Modelos Lineales Generalizados (GLM)

Víctor Aceña - Isaac Martín

DSLAB

2025-09-11

Modelos de Regresión Generalizada

Punto de Partida: La Regresión Lineal

- Es una herramienta potente para modelar una variable dependiente **continua**.
- Sin embargo, sus supuestos (normalidad, homocedasticidad) no siempre se cumplen.

El Desafío: Datos No Normales

- ¿Cómo modelamos una variable de respuesta **binaria** (ej. enfermo/sano)?
- ¿O datos de **conteo** (ej. nº de accidentes en una intersección)?

La Solución: Modelos Lineales Generalizados (GLM)

- Son una **extensión** de la regresión lineal que permite modelar respuestas con distribuciones como la Binomial o la de Poisson, utilizando **funciones de enlace** para mayor flexibilidad.

Definición de GLM

Los **Modelos Lineales Generalizados (GLM)** son una extensión de los modelos de regresión lineal que permiten manejar una mayor variedad de tipos de datos y relaciones entre variables.

Mientras que la regresión lineal clásica asume que la variable dependiente (Y) es continua y sigue una distribución Normal, los GLM permiten que Y sea:

- **Binaria:** Éxito/Fracaso, Sí/No (ej. Regresión Logística).
- **De Conteo:** Nº de eventos (ej. Regresión de Poisson).
- **Continua y Positiva** con sesgo (ej. tiempos, costos).

El marco teórico unificador de los GLM es que la distribución de la variable dependiente siempre pertenece a la **familia exponencial**.

Los 3 Componentes de un GLM

Todo Modelo Lineal Generalizado se define por la interacción de tres componentes clave:

1. **Componente Aleatorio:**
   - **Qué es:** La **distribución de probabilidad** que se asume para la variable dependiente (Y).
   - Proviene de la **familia exponencial**.
2. **Componente Sistemático:**
   - **Qué es:** La **combinación lineal** de las variables predictoras (X), que forma el **predictor lineal** (η).
3. **Función de Enlace (*Link Function*):**
   - **Qué es:** El **puente matemático** que conecta el predictor lineal (η) con la media de la variable respuesta (μ).

Componentes en Detalle: Aleatorio y Sistemático

1. Componente Aleatorio (La Distribución)

Define el tipo de datos que estamos modelando. A diferencia de la regresión lineal (solo Normal), en los GLM podemos usar otras distribuciones:

- **Distribución Binomial:** Para variables categóricas binarias (0/1, éxito/fracaso).
- **Distribución de Poisson:** Para datos de conteo (número de eventos).
- **Distribución Gamma:** Para datos continuos y positivos (como costos o tiempos).

2. Componente Sistemático (El Predictor Lineal)

Describe cómo las variables independientes se combinan linealmente. Su forma es idéntica a la de la regresión lineal:

$$\eta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Donde η es el predictor lineal y β son los coeficientes del modelo.

Componentes en Detalle: La Función de Enlace

La función de enlace (g) conecta el predictor lineal (η) con la media de la variable dependiente (μ). Es la clave de la flexibilidad del modelo.

Relación Fundamental:

$$g(\mu) = \eta$$

Esta función transforma la media de Y para que la relación con los predictores se vuelva lineal.

Logística (Logit): Para Regresión Logística.

$$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

Logarítmica: Para Regresión de Poisson.

$$g(\mu) = \log(\mu)$$

Identidad: Para Regresión Lineal estándar (el GLM más simple).

$$g(\mu) = \mu$$

Diferencias Clave: Regresión Lineal vs. GLM

Regresión Lineal Clásica

- **Distribución:** Normal.
- **Tipo de Respuesta:** Continua.
- **Relación:** Lineal y directa.
- **Función de Enlace:** Identidad.

Modelos Lineales Generalizados

- **Distribución:** Familia Exponencial.
- **Tipo de Respuesta:** Flexible (binaria, conteo...).
- **Relación:** Transformada por una función de enlace.
- **Función de Enlace:** Flexible (Logit, Log...).

Ventajas Principales de los GLM:

- **Flexibilidad:** Permiten modelar muchos más tipos de variables dependientes.
- **Interpretación Coherente:** Los coeficientes siguen siendo interpretables de forma rigurosa.
- **Evaluación Robusta:** Se pueden usar las mismas herramientas de evaluación (AIC, BIC, tests de hipótesis).

Estimación de Parámetros en GLM

La estimación en GLM representa un cambio fundamental respecto a la regresión lineal clásica.

- **Regresión Lineal:** Utiliza **Mínimos Cuadrados Ordinarios (MCO)**, un método que funciona bajo supuestos de normalidad y homocedasticidad.
- **GLM:** Necesita métodos más sofisticados debido a las distribuciones no normales y las funciones de enlace no lineales.

El enfoque para los GLM se basa en un principio unificador:

- **El Principio:** La **Máxima Verosimilitud (MLE)**, que proporciona un marco teórico coherente para toda la familia exponencial.
- **El Algoritmo: IRLS** (*Iteratively Reweighted Least Squares*), el método computacional para encontrar la solución de máxima verosimilitud.

El Método de Máxima Verosimilitud (MLE)

A diferencia de Mínimos Cuadrados, el método de **Máxima Verosimilitud** se emplea para estimar los parámetros en un GLM.

¿Por qué es necesario?

- Las distribuciones de la familia exponencial no siempre tienen una relación lineal directa con los predictores.
- La varianza de la respuesta a menudo depende de su media ($Var(Y) = V(\mu)$), violando el supuesto de homocedasticidad que requiere MCO.

Principio Fundamental de MLE

Consiste en encontrar los valores de los parámetros β que hacen **más probable** observar los datos que tenemos.

La Función de Verosimilitud ($L(\beta)$)

Para aplicar el principio de Máxima Verosimilitud, primero definimos la **función de verosimilitud**.

- **Definición:** Es la probabilidad conjunta de observar la totalidad de nuestra muestra $(y_1, \dots, y_n)$ dado un conjunto de parámetros β.
- **Fórmula:**

$$L(\beta) = \prod_{i=1}^{n} f(y_i; \theta_i, \phi)$$

  donde $f(y_i; \theta_i, \phi)$ es la función de densidad o masa de probabilidad de cada observación.
- **El Problema Práctico:** Maximizar una función que es un producto de muchos términos es computacionalmente complejo y puede ser numéricamente inestable.

La Función de Log-Verosimilitud ($\ell(\beta)$)

Para solucionar el problema de los productos, en la práctica se trabaja con el **logaritmo** de la verosimilitud.

- **Definición:** La función de log-verosimilitud es simplemente el logaritmo de la función de verosimilitud.
- **Fórmula:**

$$\ell(\beta) = \log(L(\beta)) = \sum_{i=1}^{n} \log f(y_i; \theta_i, \phi)$$

- **Ventaja Clave:** El logaritmo convierte los **productos en sumas**, lo que simplifica enormemente los cálculos matemáticos y numéricos necesarios para encontrar el máximo de la función.

La Teoría Unificadora: La Familia Exponencial

La clave de los GLM es que todas sus distribuciones (Binomial, Poisson, Gamma...) pertenecen a la **familia exponencial**.

Esto significa que todas se pueden escribir con una **forma matemática unificada**:

$$f(y;\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

donde θ es el **parámetro natural** y ϕ es el **parámetro de dispersión**.

Propiedades Derivadas de esta Forma: Esta estructura unificada permite derivar propiedades generales para todos los GLM de forma elegante:

- **Esperanza (Media):** $E(Y) = \mu = b'(\theta)$.
- **Varianza:** $\text{Var}(Y) = a(\phi)b''(\theta) = a(\phi)V(\mu)$, donde $V(\mu)$ es la **función de varianza** que caracteriza la relación media-varianza de cada distribución.

Distribuciones Comunes

Estas son las distribuciones más frecuentes en la práctica.

- **Normal**
  - **Uso Típico:** Datos continuos simétricos (es el GLM equivalente a la regresión lineal).
  - **Función de Varianza:** $V(\mu) = 1$ (varianza constante).
  - **Enlace Canónico:** Identidad ($g(\mu) = \mu$).
- **Binomial**
  - **Uso Típico:** Proporciones, datos binarios (éxito/fracaso).
  - **Función de Varianza:** $V(\mu) = \mu(1 - \mu)$.
  - **Enlace Canónico:** Logit ($g(\mu) = \log(\frac{\mu}{1-\mu})$).
- **Poisson**
  - **Uso Típico:** Conteos de eventos.
  - **Función de Varianza:** $V(\mu) = \mu$.
  - **Enlace Canónico:** Log ($g(\mu) = \log(\mu)$).

Distribuciones para Datos Asimétricos

Para datos continuos que son estrictamente positivos y tienen sesgo a la derecha.

- **Gamma**
  - **Uso Típico:** Tiempos, costos, o cualquier dato continuo positivo y asimétrico.
  - **Función de Varianza:** $V(\mu) = \mu^2$.
  - **Enlace Canónico:** Inverso ($g(\mu) = 1/\mu$).
- **Inversa Gaussiana**
  - **Uso Típico:** Tiempos hasta un evento, o datos con una asimetría aún más pronunciada que la Gamma.
  - **Función de Varianza:** $V(\mu) = \mu^3$.
  - **Enlace Canónico:** Inverso al cuadrado ($g(\mu) = 1/\mu^2$).

Conceptos Clave de la Familia Exponencial

Para entender y comparar los diferentes GLM, dos conceptos derivados de la familia exponencial son fundamentales:

1. La Función de Varianza: $V(\mu)$

- **Definición:** Es la **"firma"** de cada distribución, ya que define la relación teórica entre la media (μ) y la varianza.
- **Implicación Práctica:** Determina la **heterocedasticidad inherente** de los datos (ej. $V(\mu) = \mu$ en Poisson) y, por tanto, influye directamente en los **pesos** que el algoritmo IRLS asigna a cada observación durante la estimación.

2. El Enlace Canónico: $g(\mu)$

- **Definición:** Es la función de enlace que surge de forma **"natural"** de la estructura matemática de cada distribución.
- **Implicación Práctica:** Aunque en la práctica se pueden probar otros enlaces, el canónico suele garantizar las mejores propiedades estadísticas y una estimación computacionalmente más eficiente.

La Función de Varianza

La relación entre media y varianza es fundamental en los GLM. La función de varianza $V(\mu)$ determina la **heterocedasticidad inherente** de cada distribución e influye en la estimación del modelo.

- **Distribución Binomial:** $V(\mu) = \mu(1-\mu)$. La varianza es máxima cuando la probabilidad $\mu = 0.5$.
- **Distribución de Poisson:** $V(\mu) = \mu$. La varianza aumenta linealmente con la media.
- **Distribución Gamma:** $V(\mu) = \mu^2$. La varianza aumenta cuadráticamente con la media.

Implicación Práctica: Esta función influye directamente en los pesos del algoritmo IRLS. En regresión logística, por ejemplo, las observaciones con probabilidades cercanas a 0.5 tienen mayor varianza y, por tanto, reciben **menor peso** en la estimación.

El Algoritmo de Estimación: IRLS

Las ecuaciones de máxima verosimilitud de los GLM no tienen una solución matemática directa como en la regresión lineal. Por ello, se necesita un **algoritmo iterativo** para encontrar los coeficientes.

El método estándar es **IRLS** (*Iteratively Reweighted Least Squares*), que es una aplicación del método de Newton-Raphson.

¿Cómo funciona conceptualmente? El algoritmo aproxima el problema no lineal del GLM a una **serie de regresiones lineales ponderadas** que se resuelven de forma sucesiva.

1. Se empieza con una estimación inicial de los coeficientes β.
2. En cada paso, se calculan unos **pesos** (w_i) para cada observación. Estos pesos reflejan la "fiabilidad" de cada punto según el modelo actual.
3. Se resuelve una **regresión por mínimos cuadrados ponderada** para obtener una nueva y mejor estimación de β.
4. Se repite el proceso hasta que las estimaciones de β se estabilizan (convergen).

Propiedades de los Estimadores

Los estimadores MLE poseen propiedades **asintóticas** (se cumplen cuando $n \to \infty$) muy deseables, que los validan como el método de estimación preferido.

1. Consistencia: A medida que aumenta el tamaño de la muestra, los estimadores convergen al valor verdadero del parámetro.

$\widehat{\beta} \xrightarrow{p} \beta$ cuando $n \to \infty$.

2. Normalidad Asintótica: Para muestras grandes, la distribución de los estimadores se aproxima a una Normal multivariada.

$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}^{-1}(\beta))$

Permite construir **intervalos de confianza** y realizar **tests de hipótesis**.

3. Eficiencia: Los estimadores MLE alcanzan la cota de Cramér-Rao, lo que significa que tienen la **menor varianza asintótica posible** entre todos los estimadores insesgados.

La Matriz de Información: ¿Cuánta Certeza Tenemos?

La **matriz de información** ($\mathbf{I}(\beta)$) es un concepto clave que cuantifica la certeza de las estimaciones de nuestros coeficientes.

Interpretación Intuitiva:

- Mide la **"curvatura"** de la función de verosimilitud en su punto máximo.
- Una función muy "puntiaguda" (alta curvatura) significa **mucha información** y, por tanto, estimaciones más precisas y fiables.
- Una función más plana (baja curvatura) significa **poca información** y mayor incertidumbre.

Cálculo en la Práctica (GLM): Aunque existen definiciones teóricas basadas en las segundas derivadas de la log-verosimilitud, el algoritmo IRLS nos proporciona una aproximación computacionalmente eficiente y directa:

$$\mathbf{I}(\widehat{\beta}) \approx \mathbf{X}^T\mathbf{W}\mathbf{X}$$

donde $\mathbf{W}$ es la matriz diagonal de pesos calculada en la última iteración del algoritmo.

Cálculo de los Errores Estándar

Los **errores estándar** de los coeficientes individuales se obtienen a partir de la matriz de información.

Proceso de Cálculo:

1. **Matriz de Covarianza:** Se calcula como la **inversa** de la matriz de información.
2. **Errores Estándar:** Se obtienen como las raíces cuadradas de los elementos diagonales de esa matriz de covarianza.

Fórmula:

$$\text{SE}(\hat{\beta}_j) = \sqrt{[\mathbf{I}^{-1}(\hat{\beta})]_{jj}}$$

Uso de los Errores Estándar en Inferencia

Estos errores estándar son fundamentales para la inferencia estadística.

Aplicaciones Principales:

- **Intervalos de Confianza:**

$$\hat{\beta}_j \pm z_{\alpha/2} \cdot \mathsf{SE}(\hat{\beta}_j)$$

- **Estadísticos de Prueba (de Wald):**

$$z_j = \frac{\hat{\beta}_j}{\mathsf{SE}(\hat{\beta}_j)}$$

- **Evaluación de la Precisión** de nuestras estimaciones.

Advertencia Importante:

- Estos errores estándar **son válidos bajo los supuestos del modelo GLM**.
- Violaciones serias de estos supuestos (como **sobredispersión** en modelos de Poisson) pueden hacer que sean inadecuados.

La Deviance: Definición y Cálculo

La **deviance** es la medida principal de ajuste en GLM; es una generalización de la suma de cuadrados residuales.

Idea Fundamental: Mide la discrepancia entre la verosimilitud de nuestro **modelo propuesto** y la de un **modelo saturado** (un modelo teóricamente perfecto que ajusta cada dato).

$$D = 2 \sum_{i=1}^{n} [\ell(\text{modelo saturado}) - \ell(\text{modelo propuesto})]$$

Interpretación práctica:

- Deviance $= 0$: Modelo perfecto que ajusta exactamente todos los datos observados
- Deviance baja: Buen ajuste del modelo a los datos
- Deviance alta: Mal ajuste del modelo, sugiere que el modelo no captura adecuadamente los patrones en los datos

Interpretando la Deviance: Una Comparación Clave

La clave para interpretar la deviance es **comparar** la de tu modelo con la de un modelo base.

1. El Punto de Partida: La Deviance Nula

- **¿Qué es?** La deviance de un modelo simple, solo con intercepto (que ignora todos tus predictores).
- **Analogía:** Es el **"error máximo"** que sirve como punto de comparación.

2. El Resultado: La Deviance Residual

- **¿Qué es?** La deviance de tu modelo final, con todos los predictores incluidos.
- **Analogía:** Es el **"error restante"** después de que tus predictores han hecho su trabajo.

La **reducción de la deviance** (`Deviance Nula - Deviance Residual`) representa la **mejora en el ajuste** que se debe a tus predictores.

El Test de la Razón de Verosimilitudes (LRT)

Es la herramienta principal para **comparar modelos anidados**, basándose en el principio de parsimonia.

Estadístico de Prueba: La diferencia en las deviances sigue una distribución **chi-cuadrado**:

$$LRT = D_{\text{reducido}} - D_{\text{completo}} \sim \chi^2_{df}$$

(donde df es la diferencia en el número de parámetros).

Regla de Decisión:

- H_0: El modelo reducido es suficiente.
- Si el **p-valor** $< \alpha$, se rechaza H_0. La mejora del modelo completo es **estadísticamente significativa**.

LRT en la Práctica: Ejemplo con R

Objetivo: Determinar si añadir la variable `disp` (cilindrada) a un modelo que ya contiene `wt` (peso) mejora significativamente la predicción.

Salida del Test en R

```
Analysis of Deviance Table

Model 1: am ~ wt
Model 2: am ~ wt + disp

  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        30     19.176
2        29     17.785  1   1.3913   0.2382
```

LRT en la Práctica: Ejemplo con R

Interpretación del Resultado

El test compara la **Deviance** de ambos modelos. La hipótesis nula (H_0) es que el modelo reducido es suficiente.

- La diferencia en deviance es de **1.391** con **1** grado de libertad.
- El p-valor asociado es **0.238**.

Dado que el p-valor es mayor que 0.05, **no rechazamos la hipótesis nula**.

Conclusión: Añadir `disp` no aporta una mejora significativa. Nos quedamos con el **modelo reducido** por parsimonia.

Diagnosis de GLMs

La diagnosis de GLMs es el proceso de evaluar los supuestos del modelo y detectar problemas que puedan afectar la validez de las inferencias.

¿Por qué es diferente de la Regresión Lineal?

- A diferencia de la regresión lineal, los residuos ordinarios no son suficientes.
- Los GLM requieren herramientas especializadas debido a la **heterocedasticidad inherente** y a las diferentes distribuciones subyacentes (Poisson, Binomial, etc.).

Nuestro Enfoque: Abordaremos el diagnóstico respondiendo a tres preguntas clave, utilizando diferentes tipos de **residuos** para obtener las respuestas.

Tipos de Residuos en GLMs

Los residuos "crudos" ($y_i - \hat{\mu}_i$) no son homocedásticos, por lo que se utilizan versiones estandarizadas. Los más importantes son:

- **Residuos Pearson:**
  - Son un análogo directo a los residuos estandarizados en regresión lineal. Estandarizan el residuo crudo dividiendo por la desviación estándar predicha por el modelo.

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

- **Residuos Deviance:**
  - Son los **más recomendados** para la inspección visual en gráficos diagnósticos.
  - Su distribución se aproxima mejor a la normalidad y su varianza es más estable que la de otros residuos.

$$d_i = \text{sign}(y_i - \hat{\mu}_i)\sqrt{2[l_i(y_i) - l_i(\hat{\mu}_i)]}$$

¿La Forma del Modelo es Correcta?

Se evalúa si la estructura básica del modelo, $g(\mu) = X\beta$, es adecuada para los datos.

Herramientas de Diagnóstico:

- **Gráfico de Residuos vs. Valores Ajustados:**
  - Es la herramienta fundamental. Se grafican los residuos (idealmente, deviance) contra el predictor lineal ($\hat{\eta}_i$).
  - **Un patrón curvilíneo** es una señal clara de que la forma funcional o la función de enlace son incorrectas.
- **Gráficos de Residuos Parciales:** Evalúan si la relación es apropiada para **cada predictor individualmente**.
- **Test de Especificación de Enlace (Linktest):** Es una prueba formal. Si el predictor lineal al cuadrado ($\hat{\eta}^2$) resulta significativo al añadirlo al modelo, es evidencia de que la función de enlace está mal especificada.

¿La Distribución es Correcta?

Se evalúa si la elección de la familia de distribución (Poisson, Binomial, etc.) fue acertada, empezando por la relación entre la media y la varianza.

Sobredispersión: Ocurre cuando la varianza real de los datos es **mayor** que la media, violando el supuesto de modelos como el de Poisson ($Var(Y) = \mu$).

- **Consecuencia:** Invalida las inferencias (errores estándar demasiado pequeños, p-valores incorrectos).
- **Detección:** Se calcula el **Estadístico de dispersión ($\hat{\phi}$)**. Si $\hat{\phi}$ **es significativamente mayor que 1**, hay sobredispersión.

$$\hat{\phi} = \frac{\sum r_i^2}{n - p}$$

- **Solución:** Cambiar a un modelo más flexible. El caso clásico es pasar de **Poisson** a **Binomial Negativo**, ya que este último incluye un parámetro (α) para modelar la variabilidad extra: $Var(Y) = \mu + \alpha\mu^2$.

¿La Distribución es Correcta?

Un segundo aspecto para verificar la distribución es la forma general de los errores.

Herramienta: El **Gráfico Q-Q de residuos deviance**.

- **Concepto:** Aunque los errores de un GLM no son estrictamente normales, los **residuos deviance** sí deberían tener una distribución aproximadamente normal si el modelo está bien especificado.
- **Interpretación:**
  - Se grafican los cuantiles de los residuos deviance contra los cuantiles teóricos de una distribución normal.
  - Los puntos deberían seguir de cerca la línea diagonal.
  - **Desviaciones sistemáticas** de la línea pueden indicar que la familia de distribución asumida (Poisson, Binomial, etc.) es incorrecta.

¿Hay Observaciones que Distorsionan el Modelo?

Finalmente, buscamos identificar puntos individuales que tienen una influencia desproporcionada en los coeficientes del modelo.

Herramientas Matemáticas Clave:

- **Leverage Generalizado (h_i):** Mide el potencial de una observación para ser influyente debido a su posición en el espacio de los predictores.
- **Distancia de Cook para GLMs (D_i):** Mide la influencia global de una observación en *todos* los coeficientes.

$$D_i = \frac{r_i^2 h_i}{p(1-h_i)^2}$$

- **DFBETAS:** Mide la influencia de una observación en *cada coeficiente individual*.

Estrategia:

- La herramienta visual principal es el **gráfico de residuos vs. leverage**.
- La estrategia ante estas observaciones no es eliminarlas automáticamente, sino **investigarlas** para entender su naturaleza.

Regresión Logística: Modelando Eventos Binarios

La **regresión logística** es la herramienta fundamental de los GLM para modelar la probabilidad de ocurrencia de un **evento binario**.

Objetivo:

Modelar una variable dependiente que solo toma dos valores:

- **Éxito / Fracaso**
- **Sí / No**
- **Enfermo / Sano**

Pilares del Tema:

1. **Fundamentos:** La función sigmoide y el enlace Logit.
2. **Estimación:** Máxima Verosimilitud (MLE) y el algoritmo IRLS.
3. **Interpretación:** El concepto clave de los **Odds Ratios**.
4. **Evaluación:** Bondad de ajuste (Deviance, Pseudo R^2) y validación (Matriz de Confusión, Curva ROC).

Fundamentos: El Problema y la Función Sigmoide

El Problema: La regresión lineal puede predecir valores fuera del rango [0,1], lo cual no tiene sentido para modelar una probabilidad.

La Solución: La **función logística (o sigmoide)** transforma cualquier valor real del predictor lineal (η) en una probabilidad entre 0 y 1.

Fórmula de la Probabilidad:

$$P(Y = 1|X) = \frac{1}{1 + e^{-\eta}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots)}}$$

La curva en forma de "S" asegura que las predicciones se aplanen hacia 0 y 1 en los extremos.

Fundamentos: La Función de Enlace Logit

Para poder usar un modelo lineal ($X\beta$), necesitamos transformar la probabilidad (que está en la escala [0,1]) a una escala que vaya de $-\infty$ a $+\infty$.

La Herramienta: Esto se logra con la **función de enlace logit**.

Definición del Logit: El logit de una probabilidad p es el logaritmo de los *odds* (razón de probabilidades):

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

El Modelo Logístico Linealizado: Esta transformación nos permite expresar el modelo de forma lineal:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Estimación por Máxima Verosimilitud (MLE)

Función de Verosimilitud ($L(\beta)$): Para un resultado binario $y_i \in \{0, 1\}$, la verosimilitud se define como:

$$L(\beta) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i}$$

Función de Log-Verosimilitud ($\ell(\beta)$): Maximizamos el logaritmo de la función anterior:

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \mathbf{x}_i^T \beta - \log(1 + e^{\mathbf{x}_i^T \beta}) \right]$$

Ecuaciones de Puntuación (*Score Equations*): Para encontrar el máximo, se deriva la log-verosimilitud respecto a cada β_j y se iguala a cero:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}(y_i - p_i) = 0$$

Esto significa que la solución se encuentra cuando la suma de los residuos $(y_i - p_i)$ ponderados por cada predictor es cero.

Implementación del Algoritmo IRLS

Como las ecuaciones de verosimilitud no tienen solución analítica cerrada, se utiliza el algoritmo IRLS. Para la regresión logística, los componentes específicos son:

Pesos (w_i): El peso de cada observación es la varianza de una distribución Bernoulli, que es máxima cuando la probabilidad es 0.5.

$$w_i = p_i(1 - p_i)$$

Variable Dependiente Ajustada (z_i): Es la versión linealizada de la respuesta en cada iteración.

$$z_i = \eta_i^{(t)} + \frac{y_i - p_i^{(t)}}{w_i^{(t)}}$$

Estos componentes se utilizan en cada paso de la actualización de los coeficientes: $\beta^{(t+1)} = (\mathbf{X}^T\mathbf{W}^{(t)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{(t)}\mathbf{z}^{(t)}$.

Ejemplo: Verificación de la Estimación MLE en R

Podemos verificar que los coeficientes encontrados por `glm()` son, en efecto, los que maximizan la función de log-verosimilitud. A continuación se muestra la salida del código R:

Salida del Análisis:

```
Coeficientes estimados por glm():
(Intercept)         glu         bmi
-8.16560411  0.03433555  0.08585474

Log-verosimilitud en el óptimo: -101.4057

------------------------------------
Iteraciones necesarias: 4
¿Convergió? TRUE
```

Conclusión: La salida confirma que el algoritmo convergió en 4 iteraciones para encontrar los coeficientes que maximizan la verosimilitud del modelo.

Interpretación: Probabilidad vs. Odds

Los coeficientes β están en la escala del logit, por lo que no son directamente interpretables en términos de probabilidad. Para interpretarlos, primero necesitamos entender el concepto de **odds**.

Definición de Odds:

- El **odds** es la razón entre la probabilidad de que un evento ocurra y la de que no ocurra.

$$\text{odds} = \frac{p}{1-p}$$

- El modelo logístico es, en esencia, un modelo lineal para el **logaritmo de los odds**: $\log(\text{odds}) = X\beta$.

Ejemplo:

- Si la probabilidad de éxito $p = 0.8$.
- El odds es $\frac{0.8}{0.2} = 4$.
- **Interpretación:** El evento es **4 veces más probable** que ocurra a que no ocurra.

Interpretación: El Odds Ratio (OR)

El **Odds Ratio (OR)** es la herramienta principal para interpretar los coeficientes de una regresión logística.

- **Concepto:** Mide el cambio **multiplicativo** en los *odds* por cada incremento de una unidad en un predictor X_j, manteniendo el resto de variables constantes.
- **Cálculo:** Se obtiene exponenciando el coeficiente:

$$\text{OR} = e^{\beta_j}$$

- **Interpretación:**
  - **OR $>$ 1:** El odds aumenta (el evento se vuelve más probable).
  - **OR $<$ 1:** El odds disminuye (el evento se vuelve menos probable).
  - **OR = 1:** No hay efecto.

Ejemplo: Si $\beta_{BMI} = 0.08$, entonces $OR = e^{0.08} \approx 1.083$. Por cada unidad que aumenta el BMI, el odds de tener diabetes se multiplica por 1.083 (es decir, aumenta un **8.3%**).

Bondad de Ajuste: Deviance y Pseudo R^2

El R^2 tradicional no es aplicable en este contexto. La bondad de ajuste en regresión logística se evalúa con métricas basadas en la verosimilitud.

1. Deviance

- Compara la log-verosimilitud de nuestro modelo con la de un modelo saturado. La fórmula específica para la distribución binomial es:

$$D = 2 \sum_{i=1}^{n} \left[y_i \log \left(\frac{y_i}{\hat{p}_i} \right) + (1 - y_i) \log \left(\frac{1 - y_i}{1 - \hat{p}_i} \right) \right]$$

2. Pseudo R^2

- Son análogos al R^2 que miden la mejora en la verosimilitud del modelo en comparación con un modelo nulo (solo con intercepto).
- No representan la "proporción de varianza explicada", sino la mejora en el ajuste del modelo.

Bondad de Ajuste: Tipos de Pseudo R^2

Existen varias formulaciones para el Pseudo R^2. Las más comunes son:

- **McFadden's R^2:** Es el más utilizado.

$$R^2_{\text{McFadden}} = 1 - \frac{\ell_{\text{modelo}}}{\ell_{\text{nulo}}}$$

  - Valores entre 0.2 y 0.4 se consideran indicativos de un buen ajuste.

- **Cox-Snell R^2:**

$$R^2_{\text{Cox-Snell}} = 1 - \left(\frac{L_{\text{nulo}}}{L_{\text{modelo}}}\right)^{2/n}$$

- **Nagelkerke R^2:** Es una corrección del Cox-Snell para que su valor máximo sea 1, haciéndolo más comparable al R^2 tradicional.

$$R^2_{\text{Nagelkerke}} = \frac{R^2_{\text{Cox-Snell}}}{1 - (L_{\text{nulo}})^{2/n}}$$

Validación: La Matriz de Confusión

La validación de un modelo logístico se centra en su **capacidad de clasificación**.

La Matriz de Confusión

- Es la herramienta fundamental. Compara las clases predichas por el modelo con las clases reales.
- **Proceso:** Se convierten las probabilidades predichas ($\hat{p}_i$) en clases ("Sí" / "No") usando un **umbral de decisión** (típicamente 0.5).

Esto genera cuatro posibles resultados:

- **Verdaderos Positivos (VP):** Predijo "Sí" y era "Sí".
- **Falsos Positivos (FP):** Predijo "Sí" pero era "No" (Error Tipo I).
- **Verdaderos Negativos (VN):** Predijo "No" y era "No".
- **Falsos Negativos (FN):** Predijo "No" pero era "Sí" (Error Tipo II).

Validación: Métricas de Clasificación

A partir de la matriz de confusión, se calculan las métricas de rendimiento clave:

- **Precisión (Accuracy):**
  - $\frac{VP+VN}{\text{Total}}$
  - Proporción total de predicciones correctas. Cuidado: puede ser engañosa en datasets desbalanceados.
- **Sensibilidad (Recall o Tasa de VP):**
  - $\frac{VP}{VP+FN}$
  - De todos los positivos reales, ¿qué proporción clasificamos correctamente? Mide la capacidad para identificar los casos positivos.
- **Especificidad:**
  - $\frac{VN}{VN+FP}$
  - De todos los negativos reales, ¿qué proporción clasificamos correctamente? Mide la capacidad para identificar los casos negativos.

Validación: Curva ROC y AUC

La Curva ROC (*Receiver Operating Characteristic*)

- Es una evaluación global del rendimiento del modelo, **independiente del umbral de decisión**.
- Grafica la **Sensibilidad** (Tasa de VP) en el eje Y frente a **1 - Especificidad** (Tasa de FP) en el eje X para todos los umbrales posibles.

AUC (Área Bajo la Curva ROC)

- Cuantifica la capacidad discriminativa del modelo en un solo número (de 0.5 a 1.0).
  - **AUC = 1.0:** Clasificador perfecto.
  - **AUC = 0.5:** Clasificador inútil (equivalente al azar).
  - **Típicamente, AUC $>$ 0.8** se considera una buena discriminación.

Sí, es una idea excelente. Integrar los ejemplos prácticos es fundamental para conectar la teoría con la aplicación en R.

Caso Práctico: Resultados del Modelo Logístico

Ajustamos un modelo para predecir la diabetes en el dataset `Pima.tr`. Los resultados clave de la validación del modelo son los siguientes:

Resultados Numéricos

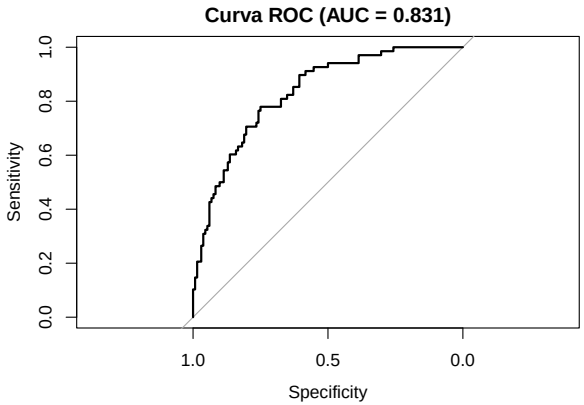
Matriz de Confusión:

Predicted	Actual No	Actual Yes
No	114	29
Yes	18	39

Métricas Clave:

- **Exactitud:** 0.785
- **AUC:** 0.831

Evaluación Visual: Curva ROC



Regresión de Poisson: Modelando Datos de Conteo

Es la técnica de GLM utilizada para modelar **datos de conteo**: una variable que representa el número de veces que ocurre un evento en un intervalo.

- **Tipo de Variable:** La respuesta toma valores enteros no negativos $(0, 1, 2, ...)$ y se asume que sigue una **distribución de Poisson**.
- **Función de Probabilidad de Poisson:**

$$P(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}$$

donde λ es la **tasa media de ocurrencia** del evento.

El Modelo de Regresión de Poisson

El objetivo del modelo es explicar la relación entre la **tasa de ocurrencia de los eventos** (λ) y un conjunto de variables predictoras X.

- **Forma Funcional:** Se utiliza una **función de enlace logarítmica** para asegurar que la tasa λ sea siempre positiva.

$$\log(\lambda) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- **Tasa Esperada:** El modelo puede expresarse en términos de la tasa esperada de eventos como:

$$\lambda = e^{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}$$

Supuestos y Limitaciones Clave

Para que el modelo sea adecuado, se deben cumplir ciertos supuestos:

- **Independencia de los eventos**.
- **Equidispersión:** El supuesto fundamental de la distribución de Poisson es que la **media es igual a la varianza**:

$$E(Y) = Var(Y) = \lambda$$

Limitaciones Comunes (Violación de Supuestos):

- **Sobredispersión:** Ocurre cuando la varianza es mayor que la media ($Var(Y) > E(Y)$). La solución es usar una **Regresión Binomial Negativa**.
- **Exceso de Ceros:** Si hay más ceros en los datos de los que predice el modelo. La solución es usar modelos **ZIP** (*Zero-Inflated Poisson*).

Interpretación de Coeficientes

Los coeficientes β están en la escala logarítmica de la tasa, por lo que para una interpretación práctica, los exponenciamos.

- **Incidence Rate Ratio (IRR):**

$$\text{IRR} = e^{\beta_j}$$

- **Interpretación:** El IRR es un **factor multiplicativo** que nos dice cuánto cambia la tasa de eventos esperada por cada incremento de una unidad en el predictor X_j.
  - **IRR > 1:** La tasa de eventos aumenta. Un IRR de 1.25 es un aumento del 25%.
  - **IRR < 1:** La tasa de eventos disminuye. Un IRR de 0.80 es una disminución del 20%.
  - **IRR $= 1$:** No hay efecto.

Estimación por Máxima Verosimilitud

La estimación se adapta a la distribución de Poisson con enlace logarítmico.

- **Función de Verosimilitud ($L(\beta)$):**

$$L(\beta) = \prod_{i=1}^{n} \frac{e^{-\lambda_i} \lambda_i^{y_i}}{y_i!}$$

  donde $\lambda_i = e^{\mathbf{x}_i^T \beta}$.
- **Función de Log-Verosimilitud ($\ell(\beta)$):**

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \mathbf{x}_i^T \beta - e^{\mathbf{x}_i^T \beta} - \log(y_i!) \right]$$

- **Ecuaciones de Puntuación:** La solución de máxima verosimilitud se encuentra cuando:

$$\frac{\partial \ell}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij}(y_i - \lambda_i) = 0$$

Ejemplo Práctico: Salida del Modelo Poisson

Objetivo: Ajustamos un modelo para predecir el `número de accidentes` en función del `tráfico` y la `visibilidad`.

Salida de Coeficientes (`summary`)

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  9.607e-04  2.316e-03   0.415  0.678
trafico      9.999e-03  1.360e-06 < 2e-16 ***
visibilidad -2.000e-01  1.012e-04 < 2e-16 ***
```

Métricas Globales del Modelo

- **Null deviance:** 1.68e+08 (con 99 g.l.)
- **Residual deviance:** 89.3 (con 97 g.l.)
- **AIC:** 1220.4

Ejemplo Práctico: Interpretación de Resultados

Significancia de los Predictores

Basado en los p-valores (`Pr(>|z|)`) de la diapositiva anterior:

- Tanto `trafico` como `visibilidad` son predictores **altamente significativos** (sus p-valores son prácticamente cero).

Interpretación de los Coeficientes (vía IRR)

Para interpretar el efecto práctico, exponenciamos los coeficientes (IRR $= e^{\beta}$):

- **IRR (tráfico):** $e^{0.01} \approx 1.01$.
  - Un aumento de 1 unidad en `trafico` **incrementa** la tasa de accidentes esperada en un **1%**.
- **IRR (visibilidad):** $e^{-0.20} \approx 0.82$.
  - Un aumento de 1 km en `visibilidad` **reduce** la tasa de accidentes esperada en un **18%**.

Implementación del Algoritmo IRLS para Poisson

Como no hay solución analítica cerrada, se utiliza el algoritmo IRLS con componentes específicos para Poisson.

- **Pesos (w_i):** El peso de cada observación es simplemente la tasa esperada.

$$w_i = \lambda_i$$

- **Variable Dependiente Ajustada (z_i):**

$$z_i = \log(\lambda_i^{(t)}) + \frac{y_i - \lambda_i^{(t)}}{\lambda_i^{(t)}}$$

- **Actualización de parámetros:**

$$\beta^{(t+1)} = (\mathbf{X}^T\mathbf{W}^{(t)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{(t)}\mathbf{z}^{(t)}$$

Propiedades Específicas de la Estimación Poisson

La estimación MLE en el modelo de Poisson tiene características particulares:

1. **Equidispersión:** El modelo asume que la varianza aumenta linealmente con la media ($E(Y_i) = \text{Var}(Y_i) = \lambda_i$).
2. **Convergencia Rápida:** Generalmente requiere menos iteraciones que la regresión logística.
3. **Estabilidad Numérica:** El enlace logarítmico garantiza automáticamente que las tasas estimadas λ_i sean siempre positivas.
4. **Interpretación Multiplicativa:** Los coeficientes se interpretan naturalmente como efectos multiplicativos sobre la tasa.

Bondad de Ajuste en Regresión de Poisson

La métrica teórica fundamental sigue siendo la **deviance**, pero en la práctica, la prueba de bondad de ajuste más importante es la **evaluación de la sobredispersión**.

- **Herramienta de Diagnóstico:** El **estadístico de dispersión ($\hat{\phi}$)** se convierte en la medida de facto del ajuste.

$$\hat{\phi} = \frac{X^2_{\text{Pearson}}}{n-p}$$

- **Interpretación:**
  - Si $\hat{\phi} \approx 1$: El supuesto de equidispersión se cumple y el ajuste es adecuado.
  - Si $\hat{\phi} \gg 1$: Hay **sobredispersión**. El modelo no se ajusta bien a la variabilidad de los datos, y se debe considerar una **Regresión Binomial Negativa**.

Validación del Modelo de Poisson

La validación se enfoca en la **capacidad de predicción**: ¿qué tan cerca están los conteos predichos de los conteos reales?

- **Proceso:** Se ajusta el modelo en un conjunto de **entrenamiento** y se evalúa su rendimiento en un conjunto de **prueba**.
- **Métricas de Validación Principales:**
  - **Raíz del Error Cuadrático Medio (RMSE):** Mide la desviación estándar de los residuos. Penaliza más los errores grandes.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum(y_i - \hat{\mu}_i)^2}$$

  - **Error Absoluto Medio (MAE):** Mide la magnitud promedio de los errores. Es menos sensible a outliers.

$$\text{MAE} = \frac{1}{n}\sum|y_i - \hat{\mu}_i|$$

- **Herramienta Visual: Gráfico de valores predichos vs. valores reales**. En un buen modelo, los puntos deben agruparse cerca de la línea diagonal $y = x$.

Ejemplo Práctico: Diagnóstico del Modelo Poisson

Este segundo ejemplo se centra en verificar dos aspectos clave de la estimación: la **convergencia del algoritmo** y el supuesto de **equidispersión**.

Comprobamos si se cumple el supuesto clave de Poisson ($media \approx varianza$).

```
Verificación de Equidispersión:

Media observada:     0.15
Varianza observada:  0.149
Razón varianza/media: 0.993
```

- **Conclusión:** La razón es **muy cercana a 1**, por lo que **se cumple el supuesto**. No hay evidencia de sobredispersión y el modelo de Poisson es adecuado.

Otros GLMs: Adaptándose a la Realidad de los Datos

Más allá de la regresión logística y de Poisson, existen otros GLM para manejar situaciones más complejas.

Estos modelos son especialmente útiles cuando los datos presentan características como:

- **Sobredispersión:** La varianza es mayor de lo esperado.
- **Sesgo:** La distribución de los datos es asimétrica.
- **Restricciones en el dominio:** La variable respuesta solo puede tomar valores positivos.

Exploraremos los tres modelos más importantes para estos casos:

- **Regresión Binomial Negativa**
- **Regresión Gamma**
- **Regresión Inversa Gaussiana**

Regresión Binomial Negativa

El Problema: Sobredispersión

- Ocurre en datos de conteo cuando la **varianza es mayor que la media**, violando el supuesto clave de la regresión de Poisson ($Var(Y) = \mu$).
- **Causas comunes:** Heterogeneidad no modelada, dependencia entre eventos o exceso de ceros.
- **Consecuencia:** La regresión de Poisson subestima los errores estándar, llevando a conclusiones incorrectas sobre la significancia de los predictores.

La Solución: El Modelo Binomial Negativo

- Es una extensión del modelo de Poisson que introduce un **parámetro de dispersión (α)** para permitir que la varianza sea mayor que la media:

$$Var(Y) = \mu + \alpha\mu^2$$

- Si $\alpha = 0$, el modelo se reduce a la regresión de Poisson.

Interpretación y Comparación

La forma funcional del modelo es la misma que la de Poisson (con enlace logarítmico), pero debemos interpretar el nuevo parámetro de dispersión y comparar ambos modelos.

Interpretación del Parámetro de Dispersión (θ)

- El software (como la función `glm.nb` en R) estima un parámetro θ, donde $\alpha = 1/\theta$.
- **Valores altos de** θ (ej. > 100): Poca sobredispersión. El modelo es similar a Poisson.
- **Valores bajos de** θ (ej. < 10): Mucha sobredispersión. El modelo Binomial Negativo es claramente más apropiado.

Comparación de Modelos (Poisson vs. Binomial Negativa)

- Se utiliza el **Criterio de Información de Akaike (AIC)**.
- Si el **AIC de la Binomial Negativa es menor** que el AIC de Poisson, debemos preferir el modelo Binomial Negativo.

Modelos para Variables Continuas No Normales

Cuando la variable dependiente (Y) es **continua**, pero **no sigue una distribución normal**, la regresión lineal clásica no es adecuada.

Este escenario es común en variables que son:

- **Estrictamente positivas** (ej. costos, tiempos).
- Tienen una distribución **sesgada a la derecha**.

Los GLM nos ofrecen alternativas como la **Regresión Gamma** y la **Regresión Inversa Gaussiana**.

Comparación: Regresión Gamma vs. Inversa Gaussiana

Regresión Gamma

- **Uso Típico:** Para variables continuas **positivas y con sesgo a la derecha** (tiempos, costos, reclamos de seguros).
- **Varianza:** Aumenta proporcionalmente al **cuadrado de la media** ($V(\mu) = \mu^2$).
- **Enlace Común:** Logarítmico ($\log(\mu) = X\beta$).

Regresión Inversa Gaussiana

- **Uso Típico:** Para tiempos de respuesta o variables con un **sesgo aún más pronunciado** que la Gamma.
- **Varianza:** Aumenta proporcionalmente al **cubo de la media** ($V(\mu) = \mu^3$).
- **Enlace Común:** Inverso al cuadrado ($1/\mu^2 = X\beta$).

Guía de Decisión: ¿Qué GLM Debo Usar?

La elección del modelo depende casi exclusivamente de la naturaleza de tu variable respuesta (Y).

- **¿Es Y binaria (0/1, Éxito/Fracaso)?** Usa **Regresión Logística**.
- **¿Es Y un conteo de eventos (nº de accidentes, nº de clientes)?**

Empieza con una **Regresión de Poisson**.

- **Importante:** Después, comprueba si hay **sobredispersión**. Si la hay ($\hat{\phi} > 1.5$), cambia a una **Regresión Binomial Negativa**.
- **¿Es Y continua, positiva y con sesgo a la derecha (tiempos, costos)?** Usa **Regresión Gamma**. Es una excelente alternativa a transformar la variable con logaritmos y usar un modelo lineal.
- **¿Es Y un tiempo hasta un evento con una asimetría muy pronunciada?** Considera una **Regresión Inversa Gaussiana**.