Maestría en Ciencia y Análisis de Datos- Universidad Mayor de San Andrés

Modelos lineales y modelos lineales generalizados

Rolando Gonzales Martinez, PhD

Fellow postdoctoral Marie Skłodowska-Curie

Universidad de Groningen (Países Bajos)

Investigador (researcher)

Iniciativa de Pobreza y Desarrollo Humano de la Universidad de Oxford (UK)

Contenido del curso

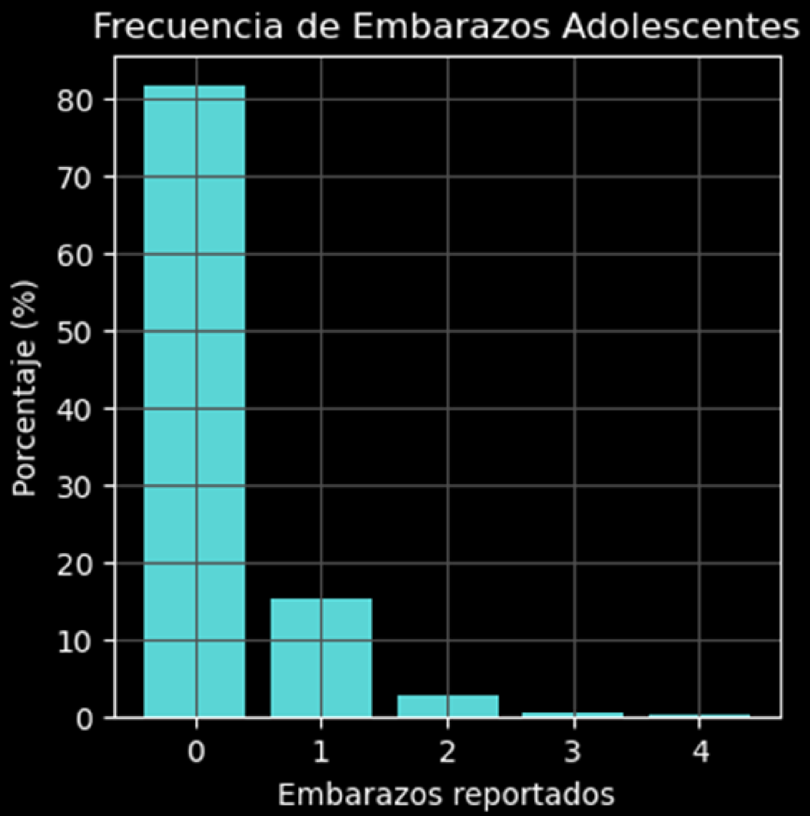
(3) Modelos Lineales Generalizados (MLG)

- Concepto de MLG.
- Funciones de enlace.
- Modelos estadísticos con distribuciones usualmente aplicadas en MLG: normal, binomial, Possion.
- Otros modelos: modeo SIR, modelos de regresión espacial y modelo logit para datos con variable dependiente desproporcionada
- Laboratorio: Implementación de MLG en problemas de regresión y clasificación.

Modelos lineales generalizados: motivación

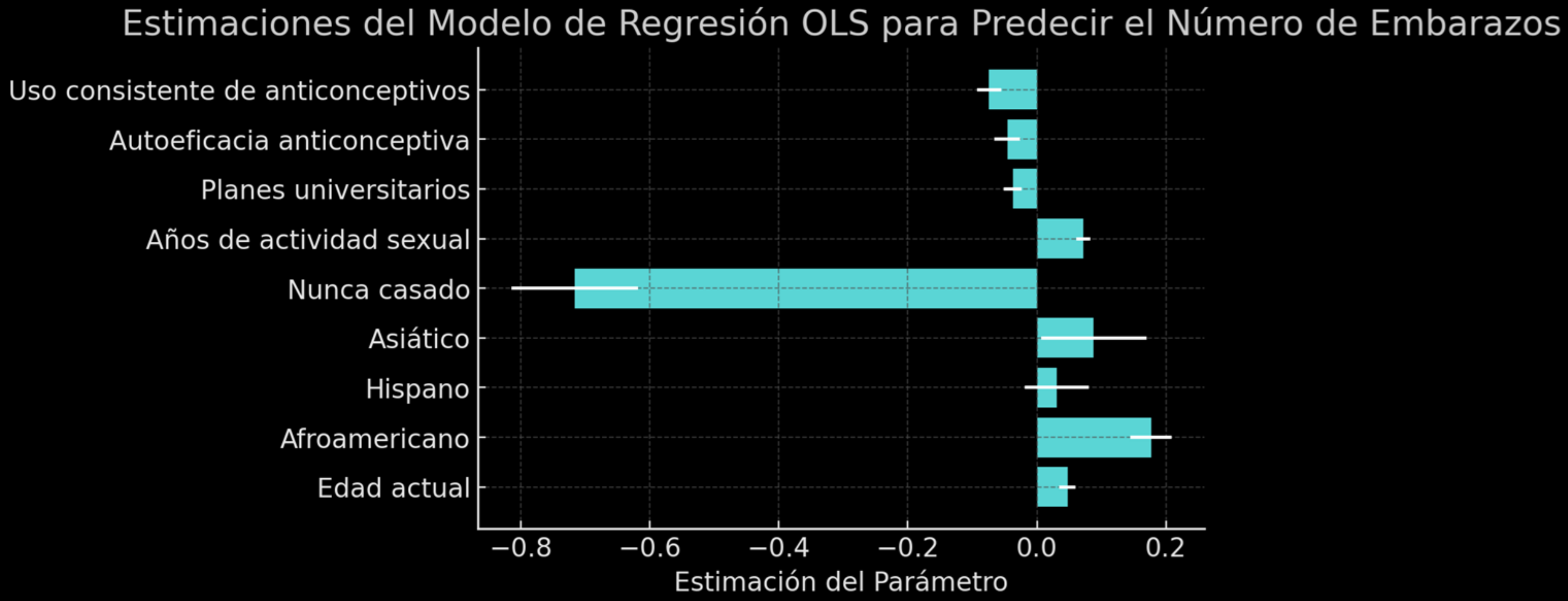
Embarazos reportados	n	%
0	1,011	81.5
1	190	15.3
2	32	2.6
3	6	0.5
4	2	0.2

Hutchinson, M. K., & Holtman, M. C. (2005). Analysis of count data using poisson regression. *Research in nursing & health*, *28*(5), 408-418.



Modelos lineales generalizados: motivación

Estimación MCO:



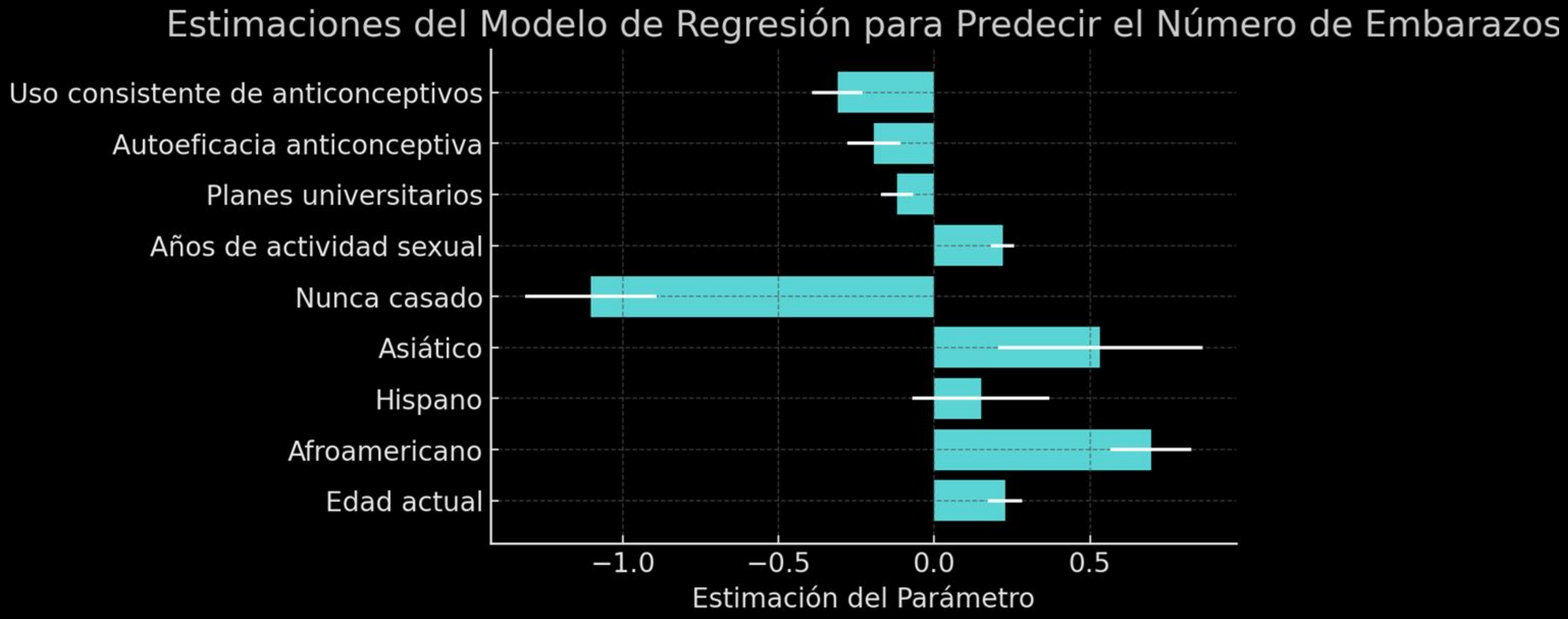
Modelos lineales generalizados: motivación

Estimación MCO:

Variable	Parametro estimado	Error Estandar	Valor t	p-value
Intercepto	0.3741	0.22441	1.67	0.0958
Edad actual	0.04718	0.01223	3.86	0.0001
Afroamericana	0.17686	0.03219	5.49	0.0001
Hispana	0.03127	0.04971	0.63	0.5295
Asiática	0.08816	0.08171	1.08	0.2808
Nunca casada	-0.71671	0.09796	-7.32	0.0001
Años de actividad sexual	0.07221	0.01092	6.61	0.0001
Planes universitarios	-0.03711	0.01402	-2.65	0.0082
Auto eficacia anticonceptiva	-0.04573	0.02005	-2.28	0.0227
Uso consistente de anticonceptivos	-0.07412	0.01879	-3.94	0.0001

Modelos lineales generalizados: motivación

Estimación de MV de un modelo de Poisson:



Modelos lineales generalizados: motivación

Estimación de MV de un modelo de Poisson:

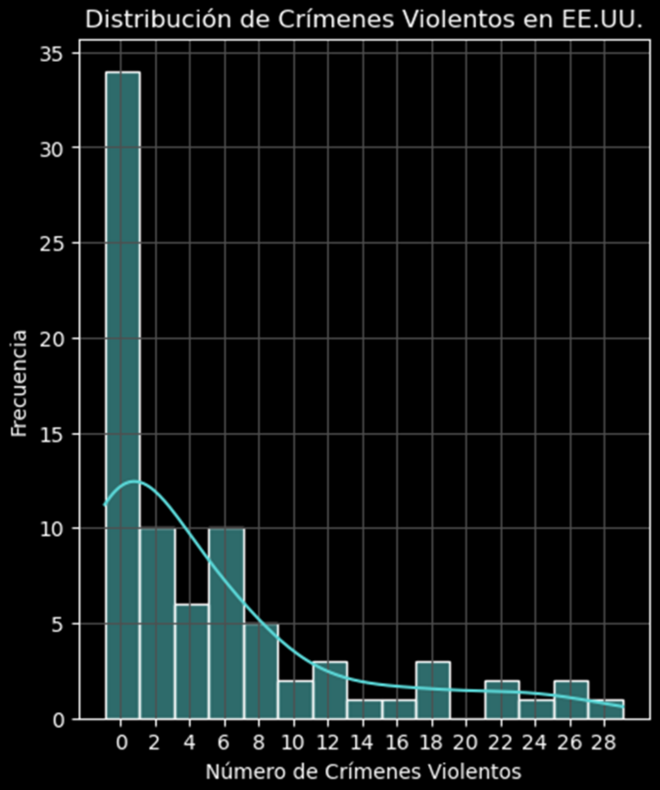
	Parámetro estimado	Error estándar	χ^2	p-value
Intercepto	-3.5182	0.9551	13.57	0.0002
Edad actual	0.2278	0.0556	16.78	<0.0001
Afroamericana	0.6958	0.1297	28.78	<0.0001
Hispana	0.1498	0.2195	0.47	0.4949
Asiática	0.5333	0.329	2.63	0.1051
Nunca casada	-1.1035	0.2122	27.04	<0.0001
Años de actividad sexual	0.2195	0.0381	33.13	<0.0001
Planes universitarios	-0.1198	0.0515	5.41	0.0201
Auto eficacia anticonceptiva	-0.1939	0.0852	5.17	0.023
Uso consistente de anticonceptivos	-0.3106	0.0812	14.62	<0.000

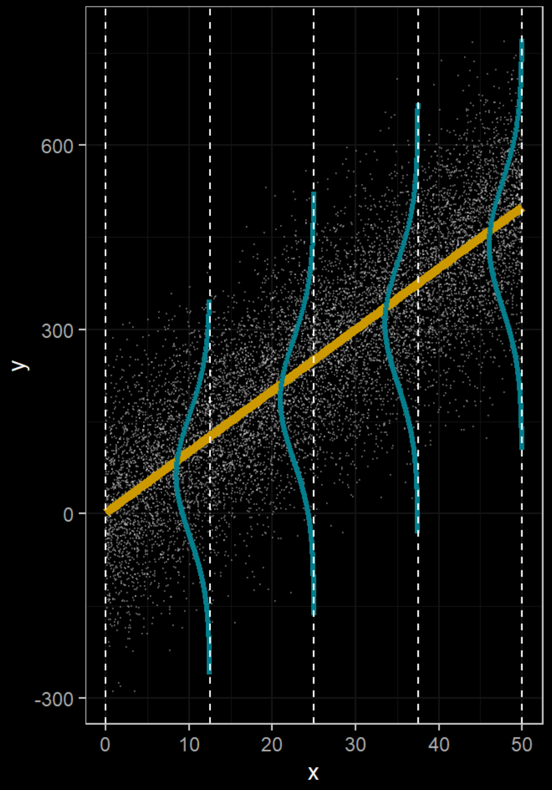
Modelos lineales generalizados: motivación

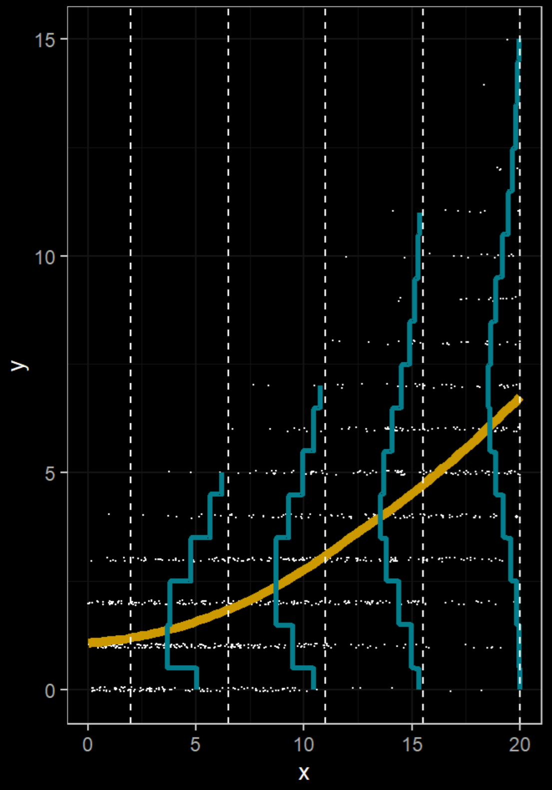
Estimación de MV de un modelo de Poisson:

	Parámetro estimado	exp(β)	ΔY (%)
Edad actual	0.2278	1.26	25.6
Afroamericana	0.6958	2.01	100.5
Hispana	0.1498	1.16	16.2
Asiática	0.5333	1.70	70.5
Nunca casada	-1.1035	0.33	-66.8
Años de actividad sexual	0.2195	1.25	24.5
Planes universitarios	-0.1198	0.89	-11.3
Auto eficacia anticonceptiva	-0.1939	0.82	-17.6
Uso consistente de anticonceptivos	-0.3106	0.73	-26.7

Modelos lineales generalizados: motivación







Modelos lineales generalizados: definición

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

μ_i es la media de la variable de respuesta Y_i.

g es la función de enlace.

η_i es el predictor lineal.

$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ son los coeficientes del modelo.

$X_{i1}, X_{i2}, \ldots, X_{ip}$ son las variables predictoras.

Modelos lineales generalizados: definición

Variable de respuesta (Y):

- Puede seguir diferentes distribuciones de probabilidad (binomial, Poisson, normal, gamma, etc.).
- No se limita a ser continua y normalmente distribuida como en los modelos lineales tradicionales.

Función de enlace g(·):

- Relaciona la media de la variable de respuesta con la combinación lineal de las variables explicativas en X.

$$\mu_i = \mathbb{E}[Y_i]$$

$$g(\mathbb{E}[Y_i]) = \mathbf{x}_i^\top \boldsymbol{\beta}$$

Modelos lineales generalizados: ejemplos

Regresión Lineal (Distribución Normal, Función de Enlace Identidad):

- Variable de respuesta: continua.
- Modelo: $Y = \beta_0 + \beta_1 X + \epsilon$
- ϵ es el término de error normalmente distribuido.

Regresión Logística (Distribución Binomial, Función de Enlace Logit):

- Variable de respuesta: binaria.
- Modelo: $\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$
- π es la probabilidad de éxito.

Modelos lineales generalizados: ejemplos

Regresión de Poisson (Distribución de Poisson, Función de Enlace Log):

- Variable de respuesta: conteos.
- Modelo: $\log(\lambda) = \beta_0 + \beta_1 X$
- λ es la tasa de conteo.

Regresión Binomial Negativa (Distribución Binomial Negativa, Función de Enlace Log)

- Variable de respuesta: conteos con sobredispersión.
- Modelo: $\log(\lambda) = \beta_0 + \beta_1 X$

Funciones de enlace

Función de enlace identidad

- **Modelo:** $Y = \beta_0 + \beta_1 X + \epsilon$
- **Función de Enlace:** $g(\mu) = \mu$
- **Inversa:** $g^{-1}(\eta) = \eta$
- **Tipo de Variable:** Continua (normalmente distribuida)

$$\mu = \beta_0 + \beta_1 X$$

Funciones de enlace

Función de enlace inversa

- **Modelo:** $\mu^{-1} = \beta_0 + \beta_1 X$
- **Función de Enlace:** $g(\mu) = \frac{1}{\mu}$
- **Inversa:** $g^{-1}(\eta) = \frac{1}{\eta}$
- **Tipo de Variable:** Continua y positiva

$$\mu = \frac{1}{\beta_0+\beta_1 X}$$

Funciones de enlace

Función de enlace potencia

- **Modelo:** $\mu^k = \beta_0 + \beta_1 X$, donde k es una constante
- **Función de Enlace:** $g(\mu) = \mu^k$
- **Inversa:** $g^{-1}(\eta) = \eta^{1/k}$
- **Tipo de Variable:** Continua y positiva

$$\mu = (\beta_0 + \beta_1 X)^{1/k}$$

Funciones de enlace

Función de enlace logit

- **Modelo:** $\operatorname{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X$
- **Función de Enlace:** $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$
- **Inversa:** $g^{-1}(\eta) = \frac{e^{\eta}}{1+e^{\eta}}$
- **Tipo de Variable:** Binaria (0 o 1)

$$\pi = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Funciones de enlace

Función de enlace gompit

- **Modelo:** $\log(-\log(\pi)) = \beta_0 + \beta_1 X$
- **Función de Enlace:** $g(\pi) = \log(-\log(\pi))$
- **Inversa:** $g^{-1}(\eta) = e^{-e^{\eta}}$
- **Tipo de Variable:** Binaria (0 o 1)

$$\pi = e^{-e^{\beta_0+\beta_1 X}}$$

Funciones de enlace

Función de enlace probit

- **Modelo:** $\Phi^{-1}(\pi) = \beta_0 + \beta_1 X$
- **Función de Enlace:** $g(\pi) = \Phi^{-1}(\pi)$, donde Φ^{-1} es la inversa de la función de distribución normal acumulativa.
- **Inversa:** $g^{-1}(\eta) = \Phi(\eta)$, donde Φ es la función de distribución normal acumulativa.
- **Tipo de Variable:** Binaria (0 o 1)

$$\pi = \Phi(\beta_0 + \beta_1 X)$$

Funciones de enlace

Función de enlace log

- **Modelo:** $\log(\lambda) = \beta_0 + \beta_1 X$
- **Función de Enlace:** $g(\lambda) = \log(\lambda)$
- **Inversa:** $g^{-1}(\eta) = e^{\eta}$
- **Tipo de Variable:** Conteos (números enteros no negativos)

$$\lambda = e^{\beta_0 + \beta_1 X}$$

Funciones de enlace

Función de enlace log-log (doble logaritmica)

- **Modelo**: $\log(\log(\mu)) = \beta_0 + \beta_1 X$
- **Función de Enlace:** $g(\mu) = \log(\log(\mu))$
- **Inversa:** $g^{-1}(\eta) = e^{e^{\eta}}$
- **Tipo de Variable:** Continua y positiva

$$\mu = e^{e^{\beta_0 + \beta_1 X}}$$

Funciones de enlace

Función de enlace arcseno de raíz cuadrática

- **Modelo:** $\arcsin(\sqrt{\mu}) = \beta_0 + \beta_1 X$
- **Función de Enlace:** $g(\mu) = \arcsin(\sqrt{\mu})$
- **Inversa:** $g^{-1}(\eta) = (\sin(\eta))^2$
- **Tipo de Variable:** Proporciones (valores entre 0 y 1)

$$\mu = (\sin(\beta_0 + \beta_1 X))^2$$

Funciones de enlace

Función de enlace Box-Cox

- **Modelo:** $\frac{\mu^\lambda - 1}{\lambda} = \beta_0 + \beta_1 X$, donde λ es un parámetro de transformación
- **Función de Enlace:** $g(\mu) = \frac{\mu^\lambda - 1}{\lambda}$
- **Inversa:** $g^{-1}(\eta) = (\lambda\eta + 1)^{1/\lambda}$
- **Tipo de Variable:** Continua y positiva

$$\mu = (\lambda(\beta_0 + \beta_1 X) + 1)^{1/\lambda}$$

Métodos de estimación: máxima verosimilitud

Los modelos lineales generalizados se estiman usualmente con máxima verosimilitud:

$$\{(y_i, \mathbf{x}_i)\}_{i=1}^{n}$$

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta})$$

$$\ell(\boldsymbol{\beta}) = \log L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log f(y_i \mid \mathbf{x}_i; \boldsymbol{\beta})$$

$$\hat{\boldsymbol{\beta}} = \arg\max_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$$

Métodos de estimación: máxima verosimilitud

Los estimadores máximo verosímiles son consistentes, asintoticamente eficientes, tienen un distribución asintótica normal e insesgadez asintótica:

$$\hat{\boldsymbol{\beta}}_{MLE} \xrightarrow{P} \boldsymbol{\beta} \qquad \lim_{n\to\infty} P(|\hat{\boldsymbol{\beta}}_{MLE} - \boldsymbol{\beta}| \geq \epsilon) = 0$$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{MLE} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, I(\boldsymbol{\beta})^{-1})$$

$$\mathbb{E}[\hat{\boldsymbol{\beta}}_{MLE}] \approx \boldsymbol{\beta} \qquad I(\boldsymbol{\beta}) = \mathbb{E}\left[-\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}\right]$$

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{MLE}) \approx \tfrac{1}{n} I(\boldsymbol{\beta})^{-1}$$

Métodos de estimación: Quasi-máxima verosimilitud

Quasi-máxima verosimilitud:

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{(y_i - \mu_i)^2}{2V(\mu_i)} + \int_{\mu_i}^{y_i} \frac{y-t}{V(t)}\, dt \right)$$

$$\frac{\partial Q(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{(y_i - \mu_i) \frac{\partial \mu_i}{\partial \beta_j}}{V(\mu_i)}$$

- No se basa en una función de verosimilitud con una distribución específica
- La derivada se utiliza para encontrar los estimadores, con métodos numéricos
- Los estimadores son consistentes y asintóticamente eficientes.

Métodos de estimación: Quasi-máxima verosimilitud

Las iteraciones del scoring the Fisher se utilizan para encontrar estimadores de quasi-máxima verosimilitud en algunos modelos estadísticos:

$$\boldsymbol{\beta}^{(0)}$$

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + I(\boldsymbol{\beta}^{(t)})^{-1} U(\boldsymbol{\beta}^{(t)})$$

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{y_i - \mu_i}{\mu_i} \mathbf{x}_i$$

$$I(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\mathbf{x}_i \mathbf{x}_i^\top}{\mu_i}$$

$$I(\boldsymbol{\beta}) = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top}\right] \qquad U(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$$

$$\mu_i = f(\mathbf{x}_i^\top \boldsymbol{\beta})$$

Métodos de estimación: Método de los momentos

Método de los momentos:

$$\mathbb{E}[Y] = \mu(\beta)$$
$$\mathbb{E}[(Y - \mu(\beta))^2] = \sigma^2(\beta)$$

$$\hat{\mu} = \tfrac{1}{n}\sum_{i=1}^{n} y_i$$
$$\hat{\sigma}^2 = \tfrac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu})^2$$

$$\hat{\mu} = \mu(\beta)$$
$$\hat{\sigma}^2 = \sigma^2(\beta)$$

Métodos de estimación: Método de los momentos

Método de los momentos: ejemplos.

Modelo lineal:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

$$\mathbb{E}[Y] = \beta_0 + \beta_1 \mathbb{E}[X]$$

$$\mathrm{Var}(Y) = \mathrm{Var}(\epsilon)$$

Regresión de Poisson:

$$\log(\lambda_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$$

$$\mathbb{E}[Y_i] = \lambda_i$$

$$\mathrm{Var}(Y_i) = \lambda_i$$

Métodos de estimación: propiedades de los estimadores

Los estimadores del método de los momentos son consistentes, asintoticamente insesgados, y tienen un distribución asintótica normal, pero en muestras pequeñas la varianza es mayor que la de los estimadores máximo verosímiles:

$$\hat{\boldsymbol{\beta}}_{MME} \xrightarrow{P} \boldsymbol{\beta}$$

$$\mathbb{E}[\hat{\boldsymbol{\beta}}_{MME}] \approx \boldsymbol{\beta}$$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{MME} - \boldsymbol{\beta}) \xrightarrow{d} \mathcal{N}(0, V)$$

$$\mathrm{Var}(\hat{\beta}_{MME}) \approx \frac{1}{n}\left(\frac{\partial \mu}{\partial \beta}\right)^{-1} \Sigma \left(\frac{\partial \mu}{\partial \beta}\right)^{-1}$$

Laboratorio: Simulaciones de Monte Carlo

- **MLMLG_0301.R:** Métodos de estimación de MLGs.
- **MLMLG_0302.R:** Simulación de consistencia de los estimadores en un modelo de regresión de Poisson
- **MLMLG_0303.R:** Simulación de insesgamiento de los estimadores en un modelo de regresión de Poisson
- **MLMLG_0304.R:** Simulación de insesgamiento asintótico de los estimadores en un modelo de regresión de Poisson
- **MLMLG_0305.R:** Simulación de la distribución asintótica normal de los estimadores en un modelo de regresión de Poisson

¿A qué software estadístico le daremos más énfasis en el resto del módulo?

R (Posit)	52%
R (AnalyticFlow)	0%
Python	48%

Interpretación de los estimadores en MLG

La interpretación de los estimadores de β en los MLGs depende de la función de enlace y de la distribución de la variable de respuesta:

Modelo	Distribución y función de enlace	Interpretación de los estimadores β
Regresión normal (gaussiana)	Distribución normal, función de enlace identidad	Cambio unitario de X está relacionado con un cambio β en Y
Regresión logistica (logit)	Distribución binomial, función de enlace logit	Cambio unitario de X está relacionado con un cambio exp(β) en el odds ratio de Y
Regresión de Poisson	Distribución de Poisson, función de enlace logarítmica	Cambio unitario de X está relacionado con un cambio logarítmico en la tasa de conteo de Y

Medidas de ajuste en MLGs: Devianza y pseudo-R^2

- La devianza mide la bondad del ajuste de un modelo comparando el modelo ajustado con un modelo saturado. Cuanto menor sea la devianza, mejor es el ajuste del modelo a los datos observados.
- El pseudo-R2 basado en la devianza indica mejoras en el ajuste respecto al modelo nulo.

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

$$D(y,\bar{y}) = 2\left[\sum_{i=1}^{n} y_i \log\left(\frac{y_i}{\bar{y}}\right) - (y_i - \bar{y})\right]$$

$$D_{\text{residual},i} = 2\left[y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right]$$

$$R^2 = 1 - \frac{D(y,\hat{\mu})}{D(y,\bar{y})}$$

Medidas de ajuste en MLGs: Devianza

Devianza para un modelo de regresión lineal gaussiano, una regresión logística y una regresión de Poisson:

$$D(y, \hat{\mu}) = \sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$$

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

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

Medidas de ajuste en MLGs: Pseudo-R^2

Existen diferentes medidas de bondad de ajuste basadas en pseudo-R^2s: Cox y Snell, Nagelkerke, McFadden, Tjur, y de Efron. Los pseudo-R^2 de Cox y Snell, MacFadden y el pseudo-R^2 basado en la devianza pueden ser negativos.

$$R^2_{CS} = 1 - \left(\frac{L_0}{L_1}\right)^{2/n}$$

$$R^2_N = \frac{R^2_{CS}}{1 - L_0^{2/n}}$$

$$R^2_{McF} = 1 - \frac{\ln(L_1)}{\ln(L_0)}$$

$$R^2_{Tjur} = \mathrm{Promedio}(\hat{y}_1) - \mathrm{Promedio}(\hat{y}_0)$$

$$R^2_{Efron} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Regresión Gaussiana

Distribución normal, función de enlace identidad:

Modelo: $Y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i$, donde $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

Variable de Respuesta: Y_i es continua.

Función de Enlace: Identidad, $g(\mu_i) = \mu_i$

Interpretación de los Coeficientes:

- Cada coeficiente β_j representa el cambio esperado en Y_i por un cambio unitario en la variable predictora x_j, manteniendo las demás variables constantes.

Regresión de Poisson

Distribución de Poisson, función de enlace logarítmica:

Modelo: $\log(\lambda_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$

Variable de Respuesta: Y_i es un conteo (número de eventos).

Interpretación de los Coeficientes:

- Cada coeficiente β_j representa el cambio en el logaritmo de la tasa de conteo ($\log \lambda_i$) por un cambio unitario en x_j.
- $\exp(\beta_j)$ es el factor multiplicativo por el cual se incrementa la tasa de conteo λ_i por un cambio unitario en x_j.

Regresión de Poisson

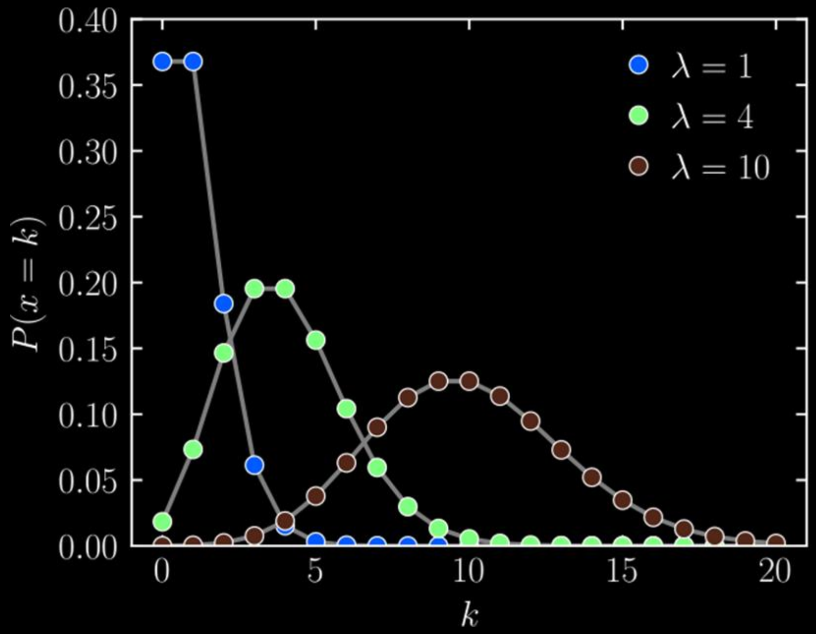
Distribución de Poisson:

- Describe el número de eventos en un intervalo fijo de tiempo o espacio.
- Función de probabilidad: $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$
- Esperanza: $\mathbb{E}[X] = \lambda$
- Varianza: $\mathrm{Var}(X) = \lambda$

Regresión de Poisson

Formalmente:

$$Y_i \sim \text{Poisson}(\lambda_i)$$
$$\log(\lambda_i) = \mathbf{x}_i^\top \boldsymbol{\beta}$$
$$\lambda_i = \exp(\mathbf{x}_i^\top \boldsymbol{\beta})$$
$$g(\lambda_i) = \log(\lambda_i)$$



Regresión de Poisson: Devianza

Devianza y devianza nula:

$$D = 2\left(\log L_{\text{sat}} - \log L\right) =$$
$$2\sum_{i=1}^{n}\left[y_i \log\left(\frac{y_i}{p_i}\right) + (1 - y_i)\log\left(\frac{1-y_i}{1-p_i}\right)\right]$$

$$d_i = 2\left[y_i \log\left(\frac{y_i}{p_i}\right) + (1 - y_i)\log\left(\frac{1-y_i}{1-p_i}\right)\right]$$

Regresión de Poisson: Interpretación de coeficientes

La fórmula general para un modelo Poisson es:

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi}$$

donde:

- λ_i es la tasa esperada de eventos para la observación i.
- β_0 es el intercepto del modelo.
- β_j son los coeficientes asociados a las variables explicativas x_{ji}.

Regresión de Poisson: Interpretación de coeficientes

Los coeficientes en un modelo Poisson se interpretan en términos de la tasa de ocurrencia del evento:

- **Coeficiente positivo ($\beta_j > 0$)**: Un aumento en x_j está asociado con un incremento en la tasa de ocurrencia del evento.
- **Coeficiente negativo ($\beta_j < 0$)**: Un aumento en x_j está asociado con una disminución en la tasa de ocurrencia del evento.
- **Coeficiente cero ($\beta_j = 0$)**: No hay asociación entre x_j y la tasa de ocurrencia del evento.

Regresión de Poisson: Interpretación de coeficientes

- Para interpretar el impacto en la tasa esperada (λ), se utiliza la exponencial de los coeficientes para obtener el cambio en la tasa esperada de eventos por unidad de cambio en x.

$$e^{\beta_j}$$

$$100 \times (e^{\beta} - 1)$$

- El cambio porcentual en la tasa esperada de eventos por unidad de cambio en la variable explicativa puede obtenerse con una fórmula adicional que resta 1 y multiplica por cien el exponencial del coeficiente estimado.
- El intercepto representa la tasa de ocurrencia esperada del evento cuando todas las variables explicativas son cero.

Regresión de Poisson: Sobredispersión

La sobredispersión ocurre cuando la variabilidad en los datos es mayor que la esperada según el modelo de Poisson:

- Supuesto de un modelo de Poisson: la varianza es igual a la media.
- Si la varianza es significativamente mayor que la media, esto sugiere sobredispersión.

La sobredispersión puede afectar la validez de los resultados del modelo (intervalos de confianza y las pruebas de hipótesis):

- Los estimadores siguen siendo insesgados y consistentes
- Los estimadores ya no son eficientes

Regresión de Poisson: Sobredispersión

Estadígrafos y test LM de sobredispersión:

- El estadígrafo basado en los residuos de Pearson debe ser igual a 1 si no existe sobredispersión o subdispersión.
- La hipótesis nula del estadigrafo Chi2 es que no existe sobredispersión.

$$\hat{\phi} = \frac{\sum(\text{Pearson residuals})^2}{n-p}$$

$$\chi^2 = \frac{\left(\sum_{i=1}^{n} \mu_i^2 - n\bar{y}_i\right)^2}{2\sum_{i=1}^{n} \mu_i^2}$$

Regresión de Poisson: Soluciones a la sobredispersión

- Modificar el modelo: Agregar variables explicativas, interacciones, transformar las variables, ajustarla presencia de valores atípicos
- Errores Estándar Robustos (estimadores sandwich)
- Modelo Quasi-Poisson: El modelo quasi-Poisson ajusta la varianza permitiendo que sea una función lineal de la media, Var(Y) = $\phi\mu$, siendo ϕ el parámetro de sobredispersión.
- Modelo binomial negativo: Modeliza explícitamente la sobredispersión mediante un parámetro adicional θ

MLG Binomial Negativo

Modeliza explícitamente la sobredispersión mediante un parámetro adicional θ que captura la variabilidad extra:

$$\mathbb{E}(Y) = \mu$$

$$\mathrm{Var}(Y) = \mu + \frac{\mu^2}{\theta}$$

$$L(\beta, \theta | y) = \prod_{i=1}^{n} \frac{\Gamma(y_i + \theta^{-1})}{\Gamma(\theta^{-1}) y_i!} \left(\frac{\theta^{-1}}{\theta^{-1} + \mu_i} \right)^{\theta^{-1}} \left(\frac{\mu_i}{\theta^{-1} + \mu_i} \right)^{y_i}$$

$$\mu_i = \exp(X_i \beta)$$

$$\frac{\partial \log L}{\partial \beta} = \sum_{i=1}^{n} X_i \left(y_i - \mu_i \frac{\theta^{-1} + y_i}{\theta^{-1} + \mu_i} \right) = 0$$

Laboratorio: Ajuste de un modelo de regresión de Poisson y Binomial Negativo en R, Python, y con AI (Gemini Advanced)

PROMPT: `estima un modelo de regresión de Poisson para predecir el número de crímenes (nv) con los matriculados y la región, interpreta el ajuste del modelo y los coeficientes del modelo`