3. Modelos lineales generalizados para datos binarios

Tarea-examen 1: Aprendizaje estadístico supervisado

Carlos Iván Canto Varela 315649888

Como visualización inicial de los datos, se presenta el siguiente diagrama de dispersión:

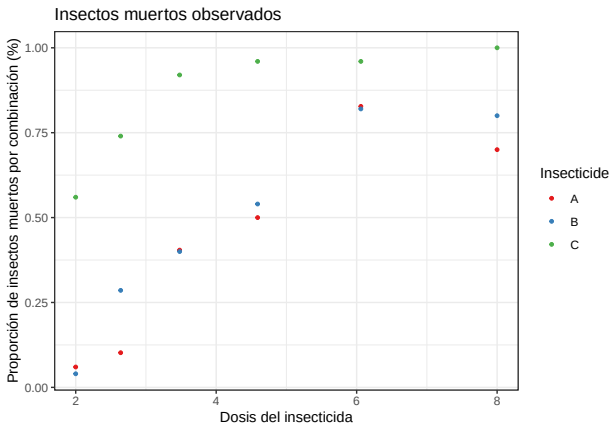

Figura 1: Diagrama de dispersión para los datos

De primera instancia, el insecticida C parece ser el más efectivo. Por otra parte, los insecticidas A y B muestran una alta similaridad percibida; quizá se podría eliminar una de estas categorías en el modelo a ajustar.

Se ajustará un modelo para datos binarios con las covariables *Insecticide*, *Deposit* y su interacción.

De esta manera, con Dep $\equiv$ dosis de insecticida, se tienen los componentes lineales para cada insecticida:

$$\begin{aligned} A: &\quad \beta_0 + \beta_3 \text{Dep}, \\ B: &\quad \beta_0 + \beta_1 + \beta_3 \text{Dep} + \beta_4 \text{Dep}, \\ C: &\quad \beta_0 + \beta_2 + \beta_3 \text{Dep} + \beta_5 \text{Dep}. \end{aligned}$$

Ningún modelo mostró disposición de cumplir los supuestos del componente lineal ni normalidad asintótica, por lo que se tiene que buscar una alternativa de regresión. Por ejemplo, aquella que incluye interacciones con la dosis en potencia cuadrada, donde los componentes sistemáticos del modelo extendido para cada insecticida son:

$$\begin{aligned} A: &\quad \beta_0 + \beta_3 \text{Dep} + \beta_4 \text{Dep}^2, \\ B: &\quad \beta_0 + \beta_1 + \beta_3 \text{Dep} + \beta_4 \text{Dep}^2 + \beta_5 \text{Dep} + \beta_7 \text{Dep}^2, \\ C: &\quad \beta_0 + \beta_2 + \beta_3 \text{Dep} + \beta_4 \text{Dep}^2 + \beta_6 \text{Dep} + \beta_8 \text{Dep}^2. \end{aligned}$$

Los modelos extendidos tendrán los puntajes

Puntajes por liga (modelo extendido)		
Liga	**Puntaje AIC**	**Puntaje BIC**
Logit	93.34989	101.3632
Probit	**93.30878**	**101.3221**
Cloglog	95.243	103.2563

De acuerdo a los criterios encontrados, los modelos extendidos son mejores que los que no consideraron al cuadrado de las dosis con una diferencia promedio de 20 puntos. Otra ventaja es que los supuestos sí se cumplen (con cierta reservación) para el modelo con liga *probit*.

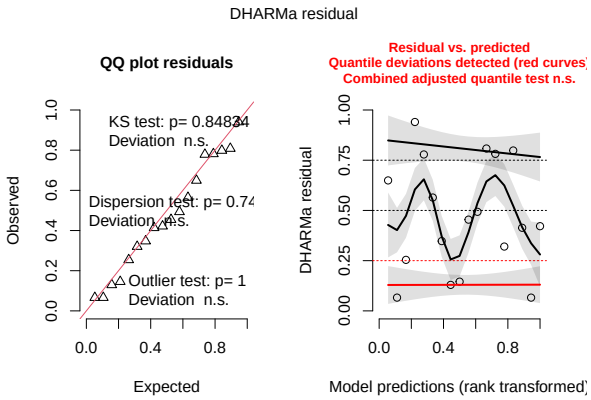

Figura 2: Gráficas de residuales simulados para los supuestos del nuevo modelo con liga probit

Dicho esto, los resultados de la estimación puntual están en la figura siguiente:

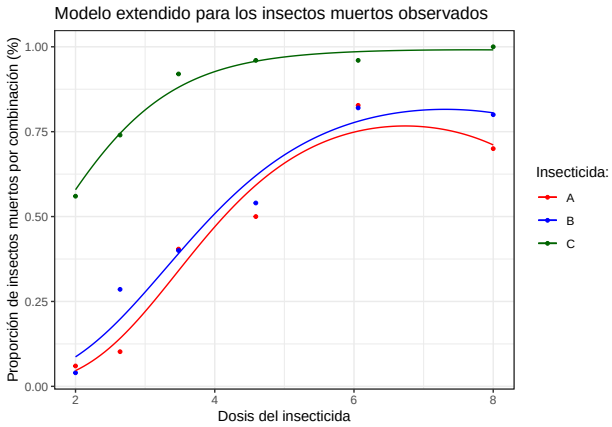

Figura 3: Curvas del modelo en el diagrama de dispersión de los datos

Según el modelo, la dosis mínima con la que se puede indicar que el 75 % de los insectos sucumbe está dada por la siguiente tabla:

Dosis necesaria para matar tres cuartos de los bichos	
Insecticida	**Dosis mínima**
A	6.03
B	5.63
C	**2.66**

Es obvio que el insecticida C tuvo los mejores resultados. ¿Se puede decir que es el mejor de todos los insecticidas? Para revisarlo se necesitan dos pruebas de hipótesis; sea $x \equiv$dosis de insecticida:

- C es mejor que A

  sólo sucede si

$$\begin{aligned} & \eta_C > \eta_A \\ \iff & \beta_0 + \beta_2 + \beta_3 x + \beta_4 x^2 + \beta_6 x + \beta_8 x^2 > \beta_0 + \beta_3 x + \beta_4 x^2 \\ \iff & \beta_2 + \beta_6 + \beta_8 > 0 \qquad \text{pues } x \geq 0. \end{aligned}$$

  Al tomar esta alternativa, la prueba a realizar es

$$H_0: \quad \beta_2 + \beta_6 + \beta_8 \leq 0 \quad \text{vs.} \quad H_a: \quad \neg H_0.$$

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: glm(formula = datum ~ Data$Insecticide * Data$Deposit + Data$Insecticide *
##     I(Data$Deposit^2), family = binomial(link = "probit"))
##
## Linear Hypotheses:
##        Estimate Std. Error z value   Pr(>z)
## 1 <= 0   2.1799     0.5334   4.086 2.19e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

- C es mejor que B:

$$\begin{aligned} & \eta_C > \eta_B \\ \iff & \beta_0 + \beta_2 + \beta_3 x + \beta_4 x^2 + \beta_6 x + \beta_8 x^2 \\ & > \\ & \beta_0 + \beta_1 + \beta_3 x + \beta_4 x^2 + \beta_5 x + \beta_7 x^2 \\ \iff & -\beta_1 + \beta_2 - \beta_5 + \beta_6 - \beta_7 + \beta_8 > 0 \qquad \text{pues } x \geq 0; \end{aligned}$$

similarmente, se tiene

$$H_0: \quad -\beta_1 + \beta_2 - \beta_5 + \beta_6 - \beta_7 + \beta_8 \leq 0 \quad \text{vs.} \quad H_a: \quad \neg H_0.$$

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: glm(formula = datum ~ Data$Insecticide * Data$Deposit + Data$Insecticide *
##     I(Data$Deposit^2), family = binomial(link = "probit"))
##
## Linear Hypotheses:
##        Estimate Std. Error z value   Pr(>z)
## 1 <= 0   1.6647     0.4916   3.386 0.000354 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Ya que en ambas pruebas se halló suficiente evidencia para rechazar que el insecticida C es peor que las alternativas, se puede afirmar que este es el mejor con una significancia de 0.05.

Previamente se mencionó que los insecticidas A y B resultaban bastante similares visualmente. Con otra prueba se puede verificar -o disentir- dicha afirmación:

$$\begin{aligned} & \eta_A = \eta_B \\ \iff & \beta_0 + \beta_3 x + \beta_4 x^2 = \beta_0 + \beta_1 + \beta_3 x + \beta_4 x^2 + \beta_5 x + \beta_7 x^2 \\ & \iff \beta_1 + \beta_5 x + \beta_7 x^2 = 0 \\ & \iff \beta_1 = \beta_5 = \beta_7 = 0; \qquad \text{pues } x \geq 0; \end{aligned}$$

de donde

$$H_0: \quad \begin{bmatrix} \beta_1 \\ \beta_5 \\ \beta_7 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad \text{vs.} \quad H_a: \quad \beta_i \neq 0 \text{ para } i \in \{1, 5, 7\}.$$

```
##
##   General Linear Hypotheses
##
## Linear Hypotheses:
##        Estimate
## 1 == 0  0.76399
## 2 == 0 -0.27608
## 3 == 0  0.02733
##
## Global Test:
##        F DF1 DF2 Pr(>F)
## 1 0.9557   3   9 0.4543
```

Con este resultado no se halló evidencia para negar que el insecticida A sea igual al B y se puede decir que tienen un desempeño similar con un valor p mayor al 5 %. Cabe mencionar que esto no afirma la igualdad ni la diferencia, sólo la similitud como causa natural de la hipótesis alternativa dada esta prueba.