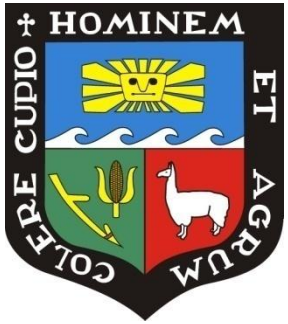

UNIVERSIDAD NACIONAL AGRARIA LA MOLINA
Dpto. de Estadística e Informática

Capítulo II.
Inferencia en los Modelos Lineales generalizados

Clase 3. Inferencia en los MLG

Plan de aprendizaje

Inicio

- Motivación
- Logros
- Saberes previos

Desarrollo

- Inferencia en los modelos lineales generalizados
- Pruebas y medidas de bondad de ajuste
- Ejercicios resueltos

Cierre

- Ejercicios propuestos
- Tarea

Motivación:

¿Cómo formular las hipótesis estadísticas en los sectores productivos?

- Hipótesis sobre las medidas estadísticas
- Hipótesis sobre los coeficientes de regresión
- Hipótesis sobre los modelos de regresión

Logros:

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

- Comprender y aplicar los métodos de inferencia en los modelos lineales generalizados.
- Resolver ejercicios sobre los métodos de inferencia en los modelos lineales generalizados aplicando el R.
- Resolver ejercicios propuestos.

Saberes previos:

- ¿Cómo se formula una hipótesis estadística?
- ¿Para qué sirve la formulación de hipótesis en los MLG?
- ¿Cómo se interpreta una hipótesis en los MLG?

Inferencia en los MLG

1. Introducción
2. Distribución muestrales
3. Pruebas y medidas de bondad de ajuste
4. Procedimiento para la prueba de hipótesis
5. Ejercicios propuestos
6. Referencias bibliográficas

1. Introducción

Los procedimientos clásicos de inferencia estadística sobre pruebas de bondad de ajuste y pruebas de hipótesis acerca de los parámetros pueden ser aplicados en forma similar en los MLG.

La aplicación del proceso de inferencia requiere el conocimiento de las distribuciones muestrales de las estadísticas score, deviance y otras y se basan generalmente en la distribución Chi-Cuadrado como prueba estadística. En el caso de que la variable respuesta sea Normal, la distribución muestral puede ser obtenida directamente, con otras distribuciones se aplicaran muestras asintóticas basadas en el Teorema del Limite Central.

La búsqueda del mejor modelo para explicar el comportamiento de la variable respuesta en función de un conjunto de variables explicativas, se basa en comparar modelos y seleccionar el más simple (principio de parsimonia).

LA BÚSQUEDA DEL MEJOR MODELO SE BASA EN COMPARAR MODELOS Y SELECCIONAR EL MÁS SIMPLE.

2. Distribuciones muestrales

Sea $Y'=(Y_1, Y_2, \ldots, Y_n)$ variables aleatorias independientes definidas en un modelo lineal generalizado con predictor lineal $x'\beta$, siendo $\beta'=(\beta_1, \beta_2, \ldots, \beta_p)$ el vector de p parámetros del modelo.

La media $E(Y)=\mu$ y la función de enlace $g(\mu)$ que permite relacionar el valor esperado de la variable respuesta con el predictor lineal $x'\beta$. Esto es:

$$E(Y_i) = \mu_i \quad \Rightarrow \quad g(\mu_i) = \eta_i = X'_i \underline{\beta}$$

2. Distribuciones muestrales

- **Distribución muestral de la variable score.** La estadística score se define como:

$$U_j = \frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n}\left[\frac{(y_i - \mu_i)}{V(Y_i)} X_{ij}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)\right] \; para \;\; j = 1,2,\ldots, p$$

La media, por propiedad es: $E(U_j) = 0\,;\; j = 1,2,\ldots, p$

La matriz de variancia-covariancia: $\Im_{jk} = E\left(U_j U_k\right) = \sum_{i=1}^{n} \frac{x_{ij} - x_{ik}}{V(Y_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$

Para un solo un parámetro β, la estadística score tiene una distribución muestral asintótica: $\frac{U}{\sqrt{\Im}} \approx N(0,1) \quad o \quad \frac{U^2}{\Im} \approx \chi^2_{(1)}$

Para un vector de parámetros $\beta' = (\beta_1, \beta_2, \ldots, \beta_p)$, entonces, el vector score $U' = (U_1, U_2, \ldots, U_p)$, se distribuye Normal multivariada asintóticamente (muestras grandes): $\underline{U} \approx N(\underline{0}, \Im)$

Para muestras grandes: $\underline{U}' \Im^{-1} \underline{U} \approx \chi^2_{(p)}$

2. Distribuciones muestrales

- **Distribución muestral del vector de coeficientes.** Se obtiene la distribución muestral del estimador de máxima verosimilitud $b'=(b_1, b_2, \ldots, b_p)$ con respecto al vector de parámetros $\beta'=(\beta_1, \beta_2, \ldots, \beta_p)$.

La media: $E(\underline{b}) = \underline{\beta}$

La matriz de variancia-covariancia: Puesto que: $\Im = E(UU')\ y\ (\Im^{-1})' = \Im^{-1}$

$$Cov(\underline{b}) = E\left[(\underline{b} - \underline{\beta})(\underline{b} - \underline{\beta})'\right] = \Im^{-1} E(UU')\Im = \Im^{-1}$$

La distribución muestral asintótica para el vector $\beta'=(\beta_1, \beta_2, \ldots, \beta_p)$, será:

$$(b-\beta)'\Im(b)(b-\beta) \approx \chi^2_{(p)}$$

donde $\Im(b)$ es la información evaluada en $\hat{\beta} = b$.

Esta es conocida como la ***Estadística de Wald.***

2. Distribuciones muestrales

- **Distribución muestral de la desviance.** La deviance conocida también como la estadística de razón log-verosimilitud es definida por:

Dónde:

$$D = 2\left[l(b_{\max}; y) - l(b; y)\right]$$

$l(b_{\max}; y)$ Es la función log-verosimilitud para el modelo saturado evaluado para el vector estimado b_{max} que corresponde al vector de parámetros β_{max}.

$l(b; y)$ Es la función log-verosimilitud para el modelo de interés para el vector estimado b que corresponde al vector de parámetros β.

Entonces se tiene que: $D \approx \chi^2_{(m-p-v)}$

v : Es el parámetro de no centralidad.
m : El número de parámetros del modelo saturado.
p : El número de parámetros del modelo de interés.

Ejemplo 1. Desviance para un modelo de regresión Poisson

Sean $Y_1, Y_2, \ldots, Y_N$ variables aleatorias independientes e idénticamente distribuidas como $Y_i \approx Poisson(\lambda_i)$,entonces la función de log-verosimilitud es:

$$l(\beta; y) = \sum_{i=1}^{N} [y_i \log \lambda_i - \lambda_i - \log y_i!]$$

Para el modelo saturado (con N parámetros), los λ_i son todos diferentes tal que se tiene $\underline{\beta} = (\lambda_1, \lambda_2, \ldots, \lambda_N)'$. Los estimadores MV son $\hat{\lambda}_i = y_i$,tal que el valor máximo de la función log-verosimilitud será:

$$l(b_{\max}; y) = \sum [y_i \log y_i - y_i - \log y_i!]$$

Para el modelo de interés con $p < N$ parámetros. El estimador MV $\underline{b}$ puede ser usado para calcular estimadores $\hat{\lambda}_i = \hat{y}_i \quad (E(y_i) = \lambda_i)$ y los valores ajustados $\hat{\lambda}_i$. Entonces el valor de la función log-verosimilitud evaluada para estos valores será:

$$l(b; y) = \sum [y_i \log \hat{y}_i - \hat{y}_i - \log y_i!]$$

Por lo tanto la Deviance $D = 2[l(b_{\max}; y) - l(b; y)] = D = 2\left[\sum_{i=1}^{N} y_i \log(\frac{y_i}{\hat{y}_i}) - \sum_{i=1}^{N} (y_i - \hat{y}_i)\right]$

Pero como $\sum y_i = \sum \hat{y}_i$, entonces la deviance

$$D = 2\sum_{i=1}^{N} y_i \log(\frac{y_i}{\hat{y}_i})$$

Ejemplo 2. Desviance para un modelo Binomial

Sean $Y_1, Y_2, \ldots, Y_N$ las variables respuestas independientes e idénticamente distribuidas como $Y_i \approx Binomial(n_i, \pi_i)$, entonces la función de log-verosimilitud es:

$$l(\beta; y) = \sum_{i=1}^{N} \left[y_i \log \pi_i - y_i \log(1-\pi_i) + n_i \log(1-\pi_i) + \log C_{y_i}^{n_i} \right]$$

Para el modelo saturado, los π_i's son todos diferentes tal que $\beta = (\pi_1 \pi_2, \ldots, \pi_N)'$ se tiene $\hat{\pi}_i = y_i / n_i$. Los estimadores MV son tal que el valor máximo de la función log-verosimilitud es:

$$l(b_{\max}; y) = \sum \left[y_i \log\left(\frac{y_i}{n_i}\right) - y_i \log\left(\frac{n_i - y_i}{n_i}\right) + n_i \log\left(\frac{n_i - y_i}{n_i}\right) + \log C_{y_i}^{n_i} \right]$$

Para modelo de interés con $p < N$ parámetros, sea $\hat{\pi}_i$ el estimador MV y sea $\hat{y}_i = n_i \hat{\pi}_i$ los valores ajustados. Entonces la función log-verosimilitud evaluada para estos valores es:

$$l(b; y) = \sum \left[y_i \log\left(\frac{\hat{y}_i}{n_i}\right) - y_i \log\left(\frac{n_i - \hat{y}_i}{n_i}\right) + n_i \log\left(\frac{n_i - \hat{y}_i}{n_i}\right) + \log C_{y_i}^{n_i} \right]$$

Por lo tanto la Deviance se expresa por: $D = 2[l(b_{\max}; y) - l(b; y)]$

$$D = 2\sum_{i=1}^{N} \left[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i) \log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right) \right], \quad con\ \hat{y}_i = n_i \hat{\pi}_i$$

Donde los $\hat{\pi}_i$ son las probabilidades estimadas del modelo de interés

3. Pruebas y medidas de bondad de ajuste

Existen una variedad de estadísticas que permiten evaluar la adecuación del modelo ajustado a un conjunto de datos.

Modelo saturado. Es el modelo donde el número de parámetros estimados e igual al número de observaciones. Utilizar este modelo implicara estimar el numero de parámetros igual al tamaño de muestra.

Modelo nulo. Este es el modelo que se utiliza como de referencia. Contiene como único parámetro (intercepto) el valor esperado para todas las observaciones.

Modelo de investigación. Es el modelo de interés o de investigación, contienen $p<n$ coeficientes asociados a cada variable predictora.

3. Pruebas y medidas de bondad de ajuste

1. **Prueba de Deviance o Desvio.** Es la medida mas usada en los MLG para evaluar el ajuste de un modelo. Es la distancia entre el logaritmo de la función verosimilitud del modelo saturado (con N parámetros) y el modelo en investigación (con p parámetros).

- **Formulación de hipótesis.** Las hipótesis que se formulan para la estadística de deviance son:

  Ho: El modelo se ajusta a los datos
  H1: El modelo no se ajusta a los datos

- **La prueba estadística:** $D_c = 2\left[l(b_{\max}; y) - l(b; y)\right]$

cuanto muy pequeño es la devianza, es porque se está obteniendo un buen ajuste como cuando se ajusta a un modelo saturado

- **La decisión estadística.** Se rechaza Ho, si: $D_c \geq \chi^2_{(N-p)}$

➢ Un valor pequeño de la Deviance, indica que para un número menor de parámetros, se obtiene un ajuste tan bueno como cuando se ajuste con un modelo saturado.

3. Pruebas y medidas de bondad de ajuste

2. **Estadística razón de log-verosimilitud.** Permite evaluar la adecuación de un modelo. Se basa en comparar el cociente de la función log-verosimilitud del modelo saturado o completo (con N parámetros) y la del modelo de interés (con p parámetros).

- **Formulación de hipótesis.** Las hipótesis que se formulan para la estadística de deviance son:

  Ho: El modelo se ajusta a los datos
  H1: El modelo no se ajusta a los datos

- **La prueba estadística:** $\lambda_c = \dfrac{l(b_{\max}; y)}{l(b; y)} \Rightarrow \log \lambda_c = l(b_{\max}; y) - l(b; y)$

- **La decisión estadística.** Se rechaza Ho, si: $\lambda_c \geq \chi^2_{(N-p)}$

➢ Valores altos sugiere que el modelo de interés es una pobre descripción de los datos, respecto del modelo saturado.

3. Pruebas y medidas de bondad de ajuste

3. **Coeficiente de determinación.** Se define como la reducción proporcional en la incertidumbre debido a la inclusión de los regresores. **El Pseudo R^2** (McFadden), se ha propuesto como una medida relativa de la mejora de la log-verosimilitud. A mayor log-verosimilitud, mejor será el modelo ajustado. Compara la log verosimilitud del modelo de interés con respecto al modelo mínimo (sólo itercepto).

$$Pseudo\ R^2 = 100 \times \left(1 - \frac{D(y,\hat{y})}{D(y,\hat{y}_0)}\right) = 100 \times \left(1 - \frac{l(b;y)}{l(b_{\min};y)}\right)$$

$D(y,\hat{y})$ y $D(y,\hat{y}_0)$ Son las desviances del modelo ajustado y el modelo nulo o mínimo (sólo con intercepto) respectivamente

El Pseudo R^2, mide el porcentaje de cuanto se reduce la desvianza del modelo nulo (con intercepto), cuando se adiciona las p variables predictoras.

3. Pruebas y medidas de bondad de ajuste

4. **Criterio de información Akaike (AIC=Akaike information criterios).** Es una medida relativa para evaluar la bondad de ajuste de los modelos estadísticos. Se selecciona el modelo con menor AIC. En general, el AIC se calcula:

$$AIC = 2k - 2\ln(L) = 2k - 2Ln(l(b;y))$$

Donde:
k= Números de parámetros del modelo
L =Valor máximo de la función verosimilitud para el modelo estimado.

5. **Residuales de Pearson.** Mide la diferencias entre el valor observado de Y y el valor predecido con el modelo ajustado. Su expresión:

$$r_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i}}$$

4. Procedimiento para probar hipótesis

El procedimiento de prueba de hipótesis con la finalidad de buscar el mejor modelo que se ajuste a la variable respuesta, se sigue como proceso de **modelo reducido**, con los siguientes pasos:

1. Especificar un modelo ***Mo*** asociado a: $H_0 : \underline{\beta}'_0 = \lfloor \beta_1, \beta_2, ..., \beta_q \rfloor$

   Especificar un modelo ***M1*** asociado a: $H_1 : \underline{\beta}'_1 = \lfloor \beta_1, \beta_2, ..., \beta_p \rfloor$

   (con q<p, el ***Mo*** como un caso particular de ***M1***)

2. Ajustar los datos a ***Mo*** y calcular la estadística: $D_0 \approx \chi^2_{(n-q)}$

   Ajustar los datos a ***M1*** y calcular la estadística: $D_1 \approx \chi^2_{(n-p)}$

3. Se prueba H0 contra H1 usando la diferencia de la estadística deviance:

$$\Delta D_c = D_0 - D_1 \approx \chi^2_{(p-q)}$$

Se rechaza Ho, si **ΔDc** > $X^2_{(p-q)}$, por lo tanto el modelo ***Mo*** (con las q variables predictoras) no se ajusta a los datos; sino, el modelo **M1** con la p variables predictoras)

Ejercicio 1.

Se tiene un experimento sobre la aplicación de dosis de un funguicida a sembríos de manzano. Se seleccionan 8 plantas de manzano y se recolectaron 15 hojas evaluando en cada hoja el número de hembras adultas de ácaros vivas. En la siguiente tabla se muestra los datos registrados por planta:

Número de hembras vivas	1	3	7	15	9	17	5	13
Dosis de funguicida (grs)	2	4	6	9	7	12	2	10

a. Formule el MLG, ajustando los datos a una regresión Poisson con función de enlace logaritmo.
b. Determine las medidas de bondad de ajuste
c. Calcule los residuales de Pearson y Deviance
d. Calcule el R^2

Solución:

a. Formule el MLG, ajustando los datos a una regresión Poisson con función de enlace logaritmo.

Variable dependiente: Y=Número de ácaros hembras
Variable predictora: X=Dosis de funguicidas (grs.)

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

Dónde: μ= Número promedio de ácaros hembras

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

Función de enlace (logaritmo):

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

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

Verificando los resultados con la función glm:

```
> #Entrada de datos
> y<-c(1,3,7,15,9,17,5,13)
> x<-c(2,4,6,9,7,12,2,10)
> #  Estimación de los coeficientes y la ecuación estimada
>  Modelo_31<-glm(y~x,family=poisson(log))
>  summary(Modelo_31)
Call:
glm(formula = y ~ x, family = poisson(log))
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.81157    0.34497   2.353   0.0186 *
x            0.17972    0.03788   4.745 2.09e-06 ***
---
    Null deviance: 30.0708  on 7  degrees of freedom
Residual deviance:  5.1706  on 6  degrees of freedom
AIC: 39.158
```

Cálculo de la deviance (Residual Deviance)

```
> #  Cálculo de la función de VM del modelo saturado: l(bmax)
> l_bmax=sum(y*log(y)-y-log(factorial(y))); l_bmax
[1] -14.99361
> #  Cálculo de la función de VM del modelo de interés: l(b)
yest<-exp(coef(Modelo_31)[1]+coef(Modelo_31)[2]*x); yest
> l_b=sum(y*log(yest)-yest-log(factorial(y))); l_b
[1] -17.57891
> #  Cálculo de la Deviance
> D=2*(l_bmax-l_b); D
[1] 5.170594
> D=2*sum(y*log(y/yest)); D# (Usando la fórmula)
[1] 5.170594
```

$$Se\ tiene: \quad l(b_{\max}; y) = \sum \left[y_i \log y_i - y_i - \log y_i ! \right] = -14.9936$$

$$l(b; y) = \sum \left[y_i \log \hat{y}_i - \hat{y}_i - \log y_i ! \right] = -17.5789$$

$$Entonces: D = 2\left[l(b_{\max}; y) - l(b; y) \right] = 2\left[-14.9936 - (-17.5789) \right] = 5.1706$$

$$D = 2\sum_{i=1}^{N} y_i \log\left(\frac{y_i}{\hat{y}_i}\right) = 2x2.5853 = 5.1706 \quad (\text{Re}\, sidual\ deviance)$$

Cálculo de la Null Deviance

```
> #  Cálculo de la Desviance Null o Modelo mínimo
> D_Null=2*sum(y*log(y/mean(y))); D_Null
[1] 30.07076
> glm(y~1,family=poisson(log))
Call:  glm(formula = y ~ 1, family = poisson(log))
Coefficients:
(Intercept)
      2.169
Null Deviance:      30.07
Residual Deviance: 30.07        AIC: 62.06
```

$$D_{NULL} = 2\sum_{i=1}^{N} y_i \log(\frac{y_i}{\overline{y}_i}) = 30.0708 \quad (R: Null\ Deviance)$$

Cálculo del Criterio de información de Akaike (AIC)

```
> #   Cálculo de AIC
> AIC=-2*sum(y*log(yest)-yest-log(factorial(y)))+2*2; AIC
[1] 39.15781
```

$$AIC = -2l(b; y) + kn,\ n = 2\ (Número\ de\ parámetros)\ y\ k = 2\ (Valor\ usual)$$

$$AIC = -2\sum(y_i \log \hat{y}_i - \hat{y}_i - \log y_i!) + 2*2 = 39.1578$$

Cálculo de los residuales de Pearson y Deviance

```
Cálculo de los residuales de Pearson y Deviance
> #  Cálculo de los residuals de Pearson
> Ri=(y-yest)/sqrt(yest); Ri
[1] -1.2390824 -0.7538107  0.1481946  1.0839907  0.3830526 -0.5571357  0.9882072
[8] -0.1580910
> Res_Per<-residuals(Modelo_31,type="pearson"); Res_Per
    1       2         3         4       5       6        7           8
-1.23908 -0.75381  0.14819  1.08399  0.38305 -0.55713  0.98820    -0.15809
> # Cálculo de los residuals de Desviance
> Di=y*log(y/yest); Di
[1] -1.17101 -1.29555  0.39203  4.18469  1.14838 -2.29539  2.19210 -0.56995
> Res_Des<-residuals(Modelo_31,type="deviance") ;Res_Des
     1         2       3         4        5       6        7         8
-1.45207 -0.80593  0.14680  1.03249  0.37482 -0.56953  0.913655   -0.15924
```

$$Pearson: \quad r_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i}} \qquad Deviance: D_i = y_i \log(\frac{y_i}{\hat{y}_i})$$

Cálculo del coeficiente de determinación:

```
> #  Cálculo del Coeficiente de determinación
> Desv_Residual=2*sum(y*log(y/yest)); Desv_Residual
[1] 5.170594
> Desv_Nulo=2*sum(y*log(y/mean(y))); Desv_Nulo
[1] 30.07076
> R2=100*(1-Desv_Residual/Desv_Nulo); R2
[1] 82.80524
> R2= (1-Modelo_31$deviance/Modelo_31$null.deviance)*100; R2
[1] 82.80524
```

$$Pseudo\ R^2 = 100x\left(1-\frac{D(y,\hat{y})}{D(y,\hat{y}_0)}\right) = 100x\left(1-\frac{l(b;y)}{l(b_{\min};y)}\right)$$

$$l(b;y) = D(y,\hat{y}) = D = 5.1706 \qquad (R: Deviance\ residual)$$

$$l(b_{\min};y) = D(y,\hat{y}_0) = D_{NULL} = 30.0708 \qquad (R: Null\ deviance)$$

$$Entonces: \quad Pseudo\ R^2 = 100x\left(1-\frac{5.1706}{30.0708}\right) = 82.8\%$$

Ejercicio 2.

Se desea estudiar la germinación de cultivos de alga en función de un factor de almacenamiento que tiene dos categorías (sin y con almacenamiento a 48 horas a 3°C) y una covariable definida como la fuerza centrifuga aplicada durante la preparación del cultivo, tomando solo uno de los tres valores: 40,150 y 350. Los resultados sobre el número de plantas germinadas se muestra en la siguiente tabla:

Factor		Fuerza centrífuga		
		40	150	350
Sin almacenamiento	Número	55	52	57
	Total	102	99	108
Con almacenamiento	Número	55	50	50
	Total	76	81	90

a. Ajuste los datos a una regresión logística. Formule el MLG
b. Realice la prueba bondad de ajuste y significación de los coeficientes. Use un nivel de significación de 0.05.

Modelo 1. Solo la covariable: Fuerza centrífuga.
Modelo 2. Adicionar el factor: Almacenamiento.
Modelo 3. Con la interacción covariable y factor.

a. Ajuste los datos a una regresión logística. Formule el MLG

Variable respuesta:

$$Y = Número\ de\ plantas\ ger\min adas \begin{cases} 1 = Ger\min ó \\ 0 = No\ ger\min ó\ (Categoría\ de\ referencia) \end{cases}$$

Variables explicativas:

$$Factor: \quad X_1 = Almacenamiento = \begin{cases} 0 = \text{Sin } almacenamiento\ (Categoría\ de\ referencia) \\ 1 = Con\ almacenamiento \end{cases}$$

$$Co\text{var}iable: \quad X2 = \log(Fuerza\ centrífuga) \quad valores: 40, 150, 350$$

$$Interacción: \quad X3 = X1 * X2$$

Componentes del MLG:

$$Componente\ aleatorio: Y_i \approx Binomial(n_i, \pi_i)$$

$$Componente\ sistemático: \eta = X'\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

$$Función\ de\ enlace: Logit\ (\log ístico)$$

Modelo logístico:

$$P(Y = 1/X2) = \pi_i = \frac{\exp(\beta_0 + \beta_2 X2)}{1 + \exp(\beta_0 + \beta_2 X2)} = \frac{1}{1 + \exp(-(\beta_0 + \beta_2 X2))}$$

Modelo logit:

$$\log it\left(P(Y = 1/X2)\right) = \log it(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_2 X2$$

Modelo 1. Solo la covariable: Fuerza centrífuga.

```
> #  Entrada de datos
> y<-c(55,52,57,55,50,50)
> n<-c(102,99,108,76,81,90)
> yy<-cbind(y,n-y)
> fuerza<-c(40,150,350,40,150,350)
> factor<-c(0,0,0,1,1,1)
> lfuerza<-log(fuerza);  fxf<-lfuerza*factor
> #  Modelo 1. Covariable: X2=log(Fuerza Centrífuga)
> Modelo1<-glm(yy~lfuerza,family=binomial(link=logit))
> summary(Modelo1)
Call: glm(formula = yy ~ lfuerza, family = binomial(link = logit))
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.0213     0.4813   2.122   0.0338 *
lfuerza      -0.1478     0.0965  -1.532   0.1255
    Null deviance: 10.4520  on 5  degrees of freedom
Residual deviance:  8.0916  on 4  degrees of freedom
AIC: 41.66
```

Modelo logístico estimado:

$$P(Y=1/\ X2)=\hat{\pi}_i=\frac{\exp(1.0213-0.1478X2)}{1+\exp(1.0213-0.1478X2)}$$

$$\log it(\hat{\pi}_i)=\log(\frac{\hat{\pi}_i}{1-\hat{\pi}_i})=1.0213-0.1478X2$$

Modelo 1. Prueba de significación de la covariable: X2=Fuerza centrífuga.

```
> #  Cálculo de la desviance
> Pi=fitted(Modelo1); Yest=n*Pi
> D = 2*sum(y*log(y/Yest)+(n-y)*log((n-y)/(n-Yest))); D
[1] 8.091578
```

$$D = 2\left[l(b_{\max}; y) - l(b; y)\right] = 2x\sum_{i=1}^{N}\left[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right] = 8.0916$$

Formulación de hipótesis:

Ho: El Modelo1 de regresión logístico se ajusta a los datos

H1: El Modelo1 de regresión logístico No se ajusta a los datos

Prueba estadística:

```
> Alfa=0.05
> Chi_Tab=qchisq(1-Alfa,Modelo1$df.residual); Chi_Tab
[1] 9.487729
> p_valor=1-pchisq(Modelo1$deviance,Modelo1$df.residual); p_valor
[1] 0.08828059
```

$$Como: D = \chi_c^2 = 8.0916 < \chi^2_{(0.05,4)} = qchisq(0.95,4) = 9.487,\ No\ se\ rechaza\ Ho.$$

$$Como: p - valor = 1 - pchisq(8.0916,4) = 0.088 > \alpha = 0.05, No\ se\ rechaza\ Ho.$$

Conclusión. Con un nivel de significación de 0.05, el Modelo1 de respuesta binaria se ajusta a los datos.

Cálculo del coeficiente de determinación (Pseudo R^2)

```
> #  Cálculo del Coeficiente de determinación
> Desv_Residual=2*sum(y*log(y/Yest)+(n-y)*log((n-y)/(n-Yest)));
Desv_Residual
[1] 8.091578
> R2= (1-Modelo1$deviance/Modelo1$null.deviance)*100; R2
[1] 22.58326
```

$$R^2 = 100x\left(1-\frac{D(y,\hat{y})}{D(y,\hat{y}_0)}\right) = 100x\left(1-\frac{Des.\text{Re}\,sidual}{Des.Null}\right) = 100x\left(1-\frac{8.0916}{10.4520}\right) = 22.6\%$$

Prueba del coeficiente de regresión:

$$H_0 : \beta_2 = 0$$

$$H_1 : \beta_2 \neq 0 \qquad Z_c = \frac{\hat{\beta}_2 - \beta_2}{S_{\beta_2}} = \frac{-0.1478 - 0}{0.0965} = -1.532 \quad .No\ se\ rechaza\ H_0.$$

Conclusión. Con un nivel de significación de 0.05, la fuerza centrífuga no es significativa al modelo.

Modelo 2. Factor: X1=Almacenamiento y covariable: X2=Fuerza centrífuga.

Modelo logístico:

$$P(Y=1/\ X1, X2)=\pi_i=\frac{\exp(\beta_0+\beta_1 X1+\beta_2 X2)}{1+\exp(\beta_0+\beta_1 X1+\beta_2 X2)}$$

```
> #  Modelo 2. Facto: X1=Almacenamiento y Covariable: X2= log(Fuerza Cent)
> Modelo2<- glm(yy~lfuerza+factor,family=binomial(link=logit))
> summary(Modelo2)
Call: glm(formula = yy ~ lfuerza + factor, family = binomial(link = logit))
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.87673    0.48701   1.800   0.0718 .
lfuerza     -0.15459    0.09702  -1.593   0.1111
factor       0.40684    0.17462   2.330   0.0198 *
---
    Null deviance: 10.4520  on 5  degrees of freedom
Residual deviance:  2.6188  on 3  degrees of freedom
AIC: 38.187
```

Modelo logístico estimado:

$$P(Y=1/\ X1, X2)=\hat{\pi}_i=\frac{\exp(0.8767+0.4068X1-0.1546X2)}{1+\exp(0.8767+0.4068X1-0.1546X2)}$$

$$\log it(\hat{\pi}_i)=\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right)=0.8767+0.4068X1-0.1546X2$$

Prueba de significación del Modelo 2.

Formulación de hipótesis:

Ho: El Modelo2 de regresión logístico se ajusta a los datos

H1: El Modelo2 de regresión logístico No se ajusta a los datos

Prueba estadística:

```
> Alfa=0.05
> Chi_Tab=qchisq(1-Alfa,Modelo2$df.residual); Chi_Tab
[1] 7.814728
> p_valor=1-pchisq(Modelo2$deviance,Modelo2$df.residual); p_valor
[1] 0.4541967
```

$$Como: D_c = \chi^2_c = 2.6188 < \chi^2_{(0.05,3)} = qchisq(0.95,3) = 7.814728,\ No\ se\ rechaza\ Ho.$$

$$Como: p-valor = 1- pchisq(2.6188,3) = 0.4542 > \alpha = 0.05,\ No\ se\ rechaza\ Ho.$$

Conclusión. Con un nivel de significación de 0.05, el Modelo2 de logístico se ajusta a los datos. Es decir, el número de plantas germinadas puede ser explicada por la fuerza centrífuga y almacenamiento.

Modelo 3. X1=Almacenamiento, X2=Fuerza centrífuga y X3=Interacción.

Modelo logístico:

$$P(Y=1/\ X1, X2, X3)=\pi_i=\frac{\exp(\beta_0+\beta_1 X1+\beta_2 X2+\beta_3 X3)}{1+\exp(\beta_0+\beta_1 X1+\beta_2 X2+\beta_3 X3)}$$

```
> Modelo3<-glm(yy~lfuerza+factor+fxf,family=binomial(link=logit))
> summary(Modelo3)
 Call: glm(formula = yy ~ lfuerza + factor + fxf, family = binomial(link =
logit))
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.23389    0.62839   0.372   0.7097
lfuerza     -0.02274    0.12685  -0.179   0.8577
factor       1.97711    0.99802   1.981   0.0476 *
fxf         -0.31862    0.19888  -1.602   0.1091
---
    Null deviance: 10.451974  on 5  degrees of freedom
Residual deviance:  0.027728  on 2  degrees of freedom
AIC: 37.596
```

Modelo logístico estimado:

$$P(Y=1/\ X1, X2, X3)=\hat{\pi}_i=\frac{\exp(0.233+1.977 X1-0.023 X2-0.318 X3)}{1+\exp(0.233+1.977 X1-0.023 X2-0.318 X3)}$$

$$\log it(\hat{\pi}_i)=\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right)=0.233+1.977 X1-0.023 X2-0.318 X3$$

Prueba de significación del Modelo 3.

Formulación de hipótesis:

Ho: El Modelo3 de regresión logístico se ajusta a los datos

H1: El Modelo3 de regresión logístico No se ajusta a los datos

Prueba estadística:

```
> Alfa=0.05
> Chi_Tab=qchisq(1-Alfa,Modelo3$df.residual); Chi_Tab
[1] 5.991465
> p_valor=1-pchisq(Modelo3$deviance,Modelo3$df.residual); p_valor
[1] 0.9862318
```

$$Como: D = \chi_c^2 = 0.0277 < \chi^2_{(0.05,3)} = qchisq(0.95,2) = 5.991465, \textit{ No se rechaza Ho.}$$

$$Como: p-valor = 1- pchisq(0.0277,2) = 0.986 > \alpha = .0.5, \textit{ No se rechaza Ho.}$$

Conclusión. Con un nivel de significación de 0.05, el Modelo3 regresión logístico se ajusta a los datos. Es decir, el número de plantas germinadas puede ser explicada por la fuerza centrífuga, almacenamiento e interacción.

Método del modelo reducido para seleccionar el mejor modelo.

```
anova(Modelo1, Modelo2, test= "Chisq")
Analysis of Deviance Table
Model 1: yy ~ lfuerza
Model 2: yy ~ lfuerza + factor
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         4     8.0916
2         3     2.6188  1   5.4727  0.01932 *
anova(Modelo2, Modelo3, test= "Chisq")
Model 1: yy ~ lfuerza + factor
Model 2: yy ~ lfuerza + factor + fxf
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         3    2.61884
2         2    0.02773  1   2.5911   0.1075
```

Probando si el almacenamiento (X1) es significativo

$$\begin{cases} H_0 : \underline{\beta}_0 = (\beta_2) & \Rightarrow D_0 = 8.0916 \approx \chi^2_{(4)} \quad (Modelo\ 1) \\ H_1 : \underline{\beta}_1 = (\beta_1, \beta_2) & \Rightarrow D_1 = 2.6188 \approx \chi^2_{(3)} \quad (Modelo\ 2) \end{cases}$$

$$\Delta D = D_0 - D_1 = 8.0916 - 2.6188 = 5.4728 \approx \chi^2_{(1)}$$

$$p - valor = 1 - pchisq(5.4728,1) = 0.019 \leq \alpha = 0.05 \quad .Se\ rechaza\ H_0.$$

Conclusión. Con un nivel de significación de 0.05, se puede afirmar que el almacenamiento y la fuerza centrífuga son significativos para explicar el número de plantas germinadas.

Probando si la interacción (X3) es significativa

$$\begin{cases} H_0 : \underline{\beta}_0 = (\beta_1, \beta_2) & \Rightarrow D_0 = 2.6188 \approx \chi^2_{(3)} \quad (Modelo\ 2) \\ H_1 : \underline{\beta}_1 = (\beta_1, \beta_2, \beta_3) & \Rightarrow D_1 = 0.0277 \approx \chi^2_{(2)} \quad (Modelo\ 3) \end{cases}$$

$$\Delta D = D_0 - D_1 = 2.6188 - 0.0277 = 2.5911 \approx \chi^2_{(1)}$$

$$p - valor = 1 - pchisq(2.5911,1) = 0.1074 \geq \alpha = 0.05 \quad .No\ se\ rechaza\ H_0.$$

Conclusión. Con un nivel de significación de 0.05, se puede afirmar que el almacenamiento y la fuerza centrífuga explican el número de plantas germinadas. Es decir, la interacción no es significativa.

Por lo tanto el Modelo 2 es significativo; es decir, el número de plantas germinadas puede ser explica por el almacenamiento y la fuerza centrífuga.

Intervalos de confianza del 95% para el Modelo 2

```
> NC=0.95
> confint.default(Modelo2, level=NC)
                  2.5 %     97.5 %
(Intercept) -0.07780037 1.83125923
lfuerza     -0.34475773 0.03557198
factor       0.06458351 0.74909689
```

5. Ejercicios propuestos.

Una empresa de especialista en análisis de encuestas con la finalidad de establecer los costos de aplicar una encuesta de opinión en Lima, desea modelar el número de encuestas realizadas por día, asumiendo que es una variable aleatoria con distribución de Poisson, en función del tiempo (horas) de su aplicación. Se extrae una muestra de 10 digitadoras y se evalúa el número de encuestas digitadas y el tiempo (minutos) por día.

Número de encuestas	20	16	15	17	24	27	21	16	22	20
Tiempo (minutos)	10	8	6.5	8	12	12	10	7	9	8

a. Formule el MLG, ajustando los datos a una regresión Poisson con función de enlace logaritmo.
b. Determine las medidas de bondad de ajuste
c. Calcule los residuales de Pearson y Deviance
d. Calcule el R^2

6. 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.