Sesión 3: Comparación de Modelos

Curso: Introducción práctica a la inferencia estadística
Escuela de Verano CIMAT 2025

Marco Antonio Aquino López
CIMAT

9 de julio de 2025

Motivación: ¿Qué modelo describe mejor los datos?

En la sesión anterior aprendimos a construir intervalos de confianza (clásico) e intervalos de credibilidad (bayesiano) para parámetros desconocidos.

Estas herramientas nos permiten cuantificar la incertidumbre de una estimación puntual, pero su utilidad depende de que el modelo elegido sea adecuado para los datos.

¿Qué pasa si el modelo está mal especificado?

Un mal modelo: Inferencia con Normal para datos de conteo

Supongamos que observamos el número de veces que ocurre un evento en un periodo fijo:

- Número de bacterias en una muestra.
- Número de errores en una página.

Este tipo de datos es discreto y no negativo. Suele modelarse con una distribución **Poisson**.

Sin embargo, alguien decide aplicar una **Normal** para hacer inferencia sobre la media.
¿Qué consecuencias puede tener esto?

Inferencia con Normal para datos Poisson

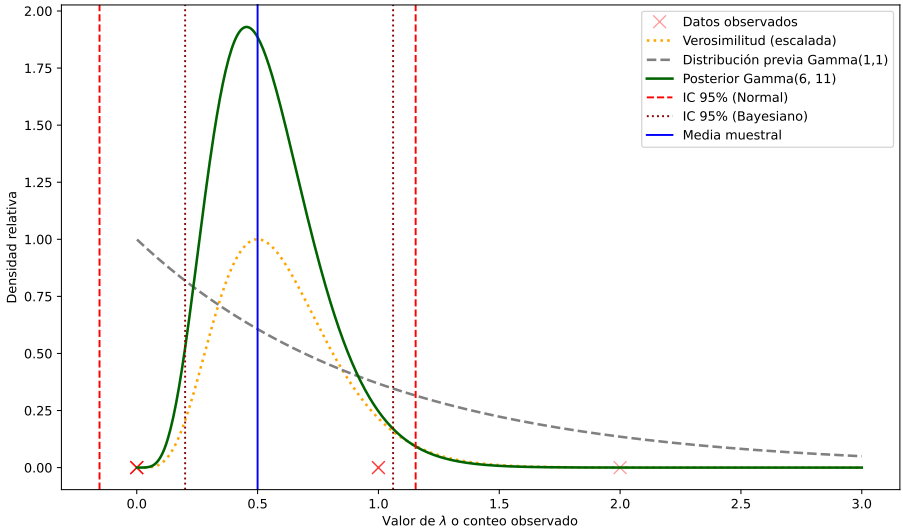
Simulamos datos desde una distribución **Poisson**, pero calculamos un intervalo de confianza como si los datos fueran normales:

- Estimador: media muestral $\bar{x}$.
- Intervalo: $\bar{x} \pm 2 \cdot \frac{s}{\sqrt{n}}$.

Problema: este intervalo puede incluir valores negativos, lo cual no tiene sentido para una tasa de ocurrencia.

Esto ilustra cómo un modelo mal especificado puede llevar a inferencias absurdas.

Visualización: Intervalo inadecuado



Solución: Usar un modelo apropiado (Poisson)

Para datos de conteo, el modelo correcto es la **distribución Poisson** con parámetro $\lambda > 0$.

- λ: tasa media de ocurrencia.
- Estimador de máxima verosimilitud: la media muestral.
- Intervalos se pueden construir con métodos exactos o aproximaciones.

También podemos hacer **inferencia bayesiana**, usando una previa adecuada (por ejemplo, Gamma).

Inferencia Bayesiana con prior Gamma(1, 1)

Si usamos una distribución previa Gamma$(1, 1)$ para λ, la posterior es:

$$\lambda \mid \text{datos} \sim \text{Gamma}(\alpha_0 + \sum x_i,\ \beta_0 + n)$$

Este intervalo siempre es positivo y coherente con el contexto del problema.

Reflexión: el modelo importa

Este ejemplo deja una lección fundamental:

- Un modelo mal especificado puede llevar a conclusiones absurdas.
- La validez de la inferencia depende de que el modelo refleje adecuadamente el fenómeno observado.
- Elegir un modelo adecuado es parte esencial del análisis estadístico.

Modelar lo incierto requiere juicio, contexto y sentido común, no solo fórmulas.

¿Y si hay una variable explicativa?

Supongamos ahora que cada observación incluye:

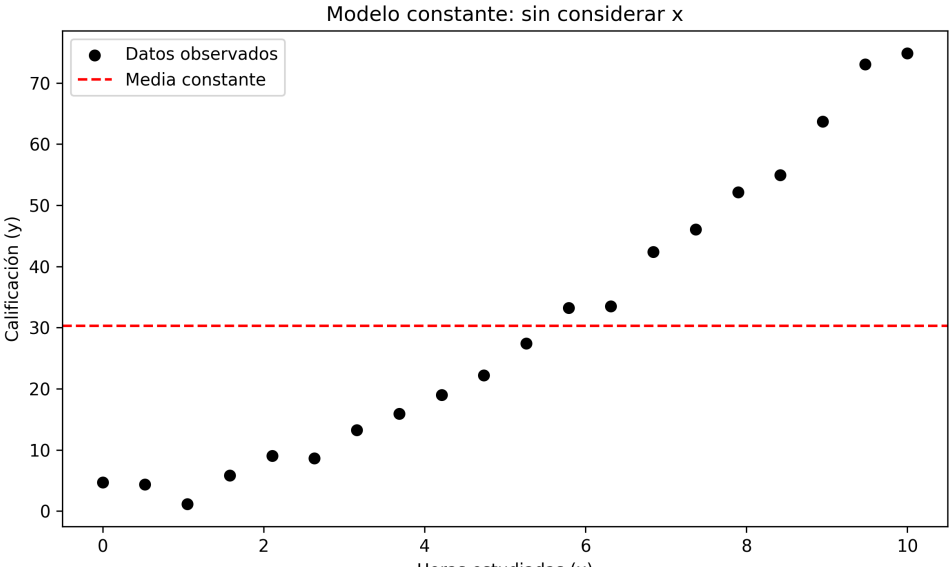
- x: número de horas que una persona estudió.
- y: calificación obtenida en un examen.

Primero ignoramos x y usamos un **modelo constante**:

- Se asume que todas las calificaciones provienen de una misma distribución normal.
- Se estima únicamente la media muestral de y.

¿Qué limitaciones tiene este enfoque?

Visualización: modelo constante vs datos



Hacia un mejor modelo: regresión lineal

Los datos muestran una tendencia: a mayor número de horas estudiadas, mayor calificación.

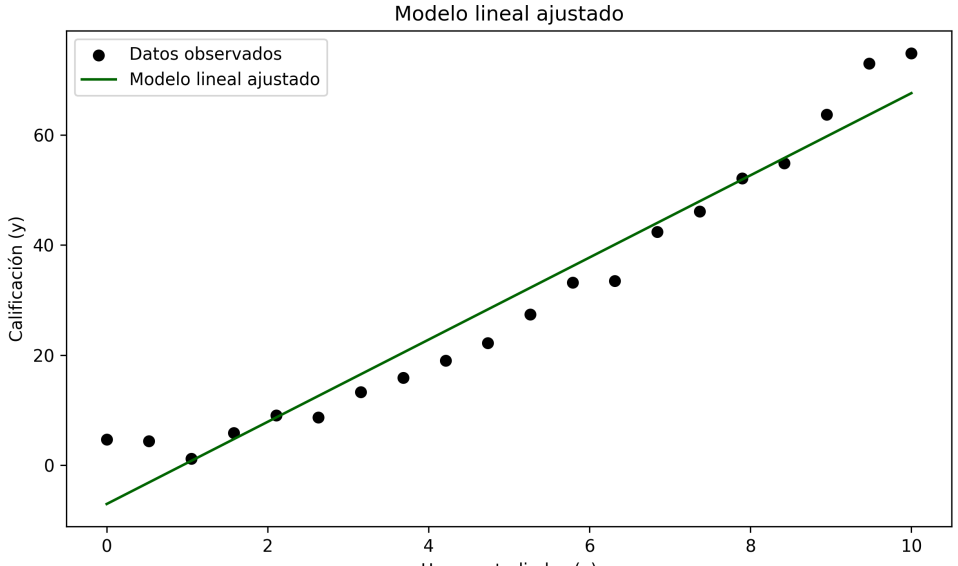
Proponemos un modelo lineal:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

- β_0: intercepto (valor esperado de y cuando $x = 0$).
- β_1: pendiente (cambio esperado en y por cada unidad en x).
- ε: error aleatorio (variación no explicada por el modelo).

Este modelo incorpora estructura y nos permite hacer predicciones.

Visualización: modelo lineal ajustado



Exploración de un modelo cuadrático

No todas las relaciones son lineales. Por ejemplo:

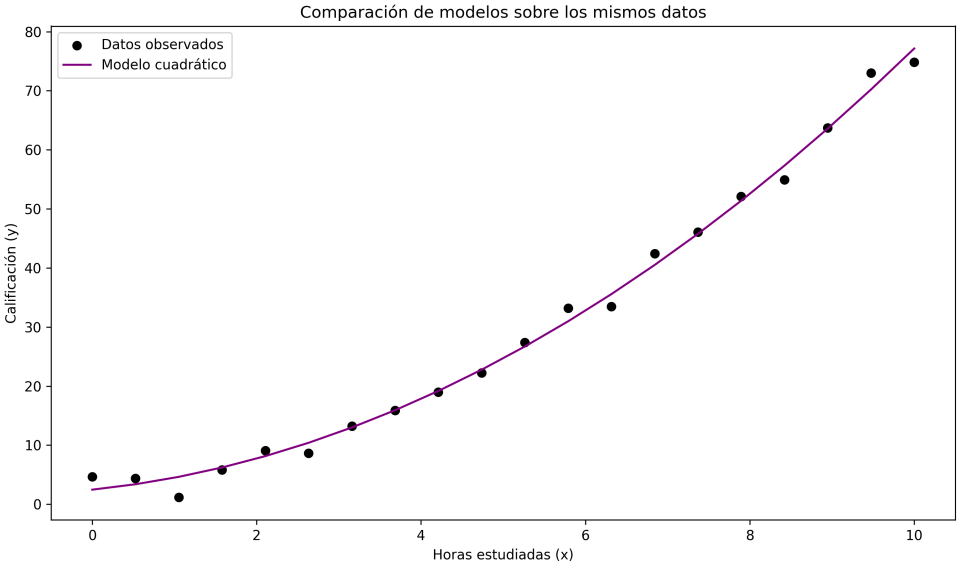
Estudiar más puede ayudar... hasta cierto punto. Después, el rendimiento podría disminuir.

Proponemos un modelo cuadrático:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

- Este modelo permite capturar formas curvadas.
- Más flexible que el modelo lineal.
- ¿Vale la pena esa mayor complejidad?

Visualización: modelo cuadrático ajustado



Evaluación cuantitativa de modelos

Para comparar modelos de forma más rigurosa, usamos:

- **Coeficiente de determinación** R^2: mide qué tanto del comportamiento de y es explicado por el modelo.
- **Criterio de información de Akaike** (AIC): penaliza modelos más complejos, buscando un equilibrio entre ajuste y simplicidad.

Estas herramientas permiten tomar decisiones más informadas al seleccionar modelos.

Coeficiente de determinación R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- Mide la proporción de la variabilidad total en y que es explicada por el modelo.
- Valores cercanos a 1 indican buen ajuste; valores cercanos a 0 indican mal ajuste.
- No penaliza modelos complejos: siempre sube o se mantiene al agregar variables.

Interpretación: "¿Qué tanto del comportamiento de la variable respuesta se explica con este modelo?"

Criterio de información de Akaike (AIC)

$$\text{AIC} = n \cdot \log\left(\frac{\text{RSS}}{n}\right) + 2k$$

- n: número de observaciones.
- RSS: suma de los cuadrados de los residuos.
- k: número de parámetros del modelo.

Objetivo: Balancear ajuste (RSS) y complejidad (número de parámetros).

Interpretación: Modelos con AIC menor son preferibles, pero diferencias pequeñas pueden no ser significativas.

Comparación de modelos: R^2 y AIC

Aplicamos R^2 y AIC a los tres modelos ajustados:

Modelo	R^2	**AIC**
Constante	0.000	149.54
Lineal	0.877	117.18
Cuadrático	0.877	117.18

Conclusiones:

- El modelo constante no explica variabilidad alguna.
- El modelo lineal mejora sustancialmente el ajuste.
- El modelo cuadrático no aporta mejora adicional significativa.

Conclusiones de la sesión

- Un modelo mal especificado puede producir inferencias absurdas (ej. tasas negativas).
- Elegir un modelo adecuado es esencial antes de hacer inferencia.
- Vimos tres modelos: constante, lineal y cuadrático.
- Usamos criterios para comparar modelos:
  - Visualización
  - Coeficiente de determinación R^2
  - Criterio de información de Akaike (AIC)
- El mejor modelo es aquel que logra buen ajuste con simplicidad.

Reflexión final y preguntas abiertas

Hoy aprendimos:

- La inferencia depende críticamente de un modelo bien elegido.
- Comparar modelos es una tarea esencial y requiere juicio.
- Herramientas como R^2, AIC y la visualización ayudan a decidir.

Preguntas para seguir pensando:

- ¿Cómo elegir un modelo cuando no conocemos la verdad?
- ¿Cuándo vale la pena usar un modelo más complejo?
- ¿Qué otras herramientas existen para comparar modelos?

¡Gracias por tu atención!