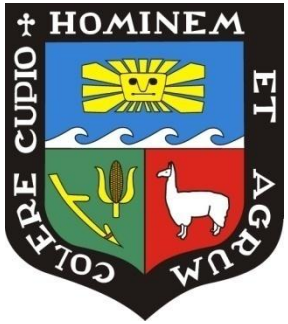


UNIVERSIDAD NACIONAL AGRARIA LA MOLINA
Dpto. de Estadística e Informática

Capítulo I.
Fundamentos de los Modelos Lineales Generalizados

Clase 1. Fundamentos de los MLG

Plan de aprendizaje

Inicio

- Motivación
- Logros
- Saberes previos

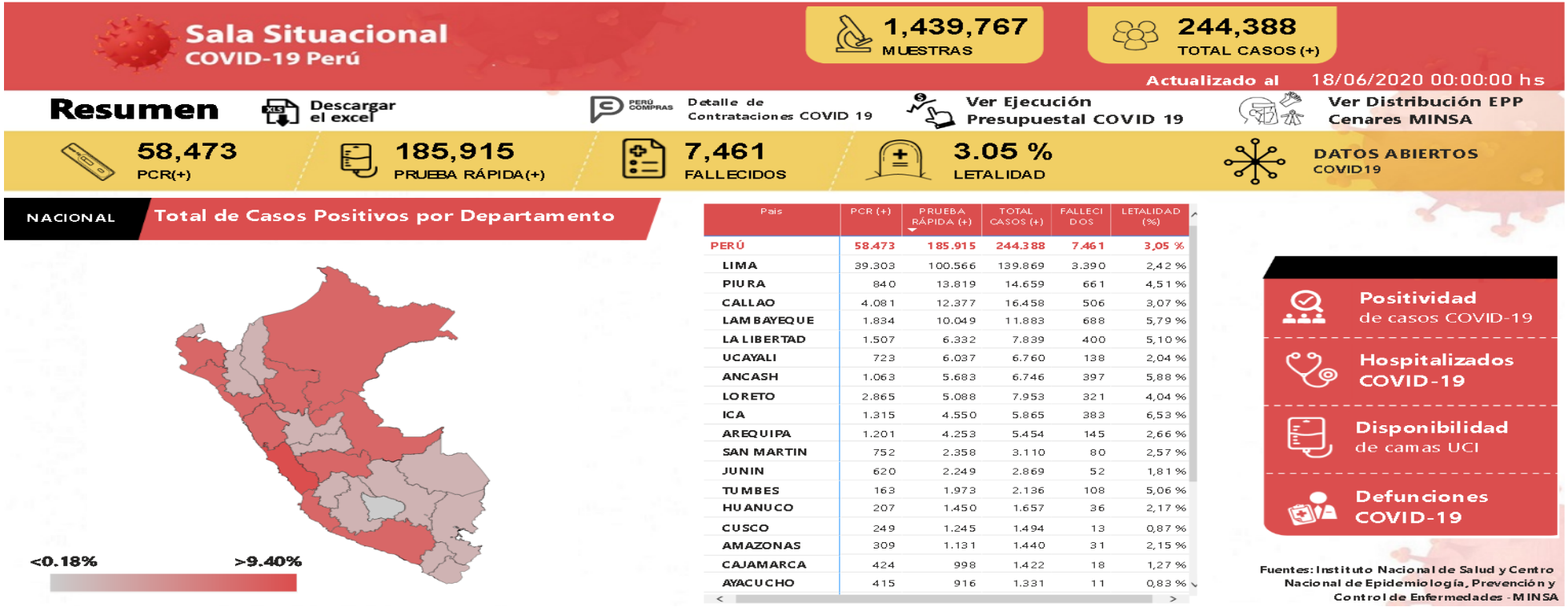
Desarrollo

- Fundamentos de los modelos lineales generalizados
- Ejercicios resueltos

Cierre

- Ejercicios propuestos
- Tarea

Motivación:



Pais	PCR (+)	PRUEBA RÁPIDA (+)	TOTAL CASOS (+)	FALLECIDOS	LETALIDAD (%)
PERÚ	58.473	185.915	244.388	7.461	3,05 %
LIMA	39.303	100.566	139.869	3.390	2,42 %
PIURA	840	13.819	14.659	661	4,51 %
CALLAO	4.081	12.377	16.458	506	3,07 %
LAMBAYEQUE	1.834	10.049	11.883	688	5,79 %
LA LIBERTAD	1.507	6.332	7.839	400	5,10 %
UCAYALI	723	6.037	6.760	138	2,04 %
ANCASH	1.063	5.683	6.746	397	5,88 %
LORETO	2.865	5.088	7.953	321	4,04 %
ICA	1.315	4.550	5.865	383	6,53 %
AREQUIPA	1.201	4.253	5.454	145	2,66 %
SAN MARTIN	752	2.358	3.110	80	2,57 %
JUNIN	620	2.249	2.869	52	1,81 %
TUMBES	163	1.973	2.136	108	5,06 %
HUANUCO	207	1.450	1.657	36	2,17 %
CUSCO	249	1.245	1.494	13	0,87 %
AMAZONAS	309	1.131	1.440	31	2,15 %
CAJAMARCA	424	998	1.422	18	1,27 %
AYACUCHO	415	916	1.331	11	0,83 %

¿Cuál será el tipo de variable dependiente, si se desea predecir?

- El número de personas contagiadas en las regiones.
- Cuando una persona tendrá el Covid 19 (Positivo o Negativo).
- La región con mayor probabilidad de tener personas fallecidas.
- El tiempo (en días) en que a una persona se le detecta el primer síntoma del Covid 19.

Logros:

Al término de la sesión, el estudiante estará en capacidad de:

- Comprender y aplicar los fundamentos de los modelos de regresión lineal en el contexto de los modelos generalizados.
- Resolver ejercicios sobre los fundamentos de los modelos lineales generalizados.
- Resolver ejercicios propuestos.

Saberes previos:

- ¿Qué son los modelos de regresión lineal?. ¿Cuál es la distribución de la variable dependiente?
- ¿Cuáles son las distribuciones discretas y continuas?
- ¿Cómo se formulan los modelos lineales generalizados en el contexto del mundo real?

Fundamentos de los modelos lineales generalizados

1. Introducción
2. Distribución de probabilidades
3. Clasificación de las variables
4. Creación de variables
5. Distribución de la familia exponencial
6. Componentes de un modelo lineal generalizado
7. Principales funciones de enlace
8. Proceso de modelamiento para un MLG
9. Ejercicios propuestos
10. Referencias bibliográficas

1. Introducción

Los modelos lineales generalizados (MLG) son una extensión del Modelo Lineal General y fueron propuestos por Nelder y Wedderburn (1972). Los MLG unifican a los modelos de regresión lineales donde la variable dependientes no necesariamente tienen una distribución Normal. En los MLG la variable respuesta puede ser tipo numérica (cuantitativa) o categórica (cualitativa), por la variable dependiente pueden asumir otras distribuciones tales como; la Binomial, la Poisson, la Logística, la Gamma, etc. En los MLG la distribución que asume la variable dependiente debe pertenecer a la familia exponencial.

En los MLG, todo el desarrollo de la inferencia estadística es una extensión del modelo que se basa en la distribución de la familia exponencial. En esta parte se presenta los principales conceptos utilizados en el modelamiento de los MLG.

1. Introducción

Características:

- La formulación de un MLG se basan en tres componentes: ***Componente aleatorio*** (variable respuesta), ***Componente sistemático*** (predictor lineal) y ***Función de enlace.***
- En los MLG, el supuesto de homocedasticidad del componente aleatorio no necesariamente se debe de cumplir.
- En los MLG se asumen que las observaciones son independientes (no correlacionadas). Los MLG asumen un único término de error (componente aleatorio).
- Los MLG permiten un modelamiento más complejo entre la variable respuesta y las variables explicativas; a través de la selección de la función de enlace que permite relacionar el predictor lineal con la media de la variable respuesta.

2. Distribución de probabilidades

Para una variable aleatoria Y, se define a $f(y;\theta)$ como la función de probabilidades, si la VA es discreta y función de densidad, si es continua; siendo θ el parámetro de la distribución.

Vector aleatorio. Se define a $Y'=(Y_1, Y_2, \ldots, Y_n)$ un vector aleatorio con n elementos, cuyos componentes son independientes e idénticamente distribuidas con $f(y_i;\theta)$ y siendo el vector de valores observados definido por $y'=(y_1, y_2, \ldots, y_n)$.

Distribución conjunta. Sea el vector aleatorio Y' con función de distribución $f(y;\theta)$ y para cada componente $f(y_i;\theta)$, entonces su distribución conjunta se define por:

$$f(y_1, y_2, \ldots, y_n; \theta) = f(\underline{y}; \theta) = \prod_{i=1}^{n} f(y_i; \theta)$$

2. Distribución de probabilidades

Función de Verosimilitud. La función de verosimilitud para una distribución f(y;θ), se define como la función L(θ;y)=f(θ;y). Se nota que L(θ;y) es algebraicamente igual a f(y;θ), dónde el parámetro θ es una variable aleatoria.

Función Log-Verosimilitud. Para la función de verosimilitud L(θ;y), se define la función de log-verosimilitud l(θ;y) como el logaritmo de L(θ;y). Esto es:

$$l(\theta; y) = \log L(\theta; y)$$

La función log-verosimilitud, se usa para hallar la estimación de parámetros por el método de máxima verosimilitud. El estimador se obtiene diferenciando la función log-verosimilitud con respecto a cada parámetro.

Ejemplo 1. Para una distribución de Poisson.

Sea el vector aleatorio $Y'=(Y_1, Y_2, \ldots, Y_n)$ de n variables aleatorias independientes e idénticamente distribuidas cada una como una Poisson. Entonces la distribución de probabilidad para un valor y_i y con parámetro θ, se expresa por:

no va el subíndice i en el teta, porque las variables aleatorias provienen de la misma distribución

$$f(y_i;\theta) = \frac{e^{-\theta}\theta^{y_i}}{y_i!}, \qquad y_i = 0,1,2, \ldots$$

La distribución conjunta será:

$$f(y_1,\ldots,y_n;\theta) = f(\underline{y};\theta) = \prod_{i=1}^{n} f(y_i;\theta) = \frac{e^{-\theta}\theta^{y_1}}{y_1!} x \ldots x \frac{e^{-\theta}\theta^{y_n}}{y_n!} = \frac{e^{-n\theta}\theta^{\sum y_i}}{\prod y_i!}$$

La función de log-verosimilitud será:

$$l(\theta; y_1, y_2, \ldots, y_n) = \log L(\theta; y_1, y_2, \ldots, y_n) = \left(\sum y_i\right)\log\theta - n\theta - \sum \log y_i!$$

3. Clasificación de las variables

En los MLG las variables dependientes y predictoras pueden ser clasificadas:

Variables dependientes

Variables cuantitativas. Toman un valor numérico. Pueden ser:
Continuas. Los valores numéricos pertenecen a un rango infinito no numerable. Están asociados a alguna unidad de medida. Son valores reales (intervalo o razón).
Discretas. Los valores numéricos pertenecen a un rango finito o infinito numerable. Están asociados a conteos. Son valores enteros.
Variables cualitativas. Toman un valor no numérico (categoría). Pueden ser:
Binarias. Presentan sólo dos categorías.
Politómicas. Tienen más de dos categorías. Pueden ser:

- **Nominales.** Las categorías no tienen un orden de importancia.
- **Ordinales.** Las categorías tienen un orden de importancia. Pueden obtenerse también por la discretización de variables cuantitativas.

Generalmente, se deben crear para una variable que tienen "a" categorías "a-1" variables binarias.

Variables predictoras

Covariable. Son variables cuantitativas.
Factor. Son variables cualitativas, cuyas distintas categorías se denominan niveles.

4. Creación de variables

La discretización. Creando por la transformación de una variable cuantitativa en cualitativa ordinal.

- **Discretización con igual longitud cada intervalo.** Para k intervalos, se calcula el tamaño de los intervalos.

  TI = (Max. –Min.)/k

Los límites inferiores (cerrado) y superiores (abiertos) de cada intervalos:

LI_1=Mínimo $LS_1=LI_1+TI$

$LI_2=LS_1$ $LS_2=LI_2+TI$

. . . $LS_k=Li_{k-1}+TI$

- **Discretización con igual número de observaciones por intervalo.** Se determinan los límites de cada intervalos.

  Percentiles. Pq

  Cuartiles. Se divide en 4 partes, con 25% cada una.

  Terciles. Se divide en 3 partes, con 33.3% cada una.

- **Discretización con intervalos arbitrarios.** Diferentes tamaños y porcentajes de observaciones en los intervalos.

4. Creación de variables

Ejemplo 2. Se tiene las edades de una muestra de 40 clientes. Discretizar con 4 intervalos con igual tamaño.

69	53	41	34
45	29	24	42
31	28	53	66
55	21	34	32
35	19	32	32
21	53	52	55
38	61	23	67
28	18	69	63
60	26	42	39
28	26	57	47

K =	4	
Max. =	69	
Min. =	18	
TI =	**12.8**	

Intervalos:	Frec. Absoluta	Frec. Relativa
[18 - 30.8 >	12	30.0
[30.8 - 43.6 >	12	30.0
[43.6 - 56.4 >	8	20.0
[56.4 - 69.2 >	8	20.0
	40	100.0

4. Creación de variables

Ejemplo 3. Discretizar con 4 intervalos con similar número de observaciones por intervalo. Usando los cuartiles: $Q_1=P_{25}$, $Q_2=P_{50}$, $Q_3=P_{75}$.

69	53	41	34
45	29	24	42
31	28	53	66
55	21	34	32
35	19	32	32
21	53	52	55
38	61	23	67
28	18	69	63
60	26	42	39
28	26	57	47

Q1 =	28.0	
Q2 =	38.5	
Q3 =	54.5	

Intervalos	Frec. Absoluta	Frec. Relativa
[18 - 28.0 >	8	20.0
[28.0 - 38.5 >	12	30.0
[38.5 - 54.5 >	10	25.0
[54.5 - 69.0 >	10	25.0
	40	100.0

- **Se observa que se obtienen intervalos más balanceados**

4. Creación de variables

La Binarización. Creando por una re categorización de una cualitativa.

En general, cuando una variable cualitativa presenta "a" categorías, se puede crear (a-1) variables binarias (dumming). Dónde una de las categorías se considera como referencial.

Ejemplo 4. Para una muestra de agricultores se ha registrado las siguientes variables:

- Los destinos de la producción (consumo, venta, trueque). Entonces se debe de crear **2** nuevas variables binarias (dummy):

Var. Cualitativa: Destino de la producción.	Var. Cualitativas: Consumo	Venta
1 = Consumo	1	0
2 = Venta	0	1
3 = Trueque	0	0

Trueque=Categoría referencial

4. Creación de variables

- Los gastos (soles) semanales en fertilización. Hacer una discretización con 4 intervalos con igual tamaño.

Var. Cuantitativa continua: Gastos	Var. Cualitativa ordinal: Rango de gastos
305	R1: [100 a 250 >
260	R2: [251 a 400 >
180	R3: [401 a 550 >
. . .	R4: [551 a 700]

Se debe de crear 3 nuevas variables binarias (dummy):

Var. cualitativa: Rango de gastos	**Var. binarias: R1**	**R2**	**R3**
R1: [100 a 250 >	1	0	0
R2: [251 a 400 >	0	1	0
R3: [401 a 550 >	0	0	1

Categoría referencial: R4: [551 a 700 >

5. Distribución de la familia exponencial

Sea Y una variable aleatoria y f(y;θ) su función de distribución de probabilidades, entonces se dice que su distribución pertenece a la familia exponencial con un único parámetro, si puede ser expresada de la forma siguiente:

$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$$

Dónde:
θ Es el parámetro de interés
Las a(y), b(θ), c(θ) y d(y) son funciones conocidas.

Si a(y)=y, la distribución toma la forma canónica y b(θ) es el parámetro natural. Si existen otros parámetros, estos son considerados como parámetros de ruido (constantes).

Propiedades distribucionales de la familia exponencial.

$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$$

La media y varianza de a(y) son:

$$E[a(y)] = -\frac{c'(\theta)}{b'(\theta)} \quad \text{y} \quad Var[a(y)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}$$

Función score. Se define como la variable aleatoria U y que depende de y. Se obtiene por la derivada de la función log-verosimilitud con respecto al parámetro θ.

$$U = U(\theta;y) = \frac{\partial l(\theta;y)}{\partial \theta} = a(y)b'(\theta) + c'(\theta)$$

La estadística score, se usada para el proceso de inferencia de los parámetros en los modelos lineales generalizados. La función score permite maximizar la función log-verosimilitud.

La media de la función score. La media de la VA score U**,** se define por:

$$E(U) = b'(\theta)E(a(Y)) + c'(\theta)$$

$$\text{Pero como: } E[a(y)] = -\frac{c'(\theta)}{b'(\theta)}, \quad \text{reemplazando se cumple: } E(U) = 0$$

La covariancia de la función score. Es conocida como la ***matriz de información.*** Es la matriz de varianzas-covarianzas de la variable aleatoria score.

$$U = U(\theta; y) = \frac{\partial l(\theta; y)}{\partial \theta} = a(y)b'(\theta) + c'(\theta)$$

Además se tiene:

$$U' = \frac{\partial U}{\partial \theta} = a(y)b''(\theta) + c''(\theta)$$

$$\text{Entonces: } E(U') = b''(\theta)E(a(y)) + c''(\theta)$$

$$= b''(\theta)\left[-\frac{c'(\theta)}{b'(\theta)}\right] + c''(\theta) = -\text{Cov}(U) = -\mathfrak{I}$$

Ejemplo 5. Demostrar que las siguientes distribuciones pertenecen a la familia exponencial.

a. La distribución Poisson.

$$Se\ tiene: \quad f(y;\mu) = \frac{e^{-\mu}\mu^{y}}{y!} \qquad y = 0,1,2,\ldots$$

$$Haciendo\ \mu = \theta\ y\ aplicando\ Ln: \quad f(y,\theta) = \exp\left[Ln\left(\frac{e^{-\theta}\theta^{y}}{y!} \right) \right]$$

$$f(y,\theta) = \exp\left[y\ln(\theta) - \theta + (-\ln y!) \right]$$

$$Donde: a(y) = y,\ b(\theta) = \ln\theta,\ c(\theta) = -\theta,\ d(y) = -\ln y!$$

Entonces la distribución de Poisson pertenece a la Familia exponencial (tiene la forma canónica).

La función log-verosimilitud será: $l(\theta, y) = y\ln\theta - \theta - \ln y!$

b. La distribución Binomial.

$$Se\ tiene: \quad f(y,\pi) = C_y^n\ \pi^y\ (1-\pi)^{n-y} \qquad y = 0,1,2,...,\ n$$

$$Haciendo \quad \pi = \theta \ \ y\ aplicando \quad Ln: \quad f(y,\theta) = \exp\left[Ln\left(C_y^n\ \theta^y\ (1-\theta)^{n-y}\right)\right]$$

$$f(y,\theta) = \exp\left[\, y\ln\theta + (n-y)\ln(1-\theta) + \ln C_y^n \,\right]$$

$$= \exp\left[\, y(\ln\theta - \ln(1-\theta)) + n\ln(1-\theta) + \ln C_y^n \,\right]$$

$$= \exp\left[\, y\ln\left(\frac{\theta}{1-\theta}\right) + n\ln(1-\theta) + \ln C_y^n \,\right]$$

$$Donde: a(y) = y, \quad b(\theta) = \ln\left(\frac{\theta}{1-\theta}\right), \quad c(\theta) = n\ln(1-\theta), \quad d(y) = \ln C_y^n$$

Entonces la distribución Binomial pertenece a la Familia exponencial (tiene la forma canónica).

La función log-verosimilitud: $l(\theta;y) = y\ln\left(\frac{\theta}{1-\theta}\right) + n\ln(1-\theta) + \ln C_y^n$

c. La distribución Normal.

$$Se\ tiene: \quad f(y;\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y-\mu)^2}{2\sigma^2}}; \quad -\infty<x<\infty, \quad -\infty<\mu<\infty, \quad \sigma^2>0$$

$$Haciendo\ \mu=\theta,\ \sigma^2=\sigma_0^2\ y\ aplicando\ Ln:$$

$$f(y;\theta)=\exp\left[\ln\left(\frac{1}{\sqrt{2\pi\sigma_0^2}}e^{-\frac{(y-\theta)^2}{2\sigma_0^2}}\right)\right]=\exp\left[\ln\left(e^{-\frac{(y-\theta)^2}{2\sigma_0^2}}\right)+\ln\left((2\pi\sigma_0^2)^{-\frac{1}{2}}\right)\right]$$

$$f(y,\theta)=\exp\left[-\frac{y^2}{2\sigma_0^2}+\frac{y\theta}{\sigma_0^2}-\frac{\theta^2}{2\sigma_0^2}-\frac{1}{2}\ln(2\pi\sigma_0^2)\right]=\exp\left[y\frac{\theta}{\sigma_0^2}-\frac{\theta^2}{2\sigma_0^2}-\frac{y^2}{2\sigma_0^2}-\frac{1}{2}\ln(2\pi\sigma_0^2)\right]$$

$$Donde: a(y)=y, \quad b(\theta)=\frac{\theta}{\sigma_0^2}, \quad c(\theta)=-\frac{\theta^2}{2\sigma_0^2}\ y \quad d(y)=-\frac{y^2}{2\sigma_0^2}-\frac{1}{2}\ln(2\pi\sigma_0^2)$$

Entonces la distribución Normal pertenece a la Familia exponencial (tiene la forma canónica).

La función log-verosimilitud: $l(\theta;y)=y\frac{\theta}{\sigma_0^2}-\frac{\theta^2}{2\sigma_0^2}-\frac{y^2}{2\sigma_0^2}-\frac{1}{2}\ln(2\pi\sigma_0^2)$

Ejercicio 1.

Distribución muestral de la estadística score para una variable respuesta distribuida como una Binomial.

Y~ Binomial(n,π). La distribución de la familia exponencial:

$$f(y,\pi)=\exp\left[\, y\ln\pi - y\ln(1-\pi) + n\ln(1-\pi) + \ln C_y^n \,\right]$$

La función log-verosimilitud: $l(y,\pi) = y\ln\pi - (n-y)\ln(1-\pi) + \ln C_y^n$

La estadística score: $U = \dfrac{\partial l(y,\pi)}{\partial \pi} = \dfrac{y}{\pi} - \dfrac{n-y}{1-\pi} = \dfrac{y-n\pi}{\pi(1-\pi)}$

Pero se sabe que : $E(y) = n\pi$ y $V(y) = n\pi(1-\pi)$

Entonces se tiene : $E(U) = E\left(\dfrac{y-n\pi}{\pi(1-\pi)}\right) = 0$

$$V(U) = \Im = V\left(\frac{y-n\pi}{\pi(1-\pi)}\right) = \frac{1}{\pi^2(1-\pi)^2}V(y) = \frac{n}{\pi(1-\pi)}$$

6. Componentes de un MLG

Sea $Y'=(Y_1, Y_2, \ldots, Y_n)$ un vector aleatorio con n elementos independientes e idénticamente distribuidas y con vector de observaciones $y'=(y_1, y_2, \ldots, y_n)$ y con vector de media definida por . Sea la matriz de datos o diseño X de orden (nxp) de valores conocidos que representan las p variables predictivas, definidas como covariables si son cuantitativas o factores si son cualitativas. Se $\beta'=(\beta_1, \beta_2, \ldots, \beta_p)$ el vector de coeficientes del modelo, cuyos elementos son desconocidos y que deben ser estimados. Entonces, los componentes de un MLG son:

- **Componente aleatorio.** Es el vector aleatorio $Y'=(Y_1, Y_2, \ldots, Y_n)$ que representa la variable respuesta, cuyos elementos son independientes e idénticamente distribuidos; con su función de distribución perteneciente a la familia exponencial y con valor esperado o media $E[Y/X]=\mu$.

6. Componentes de un MLG

- **Componente sistemático.** Está compuesto por la combinación lineal de los coeficientes asociadas a cada variable predictora, es el vector conocido como el predictor lineal:

$$\eta_{nx1} = X_{nxp}\,\beta_{px1} \Rightarrow \eta = X'\underline{\beta} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ,,, + \beta_p X_p$$

- **Función de enlace.** Es una función conocida g(.), que permite relacionar el predictor lineal $\eta = X'\beta$ con el valor esperado de la variable respuesta $E[Y/X]=\mu$. Entonces la función de enlace toma la forma:

$$g(E(Y/X)) = g(\mu) = \eta = X'\underline{\beta}$$

7. Principales funciones de enlaces

La función de enlace establece la relación funcional entre la media del vector aleatorio con el predictor lineal. Entre las más usadas:

Distribución	Función de enlace	Expresión
Normal	Identidad	$\eta = g(\mu) = \mu$
Poisson	Logarítmica	$\eta = g(\mu) = log(\mu)$
Binomial	Logit (Función logística	$\eta = g(\underline{\mu}) = log\left(\frac{\mu_i}{1-\mu_i}\right)$
Normal estándar	Probit	$\eta = g(\underline{\mu}) = \Phi^{-1}(\mu_i)$
Gamma	Recíproca	$\eta = g(\mu) = \frac{1}{\mu}$
Inversa Gausiana	Recíproca^2	$\eta = g(\mu) = \frac{1}{\mu^2}$

Ejercicio 2.

Formule el MLG y determinar sus componentes para los siguientes casos.

a. Para modelar el número de reclamos por día de una aseguradora, asumiendo que es una variable aleatoria con distribución de Poisson, en función de la edad del cliente.
b. Predecir cuando una persona tendrá el Covid 19 (Positivo) o no (Negativo), en función de su edad y tiene diabetes y región con nivel de focalización (Alto, Medio o Bajo). Formule el MLG y describa los componentes en términos del problema.

Solución:

a. Para modelar el número de reclamos por día de una aseguradora, asumiendo que es una variable aleatoria con distribución de Poisson, en función de la edad del cliente.

Variable respuesta: Y=Número de reclamos
Variable predictora: X=Edad del cliente

Componente aleatorio: $Y_i \sim Poisson(\mu_i), con\ E[Y] = \mu$

Dónde: μ= Número esperado de reclamos por día

Componente sistemático (predictor lineal): $\eta_i = X_i^{'}\underline{\beta} = \beta_0 + \beta_1 x_i$

Función de enlace (logaritmo):

$$E(Y) = g(\mu) = Log(\mu_i) = \eta_i \Rightarrow Log(\mu_i) = \beta_0 + \beta_1 x_i$$

$$Se\ tiene: \mu_i = e^{\beta_0 + \beta_1 x_i}$$

b. Predecir a una persona infectada con el Covid 19 (Positivo) o no (Negativo), en función de su edad, tiene diabetes y región con nivel de focalización (Alto, Medio o Bajo). Formule el MLG y describa los componentes en términos del problema.

Variable respuesta: Y=Número de infectados
Variable predictora: X1=Edad del paciente
X2=Tiene diabetes (Si/No)
X3=Nivel de focalización alto
X4=Nivel de focalización medio
(Nivel bajo: categoría referencial)

Componente aleatorio: $Y_i \sim Bin(n, \pi_i),\ con\ \ E[Y_i] = n\pi_i$

Dónde: π=Probabilidad que una persona tenga Covid-19 (positivo)

Componente sistemático: $\eta_i = X_i^{'}\underline{\beta} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}$

Función de enlace:

$$\log itg(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i \Rightarrow \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}$$

8. Proceso de modelamiento para un MLG

El proceso puede constar de las siguientes etapas:

1. **Formulación del modelo.** Se describe la variable respuesta y las predictoras determinando su tipo y los componentes del MLG (el aleatorio, el sistemático y la función de enlace) en términos del problema.
2. **Estimación del modelo.** Se realiza el proceso de estimación de los coeficientes y las medidas de bondad de ajuste necesarias para la validación del modelo.
3. **Validación del modelo.** Aplicar el proceso de inferencia estadística calculando las medidas de bondad de ajuste y realizando pruebas de hipótesis para verificar la validez del modelo estimado.
4. **Uso del modelo.** Con el mejor modelo estimado, se interpreta la ecuación de regresión estimada, se calcula intervalos de confianza y valores predecidos en términos de los objetivos del estudio o investigación.

9. Ejercicios propuestos.

1. Demuestra que las siguientes distribuciones de probabilidades pertenecen a la familia exponencial.
   a. Distribución binomial negativa

$$f(y,\pi) = C_{r-1}^{y+r-1}\ \pi^{r}\ (1-\pi)^{y}$$

   b. La distribución Gamma y con parámetros de escala y de forma respectivamente

$$f(y,\theta) = \frac{y^{\varphi-1}\theta^{\varphi}e^{-y\theta}}{\Gamma(\varphi)}$$

   c. La distribución exponencial.

$$f(y;\theta) = \theta e^{-\theta y}, \quad y \geq 0,\ \theta > 0$$

2. Una agencia de turismo de la ciudad del Cusco está haciendo un estudio de mercado con la finalidad de mejorar los servicios a los turistas. Co este fin, se seleccionaron al azar 200 turistas que arribaron a la ciudad del curso en julio del 2019.

Se quiere predecir gasto ($) diario, teniendo información sobre la edad, calificación del servicio de hotel (excelente, bueno y malo) y sexo.

a. Defina la variable aleatoria y su distribución de probabilidades en términos del estudio.
b. Presente la distribución de probabilidades como una familia exponencial.
c. Formule el respectivo MLG y describa los componentes en términos del estudio.

10. Referencias bibliográficas.

1. Annette J. Dobson, (2002) An Introducction to Generalized Linear Models. Second Edition, Editorial Chapman & May.
2. Peter K. Dunn and Gordon K. Smyth (2018). Generalized Linear Models with examples in R. Springer Texts in Statistics.
3. McCullagh, Meter – Nelder, J.A, (1989) Generalized Linear Models. Second Edition. Editorial Chapman & Hall.