Modelos y modelación

Todos los modelos están equivocados, pero algunos son útiles

J, Tukey & G. Box

Una teoría tiene solo la alternativa de estar errada o acertada. Un modelo tiene una tercera posibilidad: puede estar acertado pero ser irrelevante

M. Elgen

Nunca creas en un modelo hasta que haya estado validado con datos

Anónimo

Nunca creas en tus datos hasta que hayan sido validado por un modelo

?????

Modelos estadísticos:

$Z = b_0 + b_1X + \varepsilon, \qquad \varepsilon \sim NID(0,\sigma^2)$

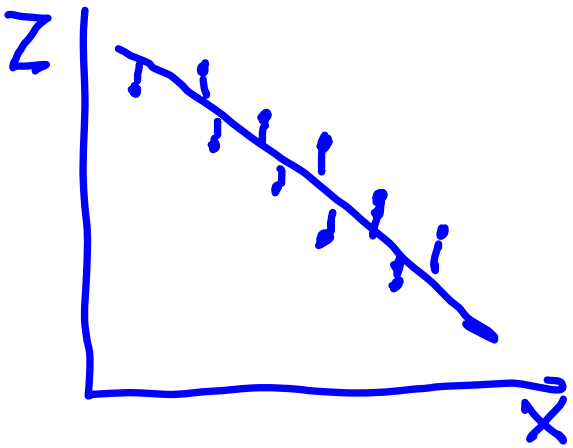
$E(Z) = b_0 + b_1X$

$\hat{Z} = b_0 + b_1X$

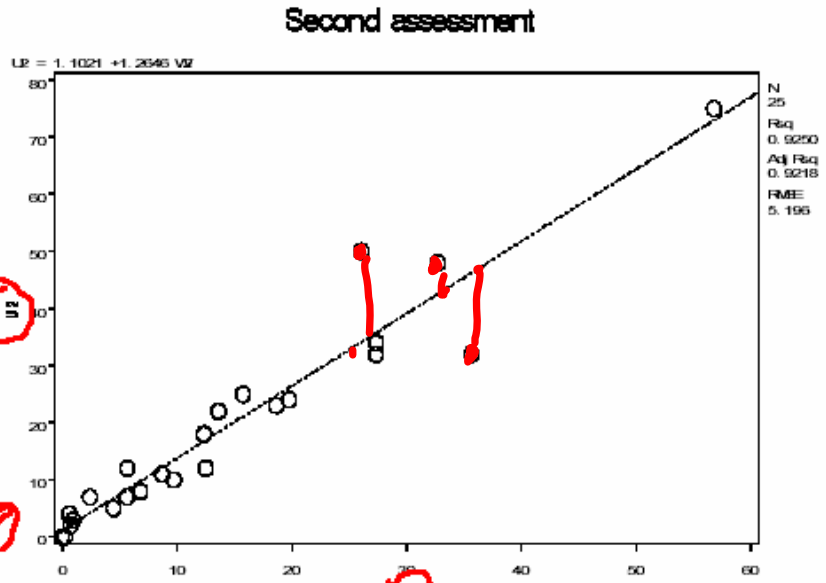
$\hat{b}_0$ y $\hat{b}_1$ corresponden a

$SSE = \sum (Z_i - \hat{Z}_i)^2$

$\hat{\sigma} = \frac{SSE}{N-2}$

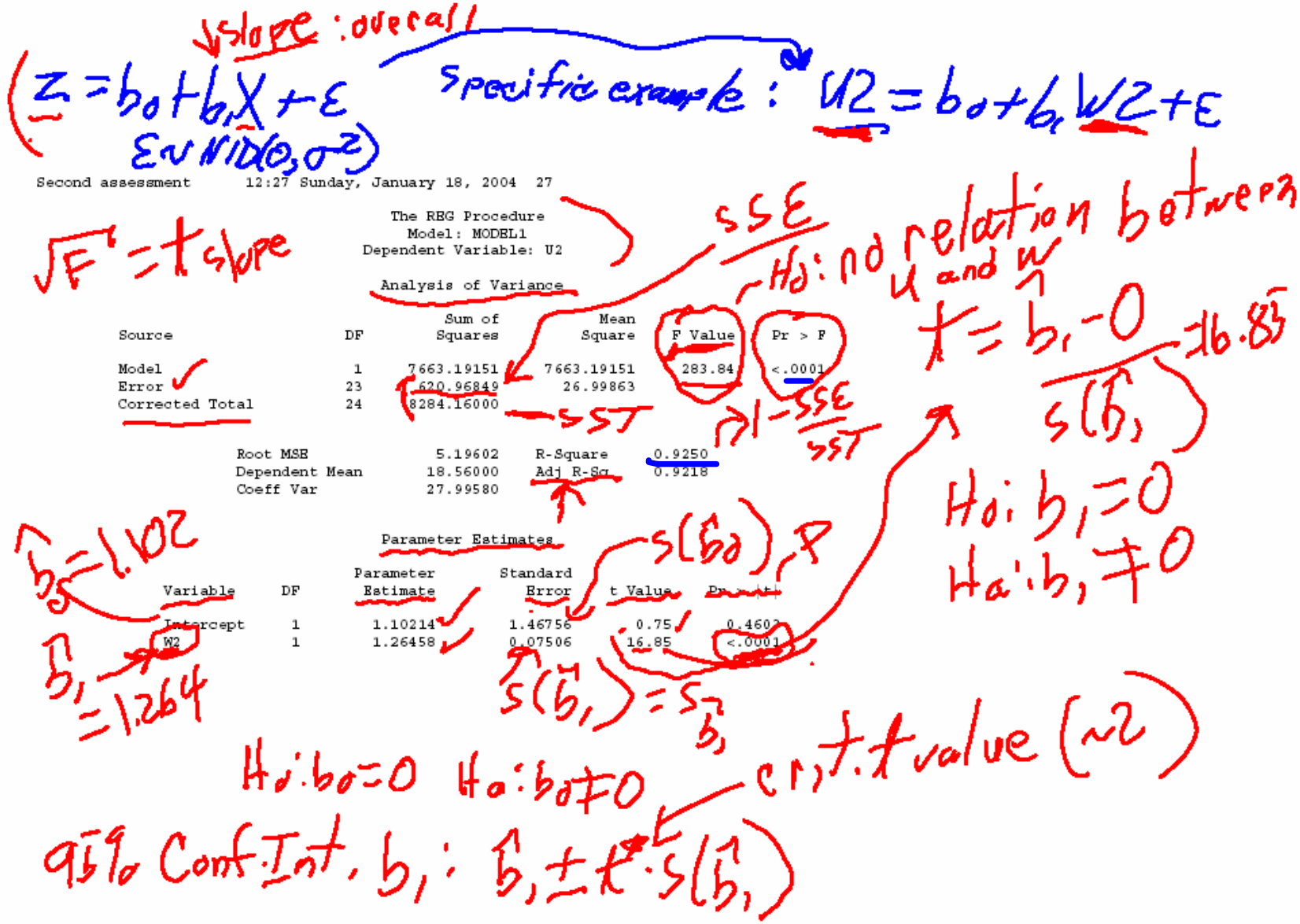


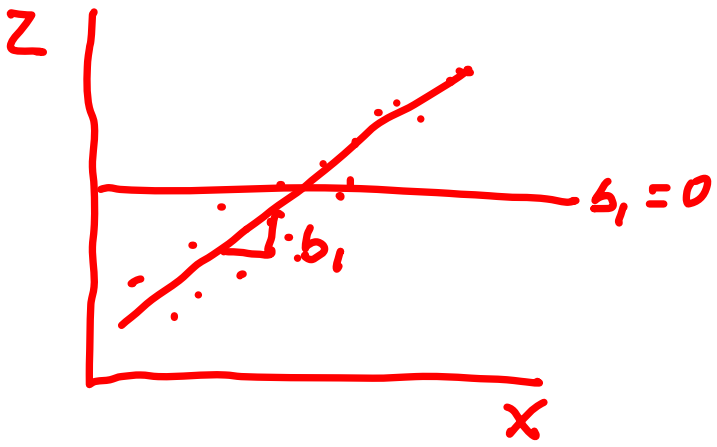
```
*-- True (W_) and measured severity (U_) at times 1 and 2;
data severity;
input W1 U1 W2 U2;
datalines;
1.0200  3   0.0000  0
0.5430  1   0.0000  0
0.0444  0   0.2464  0
3.7885  5   0.6289  4
0.0510  9   0.7205  2
1.5356  5   0.9631  3
2.5480  3   2.4224  7
1.8004  11  4.4856  5
5.7013  6   5.6566  7
2.8122  13  5.6828  12
6.4677  5   6.8447  8
6.1824  7   8.7121  11
12.6446 4   9.7677  10
10.6724 19  12.3885 18
12.5370 17  12.5156 12
11.4292 21  13.6486 22
14.6693 23  15.7731 25
21.4179 24  18.6935 23
22.9403 24  19.7970 24
23.0519 54  26.0992 50
26.2719 32  27.3776 32
29.1525 34  27.3936 34
32.2839 51  32.7840 48
38.4818 33  35.7091 32
57.6851 72  56.8222 75
;
```




```
title 'Second assessment';
proc reg data=severity;
      model U2 = W2;
plot U2*W2;
plot r.*W2;
run;
```

SAS input


slope : overall
$Y = b_0 + b_1 X + \varepsilon$
$\varepsilon \sim NID(0, \sigma^2)$
Specific example: $U2 = b_0 + b_1 W2 + \varepsilon$
Second assessment 12:27 Sunday, January 18, 2004 27
The REG Procedure
Model: MODEL1
Dependent Variable: U2
Analysis of Variance
Source DF Sum of Squares Mean Square F Value Pr > F
Model 1 7663.19151 7663.19151 283.84 <.0001
Error 23 620.96849 26.99863
Corrected Total 24 8284.16000
Root MSE 5.19602 R-Square 0.9250
Dependent Mean 18.56000 Adj R-Sq 0.9218
Coeff Var 27.99580
Parameter Estimates
Variable DF Parameter Estimate Standard Error t Value Pr > |t|
Intercept 1 1.10214 1.46756 0.75 0.4602
W2 1 1.26458 0.07506 16.85 <.0001
$\sqrt{F} = t_{slope}$
SSE
SST
H_0: no relation between U and W
$1 - \frac{SSE}{SST}$
$t = \frac{b_1 - 0}{s(b_1)}$
16.85
$H_0: b_1 = 0$
$H_a: b_1 \neq 0$
$\hat{b}_0 = 1.102$
$\hat{b}_1 = 1.264$
$s(\hat{b}_0)$, P
$s(\hat{b}_1) = s_{\hat{b}_1}$
$H_0: b_0 = 0$ $H_a: b_0 \neq 0$
CI, t, t value (~2)
95% Conf. Int. b_1: $\hat{b}_1 \pm t^* \cdot s(\hat{b}_1)$



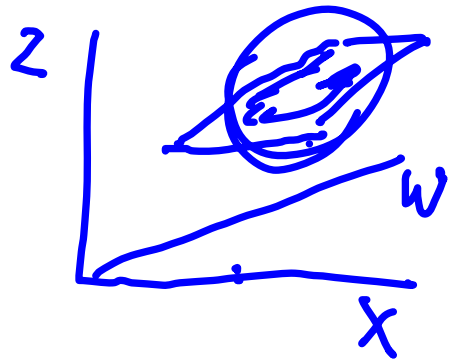
$$Z = b_0 + b_1 X_1 + b_2 W_2 + \varepsilon$$

Regresión Múltiple

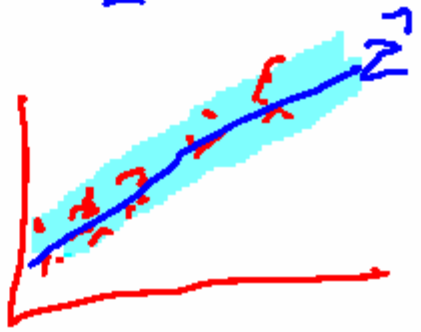
X_1: Temp

W_2: HR

N° de lesiones



Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		Residual
1	0	1.1021	1.4676	-1.9337	4.1380	-1.1021
2	0	1.1021	1.4676	-1.9337	4.1380	-1.1021
3	0	1.4137	1.4546	-1.5952	4.4227	-1.4137
4	4.0000	1.8974	1.4346	-1.0703	4.8652	2.1026
5	2.0000	2.0133	1.4299	-0.9447	4.9712	-0.0133
6	3.0000	2.3201	1.4174	-0.6121	5.2522	0.6799
7	7.0000	4.1654	1.3453	1.3824	6.9485	2.8346
8	5.0000	6.7745	1.2527	4.1831	9.3660	-1.7745
9	7.0000	8.2553	1.2058	5.7609	10.7498	-1.2553
10	12.0000	8.2885	1.2048	5.7961	10.7809	3.7115
11	8.0000	9.7578	1.1631	7.3516	12.1639	-1.7578
12	11.0000	12.1192	1.1073	9.8286	14.4099	-1.1192
13	10.0000	13.4541	1.0825	11.2148	15.6934	-3.4541
14	18.0000	16.7683	1.0446	14.6073	18.9293	1.2317
15	12.0000	16.9291	1.0437	14.7700	19.0881	-4.9291
16	22.0000	18.3618	1.0393	16.2119	20.5117	3.6382
17	25.0000	21.0484	1.0496	18.8771	23.2198	3.9516
18	23.0000	24.7415	1.1021	22.4617	27.0213	-1.7415
19	24.0000	26.1369	1.1323	23.7945	28.4794	-2.1369
20	50.0000	34.1065	1.3898	31.2316	36.9815	15.8935
21	32.0000	35.7232	1.4553	32.7127	38.7336	-3.7232
22	34.0000	35.7434	1.4561	32.7312	38.7556	-1.7434
23	48.0000	42.5600	1.7633	38.9123	46.2077	5.4400
24	32.0000	46.2590	1.9450	42.2354	50.2825	-14.2590
25	75.0000	72.9581	3.3920	65.9413	79.9749	2.0419



Análisis de la regresión por mínimos cuadrados ordinarios

- Los residuales son extremadamente importantes en el análisis y modelaje $\hat{\varepsilon} = e = z_i - \hat{z}_i$
  - Estimación del error (variación no explicada)
  - La suma de cuadrados de los residuales es la misma que la suma del cuadrado del error $SCR = \Sigma(z_i - \hat{z})^2 = \Sigma e^2$
  - Usado intensivamente para la evaluación del modelo
    - Es el modelo razonable para ajustar el set de datos?
    - Hay otros modelos más razonables?
    - Se han cumplido los presupuestos estadísticos?
    - Ver el gráfico de residuales versus X

Modelos

- Correcciones a un pobre ajuste son sugeridos por el gráfico de residuales

- Primero consideremos la modelación/estrategias de análisis
  - Descriptivos, correlativos o empíricos
  - Teóricos, mecánicos, o explicativos

Modelos empíricos

Describen datos basados en la aceptación de principios estadísticos sin usar previamente el desarrollo de teoría o conceptos para la relación entre las variables respuesta y predictora

Secuencia

- Obtención de datos (medida de la intensidad de la enfermedad y variables predictivas)
- Describir la relación usando modelos (**simples**)
- Evaluar el ajuste del modelo (residuales, estadísticos (R^2), etc.
- Usar otros modelos, si la elección del primero no es razonable
- Si el modelo seleccionado es **razonable** usarlo para:
  - Predecir, comparar, evaluar significancia, inferencia, y otros

Modelos empíricos

Comenzar con un ajuste lineal

Ver residuales vs. X = ajusta, el modelo es razonable

= no ajusta

Aplicar una transformación ej. log(Z)

= ajusta, el modelo es razonable

Transformaciones usuales:

Para la variable respuesta: Z

RC(Z), ln(Z), 1/Z

Para la variable predictora: X

RC(X), ln(X), 1/X

Modelos teóricos

Comienzan con un concepto o teoría de la realidad a ser descripta, **no con datos** El modelo desarrollado está basado en un concepto o teoría y luego se recolectan los datos. La prueba del modelo se realiza por su pertinencia de acuerdo a como ajusta el modelo con los datos y usando análisis matemático y estadístico de las propiedades del modelo

Existen interconexiones entre modelos empíricos y teóricos

- La recolección de datos y el modelo empírico puede ser la base para el desarrollo de nuevos conceptos/teoría
- Cuando un modelo teórico es desarrollado, se debe probar con los datos usando los principios del modelaje empírico para determinar si el modelo es apropiado

Modelos teóricos

Comienzan con un concepto o teoría de la realidad a ser descripta, **no con datos** El modelo desarrollado está basado en un concepto o teoría y luego se recolectan los datos. La prueba del modelo se realiza por su pertinencia de acuerdo a como ajusta el modelo con los datos y usando análisis matemático y estadístico de las propiedades del modelo

Existen interconexiones entre modelos empíricos y teóricos

- La recolección de datos y el modelo empírico puede ser la base para el desarrollo de nuevos conceptos/teoría
- Cuando un modelo teórico es desarrollado, se debe probar con los datos usando los principios del modelaje empírico para determinar si el modelo es apropiado

Últimamente, un modelo teórico es considerado superior a un modelo empírico

- Los puntos de datos observados están fuera de los límites para ser al menos algo como imprecisos, originando resultados empíricos engañosos en un estudio único

- Si nosotros realmente entendemos un fenómeno, podemos especificar el modelo (incluyendo los valores de los parámetros) antes de colectar los datos
  - Comparando a la física o química, la biología tiene pocos modelos teóricos

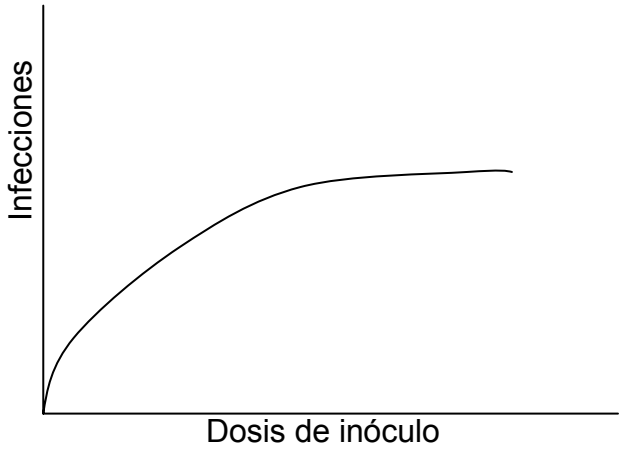
- En la práctica, muchos investigadores usan ambas aproximaciones dependiendo del estudio en cuestión

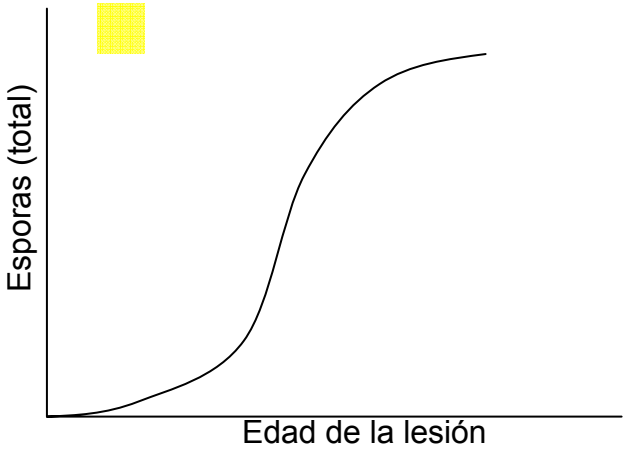
- Esfuerzos en modelación intermedia, combinando el mejor modelo empírico y el mejor modelo teórico, son comunes en epidemiología y en dinámica y ecología de las poblaciones
  - Por ejemplo, la “familia” de modelos de las curvas de progreso de enfermedades está basada en consideraciones teóricas, pero el modelo específico usado y los parámetros de los modelos pueden ser determinados a través de datos observados
  - Esta aproximación es la que se sigue en este curso

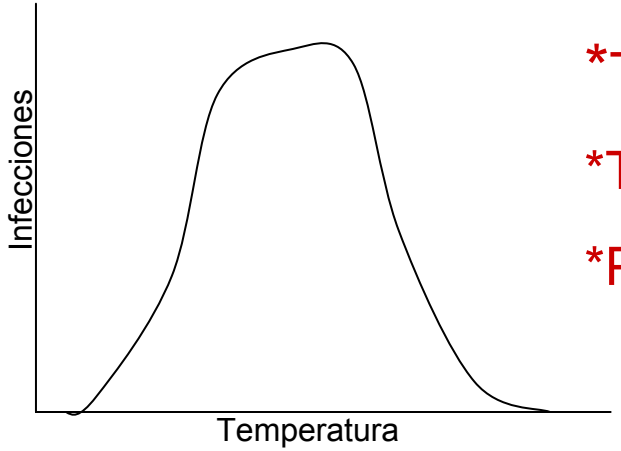
Modelos

- Basados en consideraciones conceptuales o teóricas de los mecanismos del crecimiento de las enfermedades en poblaciones (dinámica de la enfermedad) casi siempre requieren de modelos no lineales
- Las relaciones observadas a menudo sugieren modelos no lieneales
- Un modelo linear **no** necesariamente significa una línea recta
  - Pero una línea recta puede ser considerada la expresión gráfica de un modelos lineal
- Linealidad o no linealidad tiene que ver con la propiedad técnica de los modelo respecto a la forma en el cual los parámetros aparecen en él
  - El concepto de modelo lineal o no lineal para los biólogos debe quedar bien claro

Modelos: algunas relaciones







*Todas las curvas son no lineales

*Todas son intrínsecamente lineales

*Porciones de las curvas también pueden ser lineales

Modelos lineales: Ejemplos

$$Z = \alpha + \beta X$$

$$Z = \alpha + \sqrt{\beta X}$$

$$Z = \alpha + \beta_1 X + \beta_2 X^2$$

$$Z = \alpha + \beta \ln(X)$$

- Concepto: A suma de términos
- “Variable” puede ser original o una variable transformada

Modelos no lineales: Ejemplos

$$Z = \alpha\, e^{\beta X}$$

$$Z = \alpha + X^{\beta}$$

$$Z = \alpha + \beta_1 X + \beta_2 X^2$$

$$Z = 1/\,(1+\beta e^{-\gamma X})$$

e: 2.718 (base del sistema de logaritmos naturales)

- Concepto: no pueden ser escritos como una suma de términos con cada término igual al parámetro por la variable

Definición técnica:

El modelo es lineal si la derivada parcial con respecto a cada parámetro involucra solo las variables que predicen

$(Z = \alpha + \beta X\ ;\ \delta Z/\delta\beta = X)$

De otra manera, es no lineal

Consideremos

$$Z = \alpha/\beta\, X + \gamma\, X2$$

$$Z = \delta X + \gamma X_1$$

Esta función es no lineal en α y β

Pero consideremos que la relación de dos constantes es otra constante

$$\alpha/\beta = \delta$$

Por lo tanto es lineal en término de δ y γ

Modelo lineal vs. no lineal

- Están involucrados muchos parámetros (ej. para ajustar el modelo a datos)
- Ajustar modelos lineales a un set de datos es bastante directo
  - Existe una solución unica para todos los conjuntos de datos
- Es relativamente difícil ajustar modelos no lineales a datos
  - El ajuste es un proceso iterativo
    - (ej. "encontrar" el parámetro que da la mínima suma de cuadrados del error)
- Las propiedades de los estimadores de los parámetros de modelos no lineales quedan menos definidas cuando se usa un set de datos pequeño
- Aún así, los procesos biológicos son de naturaleza no lineal

Modelos no lineales

Muchos modelos no lineales pueden ser transformados (usando álgebra) en modelos lineales

Ejemplo

$Z = \alpha\, e^{\beta X}$

$\ln(Z) = \ln(\alpha\, e^{\beta X}) = \ln(\alpha) + \ln(e^{\beta X})$

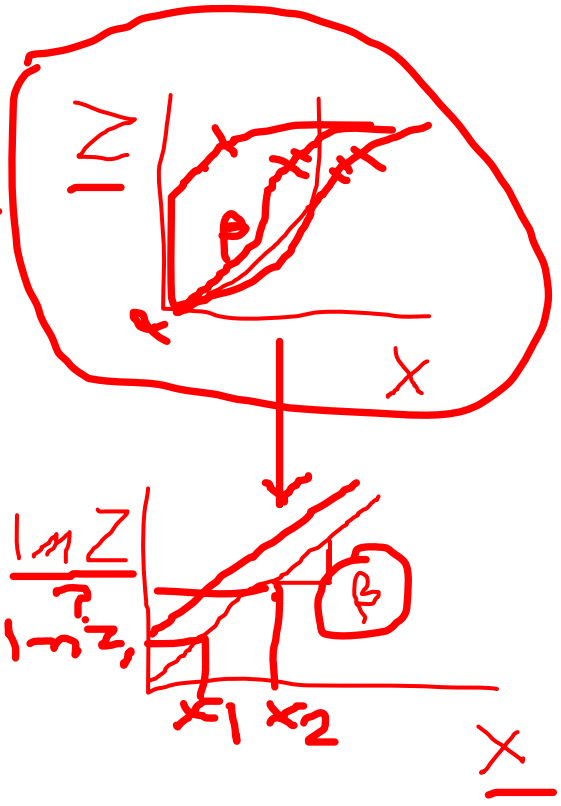
$\ln(Z) = \ln(\alpha) + \beta X$

$Z^* = \alpha^* + \beta X^*$

Nueva pendiente

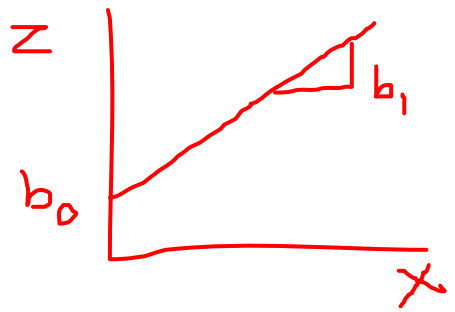
Nueva constante de intercepción

Nueva variable respuesta



Más de una manera de ver un modelo

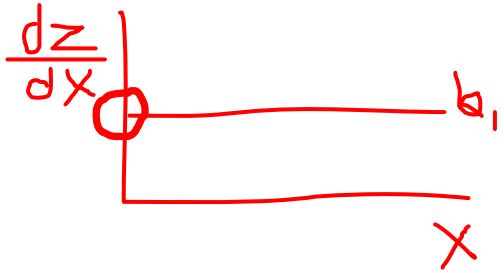
$$Z = b_0 + b_1 X$$



Si tomamos la derivada de Z respecto a: $dZ/dX = b_1$

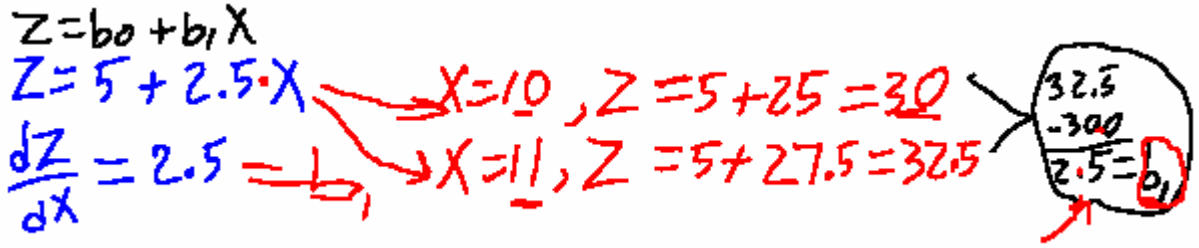
La fórmula de la derivada dice que un cambio en Z cuando cambia X es una Constante igual a b_1 (que es la pendiente de la línea Z:X)

Existen dos maneras equivalentes para describir la misma relación



Derivadas e integrales

dZ/dX es la tasa, el cambio en Z con un cambio en X. Si X es tiempo, luego se describe la tasa de cambio sobre el tiempo



Luego, la tasa es el cambio en Z cuando X cambia en una (1) unidad; con este ejemplo, es igual a la pendiente b_1

Algunas observaciones

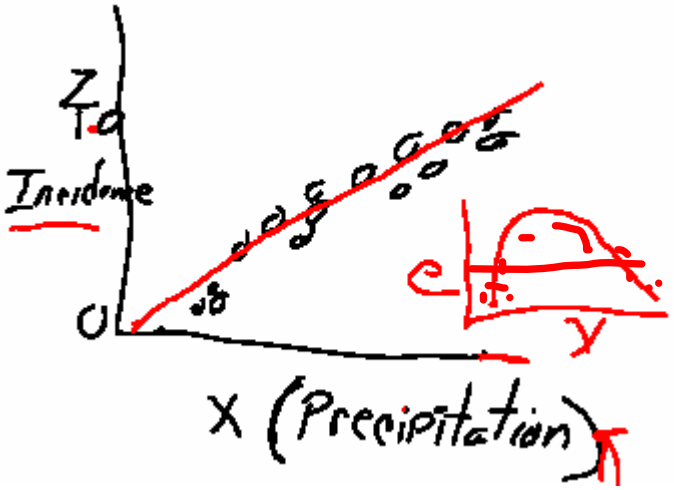
- Hay diferencia entre parámetros estimados y los parámetros reales

- Cuando hay distinciones importantes, se usará anotaciones con el “hat”

- A partir de aquí, las relaciones epidemiológicas se especificarán con modelos, normalmente comenzando con ecuaciones diferenciales (dado que expresan **dinámica**)

Ejemplo

"Phymatotrichum root rot" en algodón



Capítulo 4

Modelo empírico

Ajustar lineal simple : $Z=b_0+b_1X+e$

Resultado $Z=-0.11+0.0011X$, $R^2=0.683$

Basado en los residuales, consideramos ln(X)

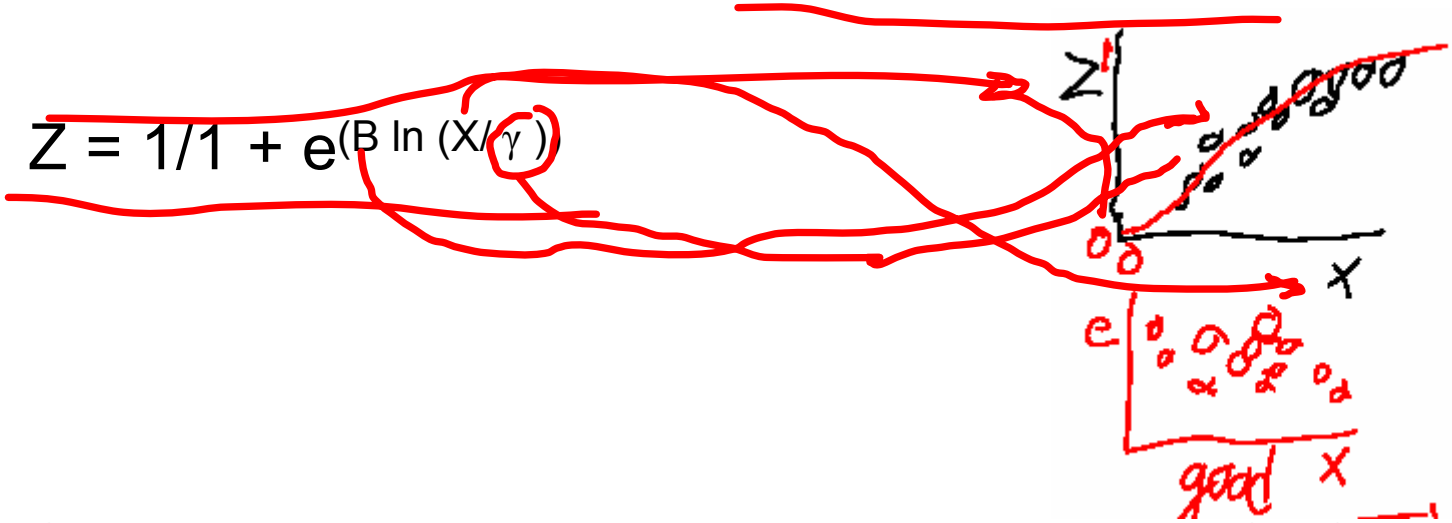
$Z = b_0 + b_1 \ln(X) + e$

$Z = -2.42 + 0.73 \ln(X)$; $R^2 = 0.744$

Todavía Z puede ser más grande que 1 o más chico que 0.01

Ahora consideremos un modelo no lineal

$Z = 1/1 + e^{(B \ln (X/\gamma))}$



Si corremos este modelo con Proc nlin de SAS empleando mínimos cuadrados

$Z = 1/1 + e^{(-0.4 \ln X/ 52.95)}$

Z en este caso no puede ser > 1 o < 0

R^2 se puede determinar de la siguiente manera:

R^2= 1-SCE/SCT =0.77

Aunque el R^2 no es mucho mejor que la transformación, esto tiene más sentido

Podemos escribir el modelo como:

$\ln (Z/(1-Z)) = -\beta \ln(X/\gamma) = -\beta (\ln(X) - \ln(\gamma))$

$= \beta \ln(\gamma) - \beta \ln(X)$

$Z^* = \alpha^* + \beta^* X^*$ ⟶ modelo lineal

Si vemos, los parámetros derivados de este modelo no son necesariamente iguales a los estimados en el modelo no lineal

$Z = -18.6 + 4.7 X$ en el no lineal, $\beta = -0.4$

$-18.6 = \beta \ln(\gamma) = -18.6$

$-4.7 \ln(\gamma) = -18.6$

$\ln(\gamma) = 3.96 \quad \gamma = e^{3.96} = 52.5$; valor cercano a 52.95 del modelo no lineal

Este ejemplo es para mostrar la modelación lineal y no lineal

Para el modelo no lineal elegido Z no puede ser >1 ó < 0

En este ejemplo, γ es el valor de X que da un Z=½=0.5; un aproximación al LD_{50}