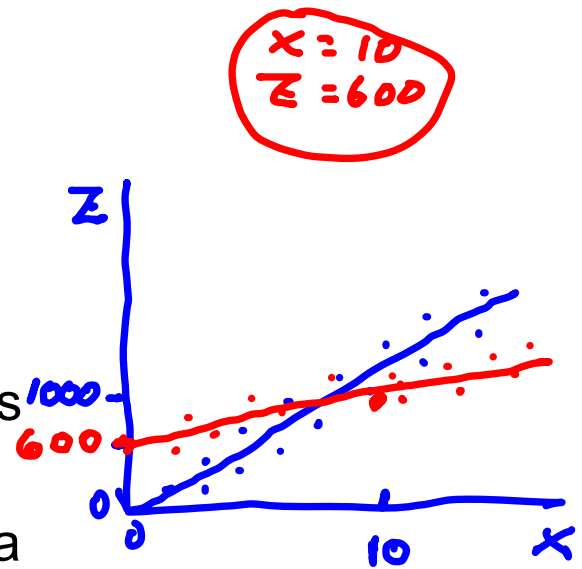


$$\hat{Z} = \hat{b}_0 + \hat{b}_1 X + e$$

Handwritten blue annotations: \hat{Z} is labeled "ajustado predicho" (circled), \hat{b}_0 is labeled "estimado", and \hat{b}_1 is labeled "estimado".

$$Z = b_0 + b_1X + e$$

- **Regresión por mínimos cuadrados**
- Equivalente a encontrar **la mejor** línea a través de los puntos
- Considerar la diferencia entre la posible línea a un X dado y un dato
 - Para la línea azul: $Z=1000$ a $X=10$, pero el **dato real es $Z=600$**
 - Para la **línea roja**: $Z=900$ siendo lo más cercano a la observación; pero tiene un ajuste pésimo a lo largo de las Xs
- Se podría intentar un gran número de líneas y encontrar la de mejor ajuste



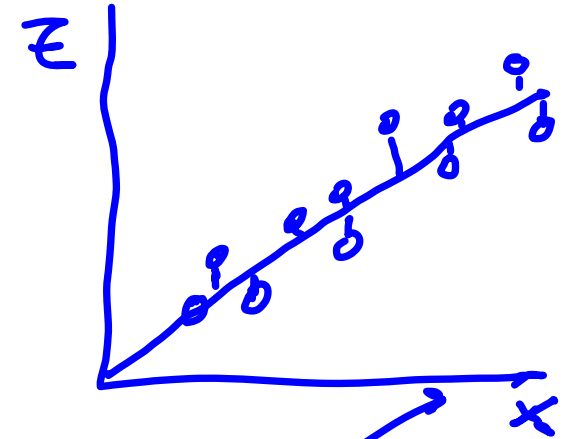
$$Z = b_0 + b_1X + e$$

•Bondad del ajuste

Se puede encontrar el total de la diferencias de cuadrados entre Z y \hat{Z}_i para cada “set” de estimación e parámetros (para cualquier línea)

$$Q = \sum (Z_i - \hat{Z}_i)^2 = \sum (Z_i - (b_0 + b_1 X_i))^2$$

subscripto para
identificar una
observación



El valor de parámetro que da el menor Q son los **Cuadrados Mínimos Estimados (CME ó LSE)**

El menor valor de Q es conocido como el **error de la suma de cuadrados (ESC ó SSE)**

Por suerte, esto se encuentra a través del cálculo (uno no tiene necesidad de intentar lotes de posibles líneas)

$$\hat{Z} = \hat{b}_0 + \hat{b}_1 X + e$$

$$Z = b_0 + b_1 X + e$$

- No se necesita intentar todos los posibles parámetros: con la regla de los mínimos cuadrados, existe una solución única (obtenida por cálculo)
- Virtualmente todos los programas estadísticos generan análisis de la regresión usando este método
- Cuando se predice Z usando los parámetros estimados, se está prediciendo la respuesta media

Mínimos cuadrados ordinarios

- Un perfecto ajuste es cuando $SCE=0$ o $CME=0$
- Un ajuste pobre: alto SCE
- Una aproximación: calcular una medida relativa de los que el modelo explica
 - Calcular la variabilidad total alrededor de la media de Z (línea horizontal)
 - Esta es la línea de “no relación”
 - Es lo mismo que una línea con pendiente = 0
 - Se denomina a esto suma total de cuadrados (STC)

Mínimos cuadrados ordinarios

- SCT/SCE es la proporción de variabilidad de Z que no es explicada por el modelo
- Luego el $R^2 = 1 - \text{SCE/SCT}$ es la proporción de variabilidad de Z que es explicada por el modelo
 - Conocido como coeficiente de determinación
 - Varía entre 0 a 1 (ó 0 a 100%)
 - R (raíz cuadrada de R^2 es conocido como coeficiente de correlación
- Nota: SCE (o CME) también es la base para un test estadístico (un **test de F**) para determinar si hay una relación entre Z y X
- Los parámetros de desviaciones estándar estimadas:
Función de CME

$$\underset{\wedge}{Z} = \underset{\wedge}{b_0} + \underset{\wedge}{b_1}X + e$$

$$Z = b_0 + b_1X + e$$

- Nota: Como una situación teórica de una línea conocida (0modelo conocido con parámetros conocidos), la línea ajustada (= modelo con parámetros estimados) no necesariamente corre a través de cada punto
- Para una línea ajustada, la diferencia entre la observación y la línea es el residual (e)
- e es una estimación de ε