Aprendizaje Automático

Modelos de regresión lineal

Prof. Rodrigo Díaz
Lic. Manuel Szewc
Lic. Luis Agustín Nieto

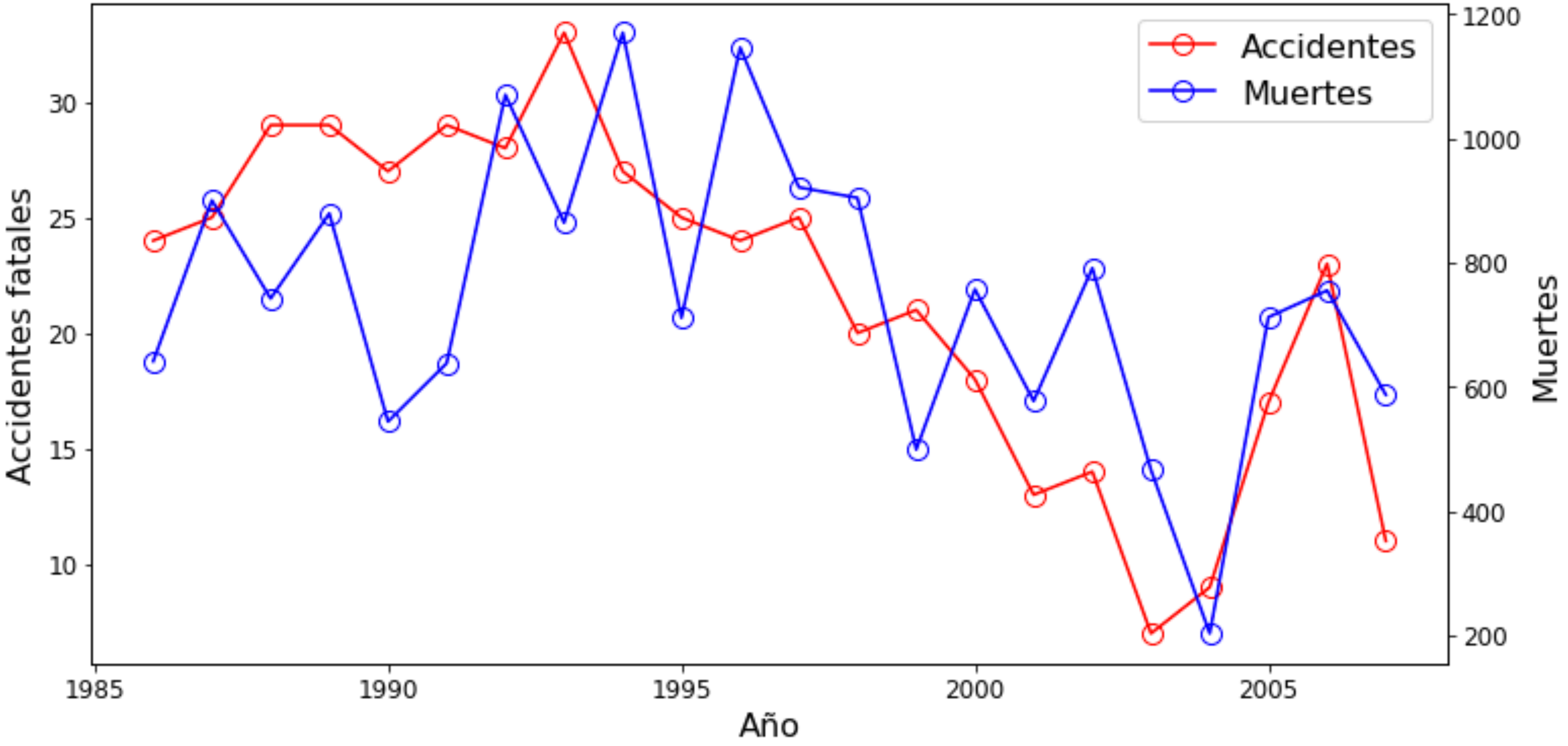
UNSAM - 25 de marzo de 2021

Agenda para hoy

- **Introducción a los modelos lineales para la regresión.**
  - ¿Qué son los modelos lineales?
  - ¿Qué modelan?
  - ¿Cómo encuentro los valores de los parámetros?
  - …
- **Diagnóstico de modelos (tal vez el martes)**
  - Residuos.
  - Palanca.
  - …

Data first

Archivo en datasets/airline_fatalities.csv



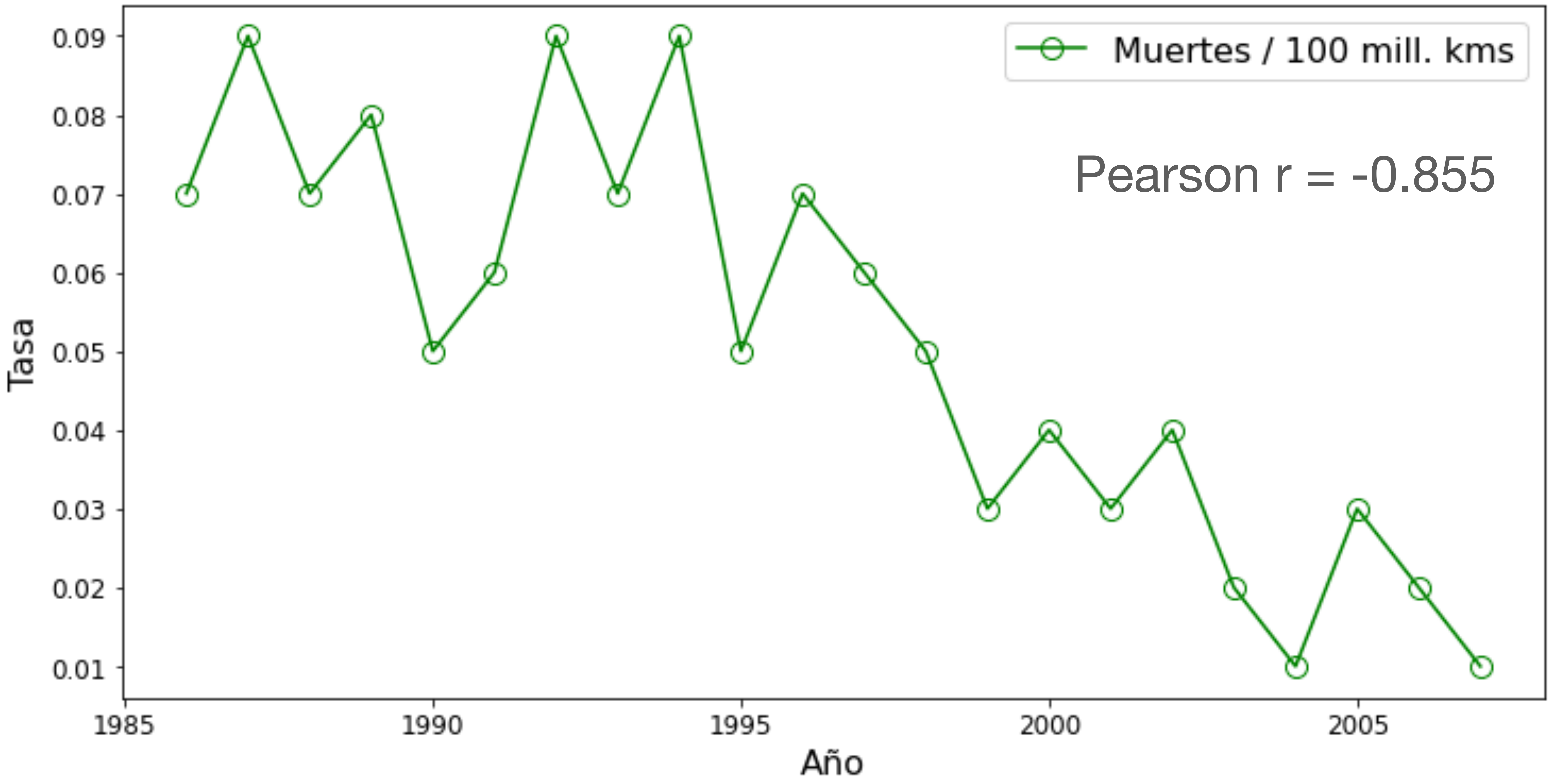
Data first

Archivo en datasets/airline_fatalities.csv

- **Algunas preguntas**
  - ¿Se va volviendo más seguro viajar?
  - ¿Cuántos accidentes fatales predecimos para 2008 a partir de estos datos?

Data first

Archivo en datasets/airline_fatalities.csv



Modelo lineal simple

Función de error

$$y(x_i, \boldsymbol{\omega}) = y_i = \omega_0 + \omega_1 x_i \qquad \boldsymbol{\omega}^T = (\omega_0, \omega_1)$$

¿Cómo encontramos los parámetros?

Desde un punto de vista frecuentista, suponemos que existen los verdaderos valores de los parámetros, y lo que buscamos, entonces, es un buen estimador. **¿Qué criterio vamos a usar?**

Buscamos los estimadores que minimizan esta función de error

$$\boxed{MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - t_i)^2}$$

Error cuadrático medio

Modelo lineal simple

Estimación de los parámetros

icas.unsam.edu.ar

Buscamos los estimadores que minimizan esta función de error

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - t_i)^2$$

derivando (con respecto a los parámetros) e igualando a cero

Error cuadrático medio

Normal equations

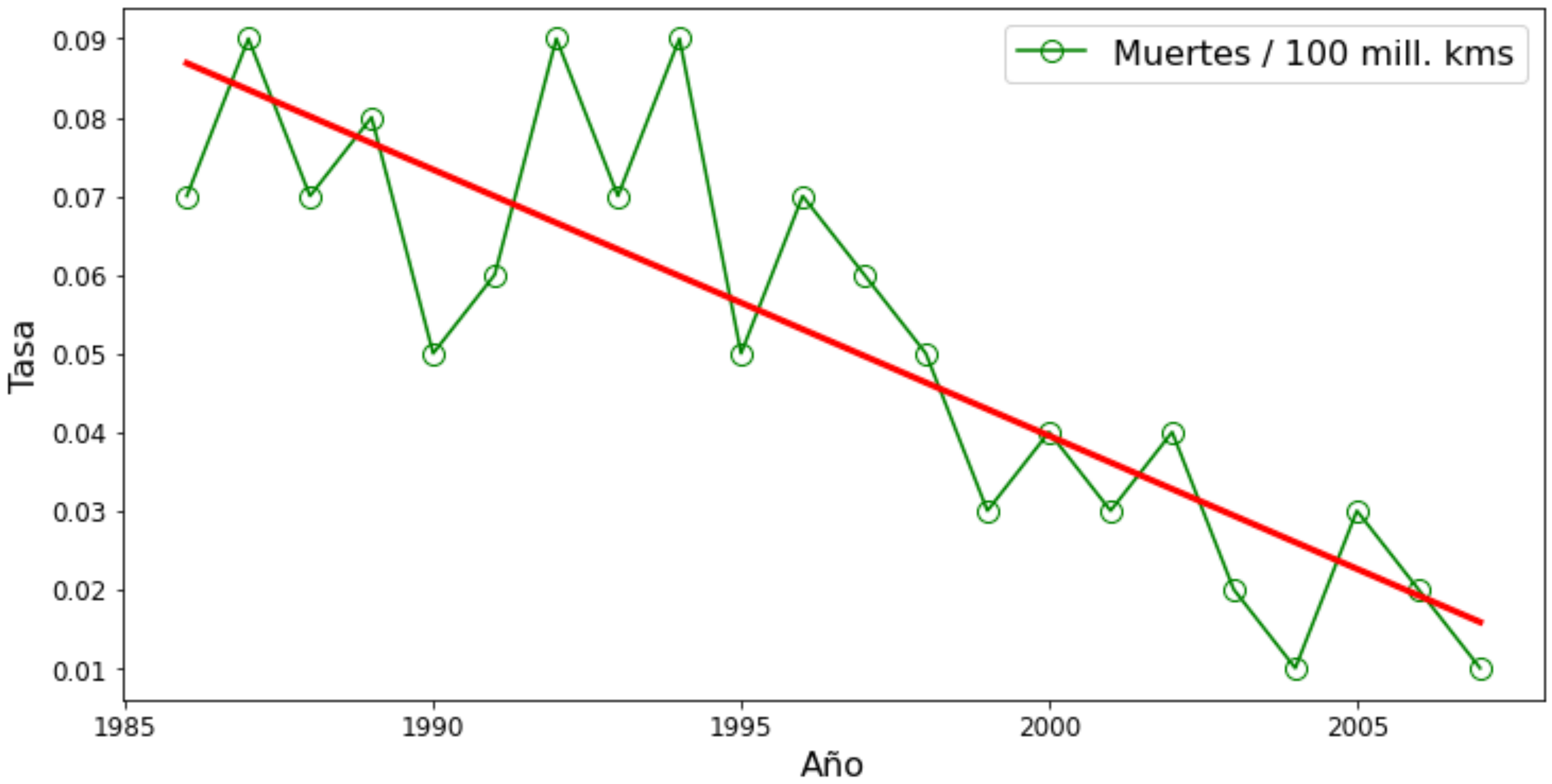
$$\begin{aligned} \textstyle\sum_{i=1}^{N}\left[t_i - (\hat{\omega_0} + \hat{\omega_1}\cdot x_i)\right] &= 0 \\ \textstyle\sum_{i=1}^{N}\left[t_i - (\hat{\omega_0} + \hat{\omega_1}\cdot x_i)\right] x_i &= 0 \end{aligned}$$

$$\begin{aligned} \hat{\omega_1} &= \textstyle\sum_{i=1}^{N}(x_i - \bar{X})(t_i - \bar{T})\left[\sum_{i=1}^{N}(x_i - \bar{X})^2\right]^{-1} \\ \hat{\omega_0} &= \bar{T} - \hat{\omega_1}\bar{X} \end{aligned}$$

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{N} x_i \ .$$

Modelo lineal simple

Estimación de los parámetros



$$\hat{\omega_1} = \sum_{i=1}^{N}(x_i - \bar{X})(t_i - \bar{T}) \left[\sum_{i=1}^{N}(x_i - \bar{X})^2\right]^{-1}$$
$$\hat{\omega_0} = \bar{T} - \hat{\omega_1}\bar{X}$$

$$\hat{\omega_1} = -0.003382$$
$$\hat{\omega_0} = 6.804$$

Modelo lineal simple

Los errores y los residuos

ICAS
icas.unsam.edu.ar

$$t_i = y_i + \epsilon_i \qquad i = \{1, \ldots, N\}$$

Para que la MSE sea válida, suponemos que

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

O sea, suponemos que los errores:

1. están distribuidos como una normal centrada en cero.
2. tienen todos la misma varianza sigma^2 Homoscedasticidad
3. son independientes

Hipótesis del modelo lineal

¿Cómo estimamos el parámetro que queda, sigma^2?

Modelo lineal simple

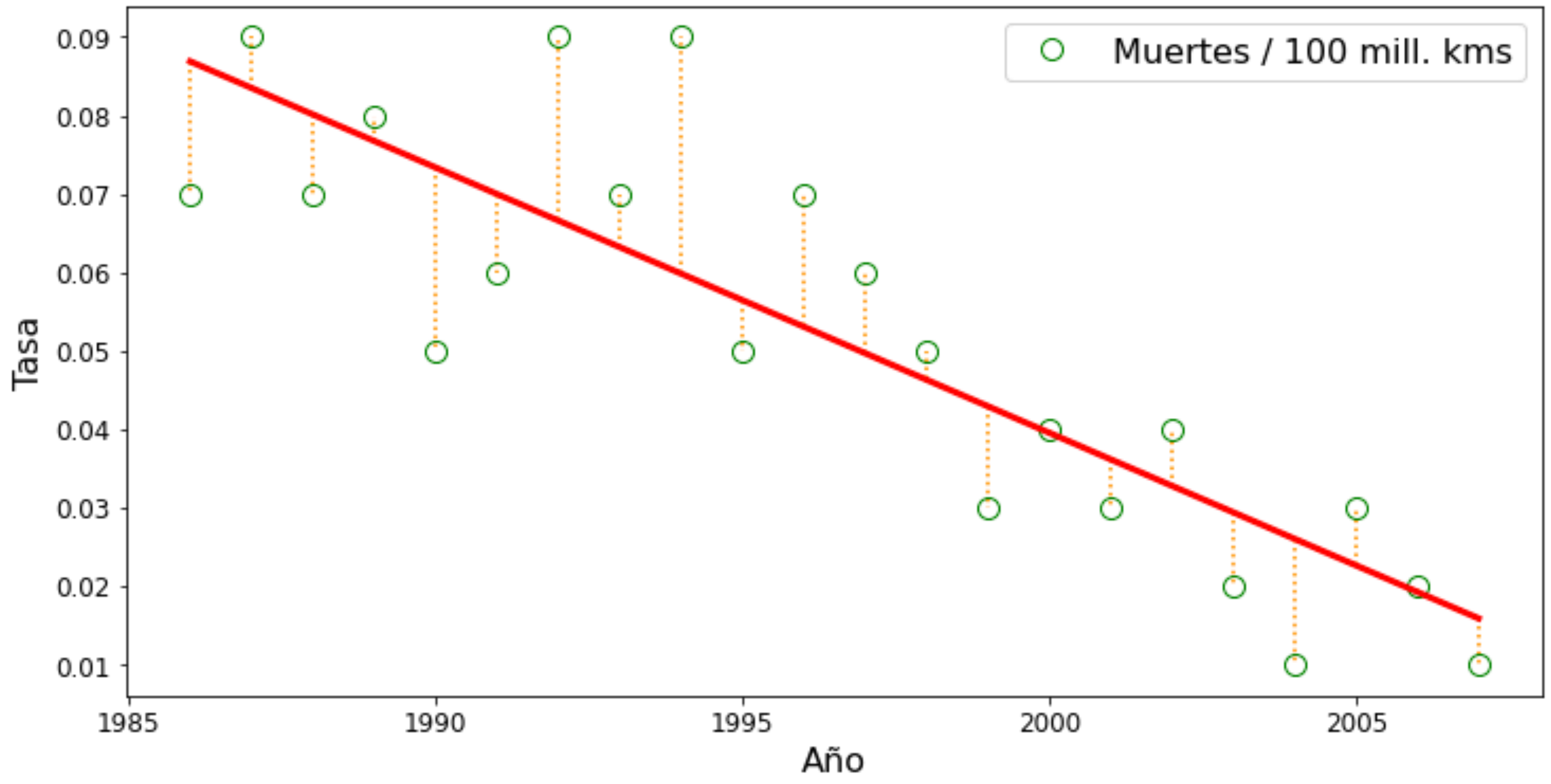
Los errores y los residuos

$$r_i = t_i - \hat{y}_i = t_i - (\hat{\omega}_0 + \hat{\omega}_1 X_i)$$

$$r_i \neq \epsilon_i$$

por ejemplo

$$\sum_{i=1}^{N} r_i = 0$$



$$\boxed{\widehat{\sigma^2} = \frac{1}{N-2}\sum_{i=1}^{N} r_i^2}$$

$$\widehat{\sigma^2} = 0.0001865$$

$$\text{MSE} = 0.0001695$$

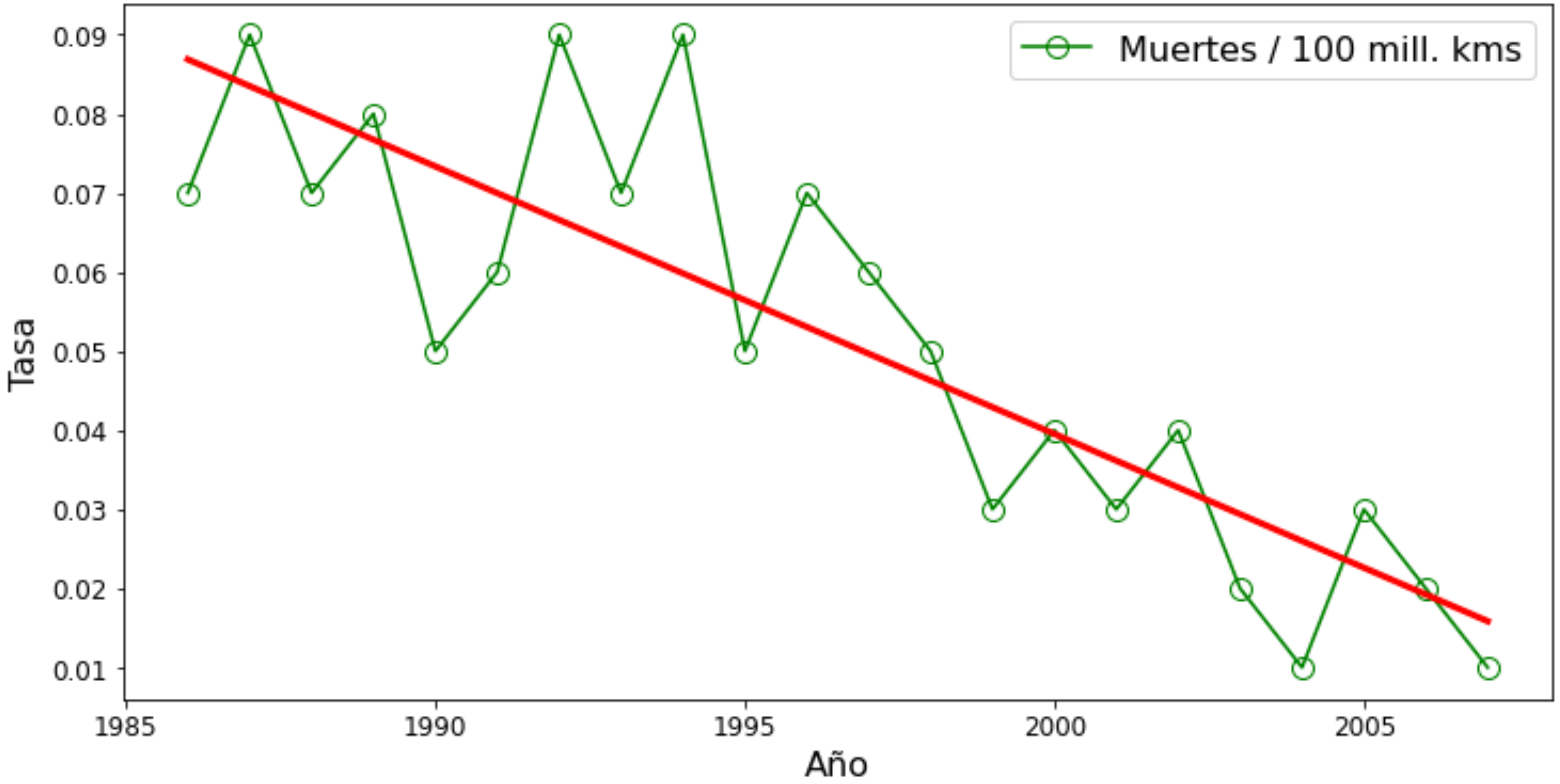
Data first

Archivo en datasets/airline_fatalities.csv

- **Algunas preguntas**
  - ¿Se va volviendo más seguro viajar?
  - ¿Cuántos accidentes fatales predecimos para 2008 a partir de estos datos?

Modelo lineal simple

Estimación de los parámetros



$$
\begin{aligned}
\hat{\omega_1} &= \sum_{i=1}^{N}(x_i - \bar{X})(t_i - \bar{T})\left[\sum_{i=1}^{N}(x_i - \bar{X})^2\right]^{-1} \\
\hat{\omega_0} &= \bar{T} - \hat{\omega_1}\bar{X}
\end{aligned}
$$

$$
\begin{aligned}
\hat{\omega_1} &= -0.003382 \\
\hat{\omega_0} &= 6.804
\end{aligned}
$$

Modelo lineal simple

Una noción del ancho de la distribución de los estimadores

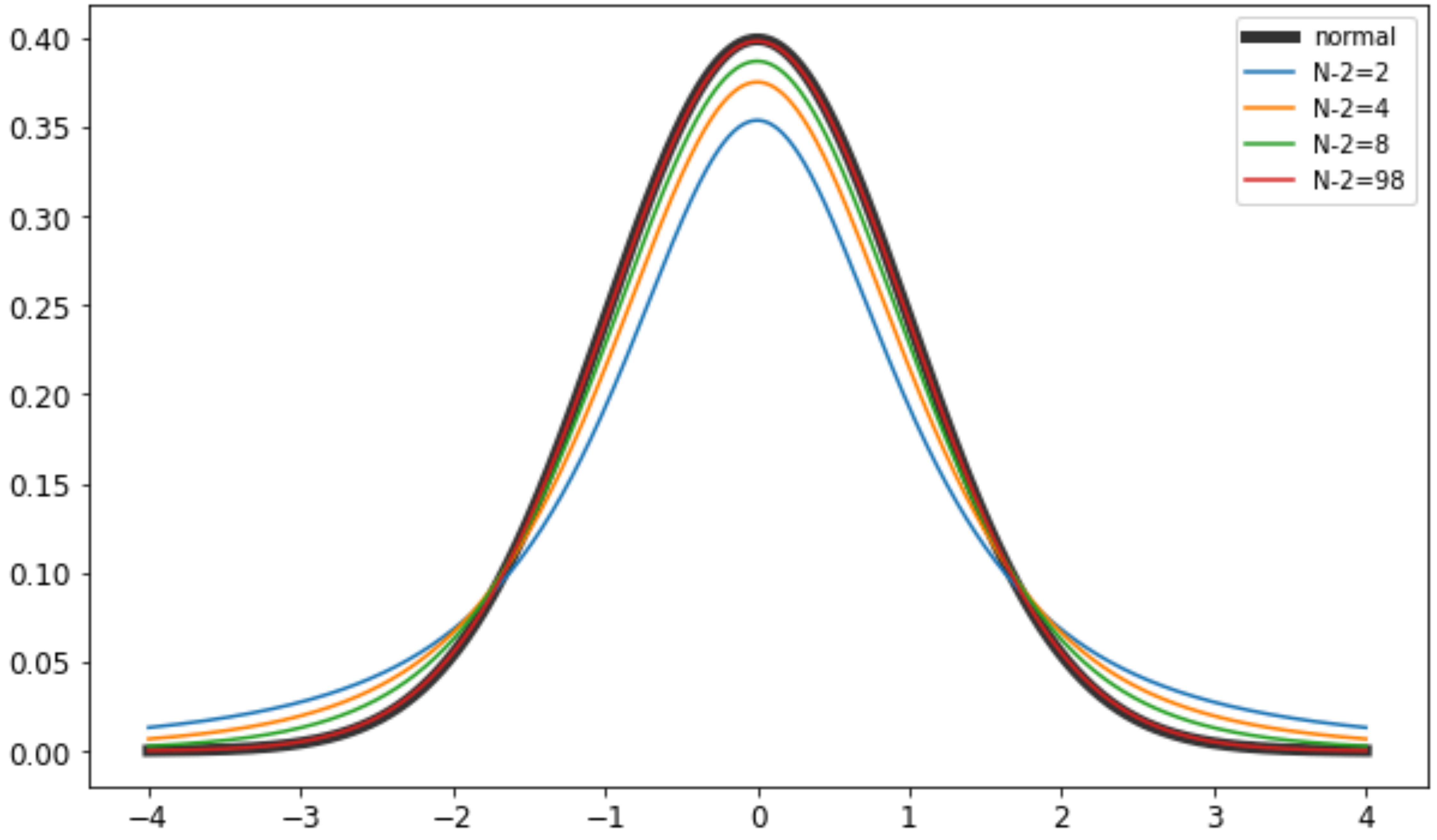
$$\mathbb{E}(\widehat{\omega_1}) = \omega_1$$

$$\mathrm{var}(\widehat{\omega_1}) = \frac{\sigma^2}{\sum_{i=1}^{N}(x_i - \bar{X})^2}$$

y se puede probar que

$$\frac{\widehat{\omega_1} - \omega_1}{\sqrt{\widehat{\mathrm{var}}(\hat{\omega_1})}} \sim \text{t-Student}_{N-2}$$

```
import scipy.stats as st
st.t(df=n-2)
```



Modelo lineal simple

Probamos la hipótesis de no cambio

ICAS
icas.unsam.edu.ar

Suponemos

$$\omega_1 = 0$$

Y definimos

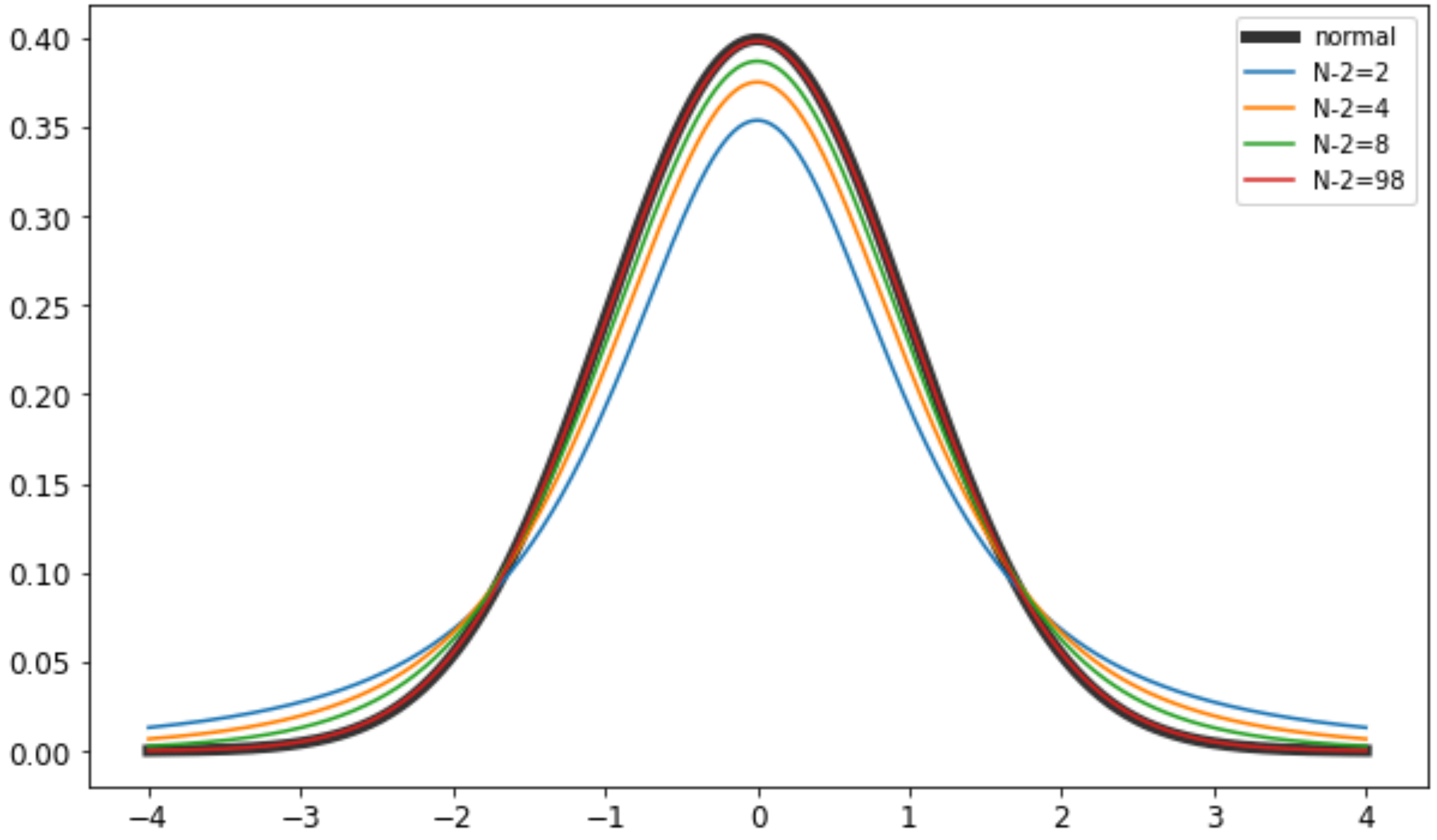
$$q = \frac{\widehat{\omega_1}}{\sqrt{\widehat{\mathrm{var}}(\hat{\omega_1})}} \sim \text{t-Student}_{N-2}$$

(si la suposición es correcta)

$$q = -7.37$$

$$p(q < -7.37) = 2.01e - 7$$

¿podemos rechazar la hipótesis de que no hay variación?
Discutamos



Data first

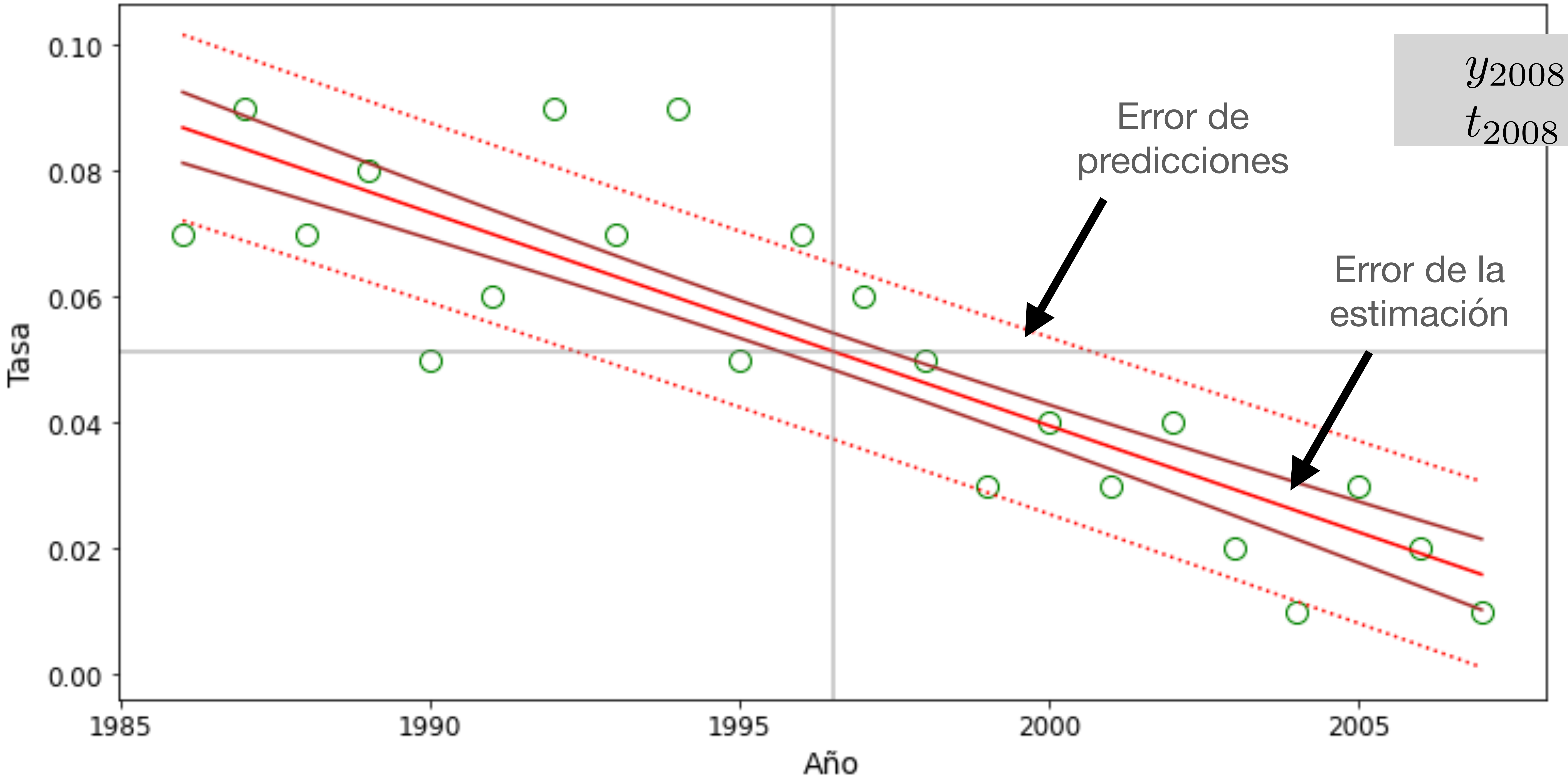
Archivo en datasets/airline_fatalities.csv

icas.unsam.edu.ar

- **Algunas preguntas**
  - ¿Se va volviendo más seguro viajar?
  - ¿Cuántos accidentes fatales predecimos para 2008 a partir de estos datos?

Error en las predicciones

Error irreducible



$$y_{2008} = 0.012 \pm 0.015$$
$$t_{2008} = 0.01$$

Extensión del modelo

Modelo lineal múltiple

Modelo lineal simple

$$y(x, w_0, w_1) = w_0 + w_1 x \quad .$$

Modelo lineal múltiple

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \cdots + w_D x_D \quad .$$

Más en general:

$$y_i(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{i=1}^{D} w_j \phi_j(\boldsymbol{x_i}) = \sum_{j=0}^{D} w_j \phi_j(\boldsymbol{x_i}) = \boldsymbol{w}^T \boldsymbol{\phi}_i$$

Modelo lineal múltiple

Notación matricial

$$y_i(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{i=1}^{D} w_j \phi_j(\boldsymbol{x_i}) = \sum_{j=0}^{D} w_j \phi_j(\boldsymbol{x_i}) = \boldsymbol{w}^T \boldsymbol{\phi}_i$$

$$y_i(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{\phi}_i \qquad i = \{1, \ldots, N\}$$

Matriz de diseño

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix} = \boldsymbol{\Phi} \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{pmatrix}$$

$$\boldsymbol{\Phi} = \begin{pmatrix} \phi_0(\mathbf{x}_1) & \phi_1(\mathbf{x}_1) & \cdots & \phi_{M-1}(\mathbf{x}_1) \\ \phi_0(\mathbf{x}_2) & \phi_1(\mathbf{x}_2) & \cdots & \phi_{M-1}(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\mathbf{x}_N) & \phi_1(\mathbf{x}_N) & \cdots & \phi_{M-1}(\mathbf{x}_N) \end{pmatrix}$$

$(N \, x \, 1) \quad = \quad (N \, x \, D) \;\; (D \, x \, 1)$

Modelo lineal múltiple

Ecuaciones normales

La resolución de las ecuaciones normales es ahora una ecuación matricula

$$\boxed{\mathbf{w}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{t}}$$

.... pero no se asusten, ahora lo vemos en la práctica.