

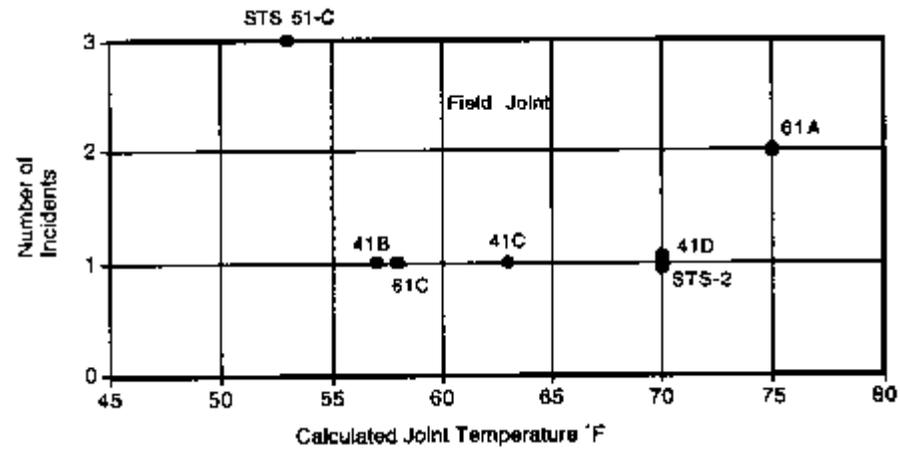
# Introducción a GLM

Modelos Estadísticos Avanzados

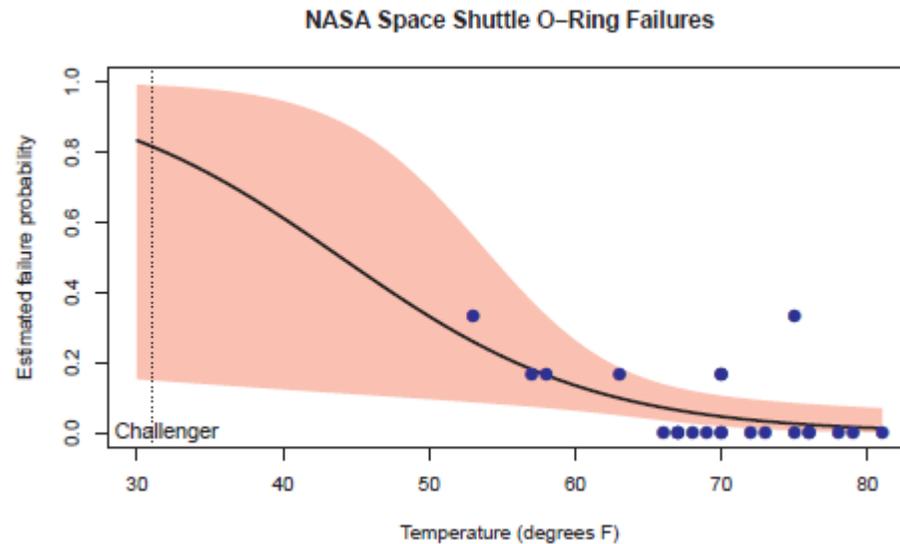
Santiago Benitez-Vieyra







Los técnicos de la NASA analizaron los datos de la relación entre la temperatura ambiental y el número de fallas en los anillos-O...



... pero excluyeron las observaciones donde no hubo fallas de los anillos-O creyendo que eran no informativas. Desafortunadamente estas observaciones eran cruciales para entender el vínculo entre la temperatura y las fallas. (tomado de <http://www.datavis.ca/gallery/>)

# ¿Por qué usar modelos lineales generalizados?

- Modelos más específicos son más eficientes y poderosos.
- Reflejan mejor la naturaleza de los datos.
- Algunos datos son *resistentes* a ser transformados (por ej. discretos, ricos en ceros).
- Las relaciones lineales a menudo fuerzan las predicciones por fuera del espacio de la respuesta (por ej. fuera del intervalo  $\{0, 1\}$  en proporciones, o predicen valores negativos en los conteos).

# ¿Por qué usar modelos lineales generales? (normales)

- Rápidos y numéricamente estables.
- Es relativamente fácil incorporar efectos aleatorios.
- Realmente muchos datos son razonablemente normales.
- Robustos para grandes sets de datos.

# Modelos...

Vamos a considerar a las teorías como generadoras de patrones de números, que en algunos casos reemplazan a los datos y que en si mismos pueden ser descritos en términos de un pequeño número de cantidades. Estas cantidades se llaman *parámetros*. (McCullagh & Nelder 1989)

$$y = \alpha + \beta x$$

En la práctica esta asociación es sólo aproximadamente lineal. Pero aún así podemos elegir valores de  $\alpha$  y  $\beta$  de alguna manera para describir la relación de  $y$  con  $x$ .

$$\hat{y} = \alpha + \beta x$$

$$y \neq \hat{y}$$

El criterio habitual para elegir  $\alpha$  y  $\beta$  es una medida de discrepancia como  $L_2$

$$L_2 = \sum (y_i - \hat{y}_i)^2$$

**Esto implica dos cosas:**

- La adición sencilla indica que las observaciones son independientes.
- La resta sencilla indica que cada desviación tiene el mismo peso independientemente del valor de  $\hat{y}_i$ , es decir que la varianza de cada observación es independiente de la media.

# Componentes de un Modelos Lineal Generalizado

# 1. Un componente aleatorio

La variable respuesta y su distribución de probabilidad: la familia exponencial natural

Por ej.

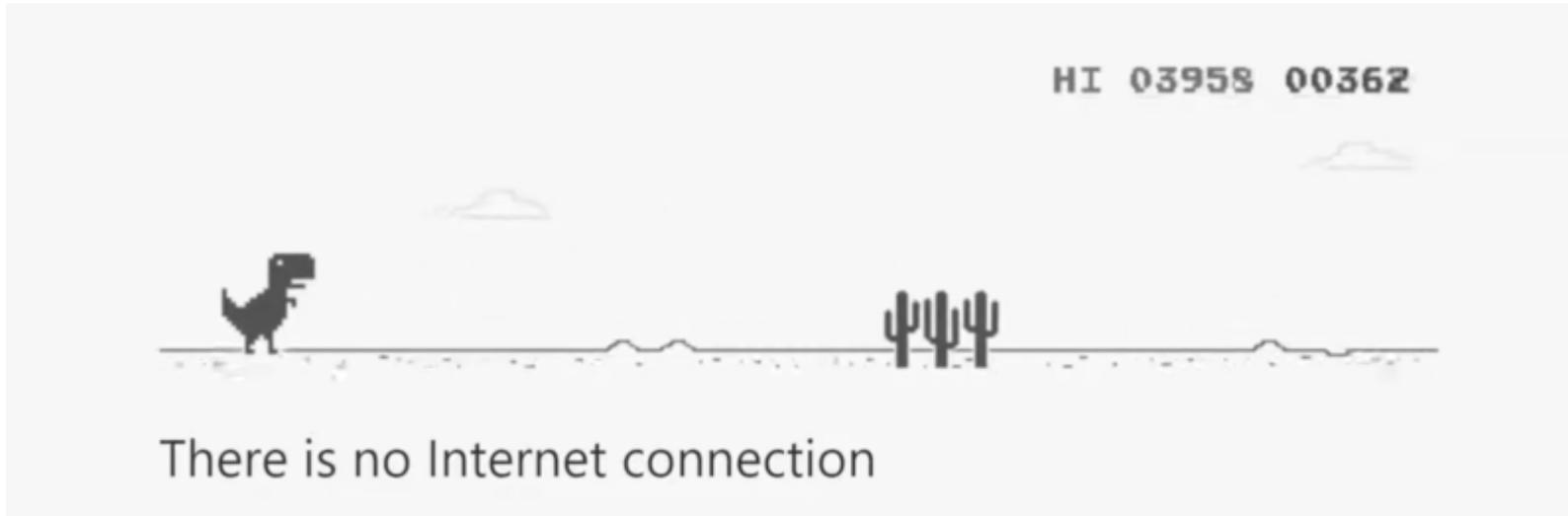
$$y_i \sim N(\mu_i, \sigma^2)$$

$$E(y_i) = \mu_i$$

$$\text{var}(y_i) = \sigma^2$$

Pero hay muchas otras posibilidades además de la normal

<https://www.youtube.com/watch?v=TktiBhPdZYE>



## La familia exponencial natural

- Abarca una serie de distribuciones como Normal, Poisson, Binomial, Gamma, Binomial Negativa, Multinomial y varias otras.
- La elección de la distribución es **a priori** y depende de **la forma en que se generaron los datos**.
- Cada distribución se caracteriza por la relación entre su **parámetro de posición** y su **parámetro de dispersión**.

# Forma en que se generan los datos

Distribución	Origen
Normal	Caracteres métricos y continuos.
Poisson	Valores enteros entre 0 y un número indeterminado.
Binomial	Conteos distribuidos entre 0 y un número finito conocido. Proporciones.
Bernoulli	Caso especial del anterior. Toman sólo valores de 0 y 1.
Binomial Negativa	Número de experimentos de Bernoulli hasta la consecución del k-ésimo éxito.
Gamma	Similar a la normal, pero no posee valores negativos. Generalmente asociada al tiempo hasta el cierto evento.

---

## Relacion entre el parámetro de posición y el de dispersión.

Distribución	Posición	Dispersión
Normal	$E(Y) = \mu$	$var(Y) = \sigma^2$
Poisson	$E(Y) = \mu$	$var(Y) = \mu$
Binomial	$E(Y) = N\pi$	$var(Y) = N\pi(1 - \pi)$
Bernoulli	$E(Y) = \pi$	$var(Y) = \pi(1 - \pi)$
Binomial Negativa	$E(Y) = \mu$	$var(Y) = \mu + \mu^2/k$
Gamma	$E(Y) = k/\lambda$	$var(Y) = k/\lambda^2$

---

# NOTA: Algunas distribuciones que no pertenecen a la familia exponencial natural

**Distribución**

**Proceso**

**Log- Normal**

“Ruido multiplicativo” (por oposición al modelo infinitesimal de Fisher, que origina distribuciones normales por pequeños efectos aditivos).

**Beta**

Variable continua acotada a  $(0, 1)$ . Típicamente proporciones de otras variables continuas.

---

## 2. Un componente sistemático

las variables predictoras en el modelo. Pueden ser continuas, categóricas, funciones polinomiales, interacciones, etc.

Función  $\eta$  formada por una combinación lineal de las variables predictoras  $X$ .

$$\eta(X_{i1}, \dots, X_{ik}) = \beta_0 + \beta_1 X_{i1} + \dots + \beta_k X_{ik}$$

### 3. Una función de enlace.

$$g(\mu_i) = \eta(X_{i1}, \dots, X_{ik})$$

- Propias de cada distribución (algunas permiten más de una).
- El enlace debe ser reversible.
- El enlace acota la respuesta a la distribución de la variable respuesta.
- Para la distribución normal, el enlace es la identidad.

$$E(Y_i) = \mu_i = \eta(X_{i1}, \dots, X_{ik}) = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki}$$

<b>Distribución</b>	<b>enlace (canónico en negrita)</b>
<b>Normal</b>	<b>identidad, log, inverso.</b>
<b>Poisson</b>	<b>log, identidad, raíz cuadrada.</b>
<b>Binomial</b>	<b>logit, probit, complemento log-log.</b>
<b>Binomial Negativa</b>	<b>log, identidad, raíz cuadrada.</b>
<b>Gamma</b>	<b>inverso, identidad, log.</b>

---

# Estimación por máxima verosimilitud.

- El interés es obtener un set de parámetros que maximice la probabilidad de observar los datos.

$$P(\text{datos}|\text{parametros})$$

$$P(y_1 = \hat{y}_1, y_2 = \hat{y}_2, \dots, y_n = \hat{y}_n)$$

$$L = \prod f(x_i|\theta)$$

- Para esto rastreo en el espacio de parámetros, cuáles de son los más verosímiles para mi set de datos.

$$P(\text{datos}|\text{parametros})$$

$$L(\text{parametros}|\text{datos})$$

¿Qué grado de apoyo tienen los parámetros que obtuve?

- El grado de apoyo es siempre comparativo.
- El grado en que los datos apoyan el valor de un set de parámetros particular, respecto al apoyo que hacia otro set de parámetros es igual al **cociente de sus verosimilitudes**.
- Para ver la significancia de un parámetro basta con comparar la verosimilitud de un modelo con ese parámetro incluido y la de un modelo donde falte.

$$\frac{L(H_1)}{L(H_0)}$$

## ¿Qué modelos comparar?

- **Modelo nulo:** No ofrece ninguna explicación para los datos, se expresa como  $y \sim 1$  y, gráficamente en una regresión, como una línea horizontal. Es el modelo con la mínima verosimilitud.
- **Modelo saturado:** Tiene un parámetro para cada dato, por lo que explica total y absolutamente todas las observaciones. Tiene la máxima verosimilitud posible pero no sirve para nada, salvo algunas excepciones. Gráficamente, en una regresión, es una línea que conecta todos los puntos entre sí.
- **Modelo corriente:** es el que intentamos buscar, aquel que explique la mayor parte de la variación de los datos, pero que use el menor número de parámetros posibles.

¿Cómo conozco el ajuste del modelo?

$$Dev = -2\log(L(\text{Modelo}_{\text{corriente}})/L(\text{Modelo}_{\text{saturado}}))$$

$$Dev \sim \chi^2$$

- Modelo nulo: tiene la máxima devianza.
- Modelo saturado: Tiene devianza = 0
- Modelo corriente: es el que deja una devianza residual lo más pequeña posible.

# CONCEPTOS CLAVE

- Elegimos *a priori* una distribución para la variable respuesta, basada en su naturaleza y su relación media-varianza.
- Elijo una función de enlace que proyecta la predicción en el espacio de la variable respuesta.
- Construyo un modelo con variables predictoras.
- Obtengo la *Devianza* del modelo que me interesa y la comparo con la del modelo saturado.

**END**