9. ANÁLISIS DE REGRESIÓN

La regresión es una técnica estadística utilizada para simular la relación existente entre dos o más variables. Por lo tanto, se puede emplear para construir un modelo que permita predecir el comportamiento de una variable dada. En la construcción de un modelo en R, el operador ~ se utiliza para definir una fórmula. La forma para un modelo lineal ordinario es:

var respuesta ~ ope_1 term_1 ope_2 term_2 ...

donde **var respuesta** es un vector o una matriz que definen, respectivamente, la o las variables respuesta; **ope_i** es un operador, bien + o bien –, que implica la inclusión o exclusión, respectivamente, de un término en el modelo. El primer (ope_1), de ser +, es opcional, no es completamente necesario; **term_i** es un término de uno de los siguientes tipos:

- una expresión vectorial, una expresión matricial o el número 1
- un factor
- una expresión de fórmula consistente en factores, vectores o matrices conectados mediante operadores de fórmula.

En todos los casos, cada término define una colección de columnas que deben ser añadidas o eliminadas de la matriz del modelo. Un 1 significa un término independiente y está incluido siempre, a no ser que sea eliminado explícitamente. A continuación se muestran algunos ejemplos de modelos estadísticos, teniendo en cuenta que **y**, x_0, x_1, x_2, ... son variables numéricas, que **X** es una matriz y que **A** especifica los factores:

- **y ~ x** o **y ~ 1 + x** Ambos definen el mismo modelo de regresión lineal de **y** sobre **x**. El primero contiene el término independiente implícito y el segundo, explícito.
- **y ~ 0 + x, y ~ -1 + x** o **y ~ x - 1** Regresión lineal de **y** sobre **x** sin término independiente; esto es, que pasa por el origen de coordenadas.
- **log(y) ~ x_1 + x_2** Regresión múltiple de la variable transformada, log(y), sobre x_1 y x_2 (con un término independiente implícito).
- **y ~ Poly(x, 2)** o **y ~ 1 + x + I(x ^ 2)** Regresión polinomial de **y** sobre **x** de segundo grado. La primera forma utiliza polinomios ortogonales y la segunda utiliza potencias de modo explícito.
- **y ~ X + Poly(x, 2)** Regresión múltiple de **y** con un modelo matricial consistente en la matriz **X**, términos polinomiales en **x** de segundo grado.
- **y ~ A** Análisis de varianza de entrada simple de **y**, con clases determinadas por A.
- **y ~ A * x, y ~ A / x** ó **y ~ A / (1 – x) - 1** Modelos de regresión lineal simple separados de **y** sobre **x** para cada nivel de A. La última forma produce estimaciones explícitas de tantos términos independientes y pendientes como niveles tiene A.

9.1 MODELO LINEAL

El comando **lm()** es utilizado para ajustar modelos lineales mediante mínimos cuadrados:

lm(formula, data, model = TRUE)

Los argumentos utilizados en el comando **lm()** son: **fórmula** se refiere a la fórmula utilizada en la construcción del modelo a ajustar; **data**, al argumento opcional si los datos provienen de un data.frame que contiene las variables involucradas en el modelo; **model**, al argumento lógico, si es TRUE arroja los componentes del modelo ajustado. Además, hay más argumentos que se pueden consultar en la ayuda interactiva. El comando **lm()** arroja algunos resultados simples; las funciones summary y anova son utilizadas para obtener resumen y tabla de análisis de varianza, respectivamente, del modelo ajustado.

Ejemplo: El supervisor de mantenimiento de una línea de autobuses cree que existe una relación entre el costo anual de mantenimiento de las unidades y los años que llevan de operación, y considera que si tal relación existe podrá hacer un mejor pronóstico de presupuesto. Los datos tomados por el supervisor sobre 15 autobuses de la empresa se muestran en la siguiente tabla, x=tiempo de operación en años, y=costo de mantenimiento:

Tabla 9. Datos sobre autobuses

x	8	5	3	9	11	2	1	8	12	4	7	10	6	3	9
y	8.6	6.8	4.7	7	11	2.2	3.2	6.5	10.5	5.6	6.8	10.8	6.2	5	8

Modelo ajustado:

```
> # Tiempo de operación (años)
> x=c(8,5,3,9,11,2,1,8,12,4,7,10,6,3,9)
>
> # Costo de matenimiento
> y=c(8.6,6.8,4.7,7,11,2.2,3.2,6.5,10.5,
+      5.6,6.8,10.8,6.2,5,8)
> # Modelo lineal
> lm( y ~ x, model = TRUE)

Call:
lm(formula = y ~ x, model = TRUE)

Coefficients:
(Intercept)            x
      2.215        0.711
```

Imagen 162. Salida R Modelo ajustado

Como se observa, el comando **lm()** arroja resultados sencillos; de necesitar aspectos más específicos del modelo se deben usar funciones extractoras de información del modelo; la descripción de algunas funciones es:

- *formula*(lm(modelo)): Extrae la fórmula del modelo.
- *coefficients*(lm(modelo)): Extrae la matriz de coeficientes de regresión. Forma reducida: *coef*(lm(modelo)).
- *summary*(lm(modelo)): Imprime un resumen estadístico completo de los resultados del análisis de regresión.
- *anova*(lm(modelo)): Compara un submodelo con un modelo externo y produce una tabla de análisis de varianza.
- *residuals*(lm(modelo)): Extrae la matriz de residuos, ponderada si es necesario. La forma reducida es **resid**(lm(modelo)).
- *plot*(lm(modelo)): Crea cuatro gráficos que muestran los residuos, los valores ajustados y otros gráficos de diagnósticos para examinar la calidad del modelo.
- *predict*(lm(modelo)): El resultado es un vector o matriz de valores predichos correspondiente a los valores de las variables de los datos.
- *deviance*(lm(modelo)): Suma de cuadrados residual, ponderada si es lo apropiado.
- *step*(lm(modelo)): Selecciona un modelo apropiado añadiendo o eliminando términos y preservando las jerarquías. Se devuelve el modelo que en este proceso tiene el máximo valor de AIC (Criterio de información de Akaike).

Teniendo en cuenta estas funciones extractoras, se presentan algunas ilustraciones para el modelo ajustado, usando el ejemplo de mantenimiento de la línea de autobuses. Para obtener un resumen del modelo se utiliza el comando summary():

```
> # Resumen del modelo
> summary( lm( y ~ x, model = TRUE ) )

Call:
lm(formula = y ~ x, model = TRUE)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6139 -0.5029  0.2744  0.6747  1.4751

Coefficients:
            Estimate Std. Error  t value Pr(>|t|)
(Intercept)  2.21458    0.56945    3.889  0.00186 **
x            0.71103    0.07778    9.141 5.05e-07 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9953 on 13 degrees of freedom
Multiple R.squared: 0.8654,        Adjusted R-squared: 0.855
F-statistic: 83.57 on 1 and 13 DF,  p-value: 5.045e-07
```

Imagen 163. Salida R Resumen del Modelo ajustado

Para obtener la tabla de análisis de varianza se utiliza el comando anova(), como se muestra:

```
> # Tabla de análisis de varianza para el modelo
> anova( lm( y ~ x, model = TRUE ) )
Analysis of Variance Table

Response: y
          Df Sum Sq Mean Sq F value     Pr(>F)
x          1 82.779  82.779  83.567  5.045e-07 ***
Residuals 13 12.877   0.991
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Imagen 164. Salida R Anova del Modelo ajustado

Mediante el comando plot() se generan cuatro gráficos en una ventana interactiva, en la cual se pasa de un gráfico a otro con hacer un clic o dar un enter.

```
> # Gráfico para analizar el modelo
> plot(lm(y~x,model=TRUE))
```

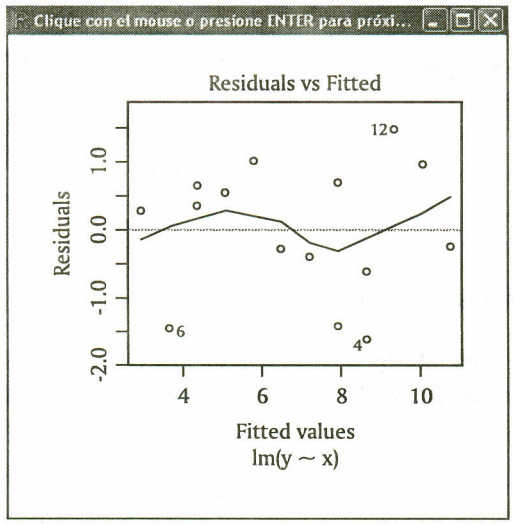

Imagen 165. Salida R Gráficos para el Modelo ajustado

9.2 MODELOS LINEALES GENERALIZADOS (GLM)

Cuando la variable respuesta es discreta o categórica, el modelo lineal clásico no es apropiado. Nelder y Wedderburn, en 1972, extendieron la teoría de los modelos lineales a una familia más amplia: la familia exponencial de densidades, denominándolos Modelos Lineales Generalizados (GLM), (Demétrio, 2001). Los GLM pueden ser usados cuando se tiene un único vector aleatorio Y asociado a un conjunto de variables explicativas o covariables $X_1, X_2,..., X_p$. Un GLM se define a través de tres componentes (Demétrio):

Componente aleatorio: Representado por un conjunto de variables aleatorias independientes $Y_1, Y_2, \ldots, Y_n$ provenientes de una misma distribución que hace parte de la familia exponencial de densidades. La familia exponencial de densidades fue propuesta por Pitman, Koopman y Darmois (Demétrio). Una distribución pertenece a esta familia si su función de densidad se puede llevar a la forma:

$$f\ (y,\theta,\phi) = \exp\left[\frac{1}{a(\phi)}(y\theta - b(\theta) + c\ (y,\phi))\right]$$

Ecuación 20. Familia exponencial de densidades

Componente sistemático: Para un GLM se considera un conjunto pequeño de parámetros β_1, $\beta_2, \ldots, \beta_p$ tal que la combinación lineal de los â's es igual a:

$$\eta_i = X_i^t \beta$$

Ecuación 21. Componente sistemático

Función de enlace: Función que relaciona la media con el predictor lineal, es decir, enlaza el componente aleatorio con el componente sistemático:

$$g\ (\mu_i) = \eta_i = X_i^t \beta$$

Ecuación 22. Función de enlace

Siendo g(*) una función monótona derivable y $\mu_{it} = E(Y_i)$.

Cada distribución para la variable respuesta admite una variedad de funciones de enlace para conectar la media con el predictor lineal. La siguiente tabla recopila las que están disponibles automáticamente en R.

Tabla 10: Funciones de enlace

Nombre de la familia	Función de enlace
binomial	logit, probit, cloglog
gaussian	Identity
Gamma	identity, inverse, log
inverse.gaussian	1/mu ^ 2
poisson	identity, log, sqrt
quasi	logit, probit, cloglog, identity, inverse, log, 1/mu ^ 2

El comando **glm()** permite ajustar un modelo lineal generalizado y tiene la siguiente estructura.

glm(formula, family=familia(link = "función de enlace"), data=)

Los argumentos utilizados en el comando anterior son: **formula**, que especifica el modelo; **family**, que especifica la familia, y la función de enlace (**link=** "") que se desee utilizar, y **data=**, que determina el conjunto de variables presentes en el modelo; estas pueden provenir de un data.frame, lista u otro ámbito permitido. Estos no son los únicos parámetros disponibles; para mayor información consulte la ayuda interactiva help(glm).

9.3 EJERCICIOS

9.3.1 En una encuesta realizada a 13 familias de la región se observaron las variables: número de integrantes de la familia (N) y gasto en alimentación por familia en miles (G); los resultados se muestran en la siguiente tabla:

N	3	2	5	4	6	3	2	4	5	5	6	4	3
G	150	120	180	180	210	160	90	150	200	150	225	170	100

a. Obtenga el modelo lineal que explica el gasto en alimentación de las familias en función de su tamaño.
b. Obtenga un resumen estadístico completo de los resultados del análisis de regresión.
c. Calcule los residuales y los valores predichos.

9.3.2 Un Supermercado ha decidido abrir una sucursal en la ciudad y decide estudiar el número de cajas registradoras que va a instalar, para evitar colas innecesarias a la hora de pagar. Para ello se obtuvieron los siguientes datos procedentes de otras sucursales en diferentes ciudades acerca del número de cajas registradoras (variable C) y del tiempo medio de espera en segundos (variable T).

N.° cajas	10	11	12	14	22	12	18	20
Tiempo	58	49	33	42	20	32	26	23

a. Obtenga la ecuación que explica la relación entre estas variables.
b. Obtenga la tabla de análisis de varianza.
c. Construya gráficos que le permitan realizar el diagnóstico del modelo construido.
d. Si se instalan 9 cajas registradoras, ¿cuál será el tiempo de espera?