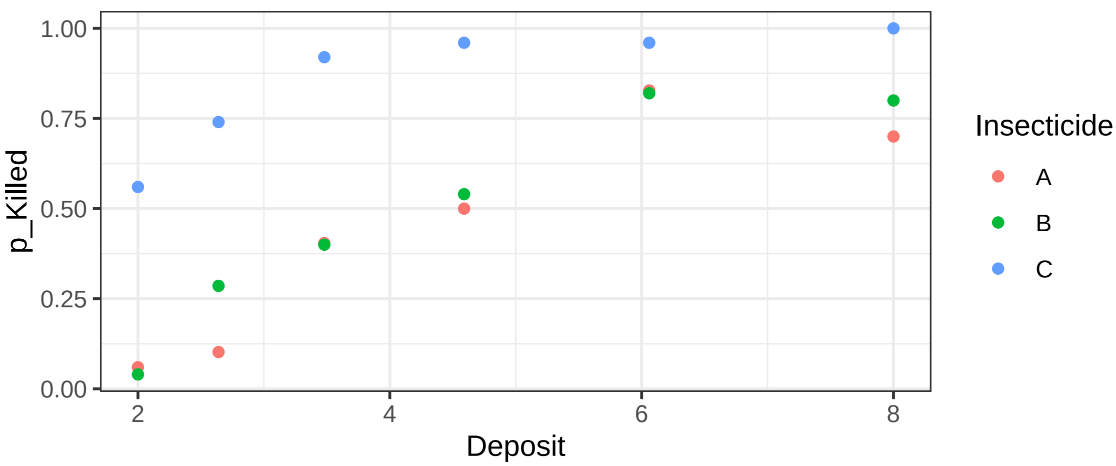
3. Modelos lineales generalizados para datos binarios

La base de datos Preg3B.csv contiene información sobre 862 insectos que fueron expuestos a diferentes dosis en mg (Deposit) de tres insecticidas (Insecticide). La asignación a una dosis y al tipo de insecticida se realizó de forma aleatoria. Después de seis días se analizó si los insectos se habían muerto, de manera que la base de datos contiene también el número de insectos muertos (Killed) y el número total de insectos expuestos (Number) por cada dosis e insecticida. Dado que se asume que el costo de los insecticidas es el mismo, el objetivo del análisis es identificar para cada insecticida qué dosis es la mínima con la que se puede indicar que el 70% de los insectos se muere, así como si considerando la menor de esas tres dosis se puede afirmar que un insecticida es el mejor comparado con el resto. El evento de interés es si el insecto muere o no (died).

i) Gráfica de dispersión de dosis del insecticida y la proporción de insectos muertos.

Se presenta una gráfica de dispersión en donde en el eje x se incluye la dosis del insecticida (Deposit) y en el eje y la proporción de insectos muertos observados (p_Killed) para cada combinación dosis-insecticida (Deposit-Insecticide), distinguiendo con un color el insecticida asociado. Se puede observar que el insecticida C tiene una mayor tasa de mortalidad para todas las seis dosis consideradas (solamente la primera dosis es menor a 70%). Para el caso de los insecticidas A y B, los resultados son muy parecidos, aunque marginalmente parece que el insecticida A tiene menor tasa de mortalidad, al menos de manera evidente en tres dosis distintas.



ii) Ajuste modelos para datos binarios 1

Ajustaremos modelos para datos binarios (ligas: logit, probit, y cloglog) en donde se incluyen como covariables a Insecticide y lnD (lnD = ln(Deposit)), así como su interacción. Se calcularon los tres modelos con interacciones y se muestran en el siguiente Cuadro. De acuerdo con el criterio AIC el modelo más adecuado es el de la liga probit, cuyo AIC fue de 789.28 (el del logit de 789.44 y cloglog de 800.46). Los términos de las interacciones no son significativas para los tres modelos (no se rechaza la hipótesis nula de que los coeficientes son cero), mientras que para el intercepto, InsecticideC y lnD sí se rechaza la hipótesis nula. Esto sugiere que podría ser más adecuado el modelo reducido.

Adicionalmente, se calcularon los tres modelos (ligas logit, probit y cloglog) reducidos, sin las interacciones Insecticide-lnD. Todos tienen un menor AIC, en particular el modelo probit. Se puede observar que en estos casos incluso InsecticideB podría ser estadísticamente significativo si consideramos un nivel de significancia estadística del 10%. Si consideramos el modelo reducido, el modelo probit tiene un mejor desempeño por su AIC y por ser más parsimonioso, con componente lineal o sistemático $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 = -2.623 + 0.209 InsecticideB + 1.672 InsecticideC + 1.690 lnD$.

La prueba de hipótesis global con la chi-cuadrada del modelo **probit reducido** muestra un valor Chisq de 264.5619 y un p-value muy pequeño (Pr(>Chisq): 4.633875e-57), mucho menor a 0.05, es decir se rechaza la hipótesis nula, por lo que podríamos proceder con el análisis de los supuestos del modelo. Antes de continuar, revisaremos en el siguiente inciso algunos modelos que incluyan $(lnD)^2$, y veremos si tienen menor AIC.

iii) Ajuste modelos para datos binarios 2

A continuación incluiremos, adicional a los términos de las covariables anteriores, a la interacción de Insecticide con el término cuadrádico (lnD)^2. Para las ligas logit y probit, ninguna de las intersecciones con lnD y (lnD)^2 rechazan la

Table 1:

	Dependent variable:					
	died					
	logistic	*probit*	*glm: binomial link = cloglog*	*logistic*	*probit*	*glm: binomial link = cloglog*
	(1)	(2)	(3)	(4)	(5)	(6)
InsecticideB	0.188	0.105	0.260	0.349*	0.209*	0.249*
	(0.722)	(0.400)	(0.530)	(0.206)	(0.120)	(0.135)
InsecticideC	2.110***	1.505***	2.350***	2.840***	1.672***	1.706***
	(0.790)	(0.433)	(0.485)	(0.254)	(0.141)	(0.151)
lnD	2.727***	1.634***	1.861***	2.887***	1.690***	1.714***
	(0.349)	(0.194)	(0.234)	(0.224)	(0.122)	(0.134)
InsecticideB:lnD	0.111	0.072	−0.004			
	(0.487)	(0.270)	(0.319)			
InsecticideC:lnD	0.661	0.137	−0.486			
	(0.671)	(0.347)	(0.327)			
Constant	−4.231***	−2.543***	−3.377***	−4.461***	−2.623***	−3.138***
	(0.524)	(0.289)	(0.392)	(0.356)	(0.194)	(0.238)
Observations	862	862	862	862	862	862
Log Likelihood	−388.721	−388.640	−394.229	−389.246	−388.727	−395.786
Akaike Inf. Crit.	789.443	789.280	800.458	786.491	785.454	799.571
Note:						*p<0.1; **p<0.05; ***p<0.01

hipótesis nula, es decir ninguna aparece estadísticamente significativa porque el p-value asociado es mayor a 0.05. Para el caso del cloglog, la única intersección estadísticamente significativa al 5% de significancia estadística es InsecticideC:lnD. En los tres modelos se rechaza la hipótesis nula para el intercepto, InsecticideC, lnD y lnD2. Los AIC son 786.61, 786.92 y 786.06 para los modelos con liga logit, probit y cloglog, respectivamente, lo que indica que el mejor modelo por el criterio AIC es el de la liga cloglog.

Adicionalmente, se procedió a hacer un modelo reducido con sólo efectos principales, sin estas interacciones y el resultado es que hay menores AIC para los tres modelos considerando las variables explicativas Insecticide, lnD y lnD2, sin las interacciones. Por ejemplo, el menor AIC es de 780.01 para el caso de la liga probit, con componente lineal o sistemático $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 = -3.920 + 0.195 InsecticideB + 1.701 InsecticideC + 3.775 lnD - 0.750(lnD)^2$.

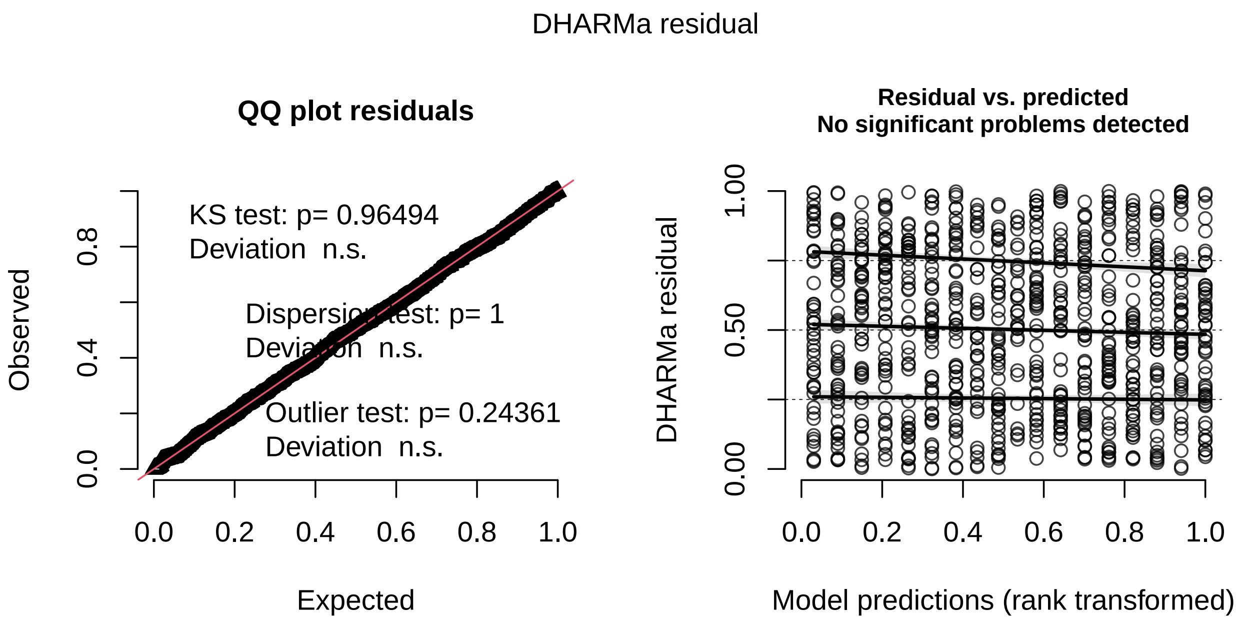
La prueba de hipótesis global con la chi-cuadrada del modelo **probit reducido** muestra un valor Chisq de 254.2325 y un p-value muy pequeño (Pr(>Chisq): 7.974736e-54), mucho menor a 0.05, es decir se rechaza la hipótesis nula, por lo que podemos proceder con el análisis de los supuestos de este modelo reducido más sencillo con el menor AIC.

Se puede notar que una ventaja de introducir el componente $(lnD)^2$ es que los AIC disminuyeron, por lo que nos quedamos con este modelo probit reducido, para los análisis subsecuentes.

En la prueba de normalidad `Lilliefors (Kolmogorov-Smirnov) normality test` tenemos que el p-value es de 0.273946359109463, por lo que no se rechaza la hipótesis nula de normalidad. Por otra parte, pa la prueba de normalidad de `Shapiro-Wilk normality test` el p-value es de 0.331984993533857, lo que también no rechaza la hipótesis nula de normalidad. Esto se observa en la siguiente Gráfica.

Table 2:

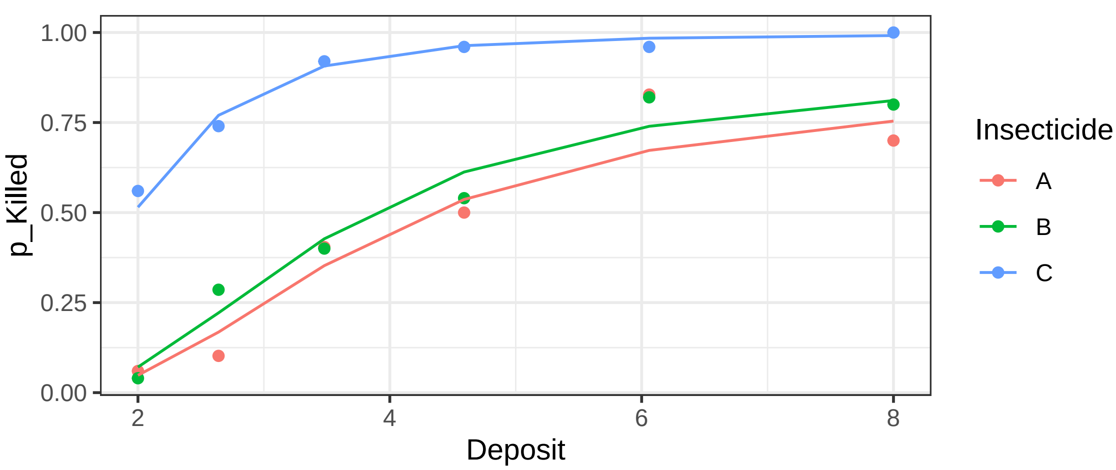
	Dependent variable:					
	died					
	logistic	*probit*	*glm: binomial link = cloglog*	*logistic*	*probit*	*glm: binomial link = cloglog*
	(1)	(2)	(3)	(4)	(5)	(6)
InsecticideB	2.013	0.679	1.973	0.325	0.195	0.221*
	(2.589)	(1.361)	(2.085)	(0.204)	(0.120)	(0.133)
InsecticideC	6.150**	2.934**	6.139***	2.976***	1.701***	1.663***
	(2.684)	(1.388)	(1.825)	(0.271)	(0.145)	(0.152)
lnD	9.085***	4.717***	8.599***	6.813***	3.775***	4.117***
	(2.778)	(1.474)	(2.173)	(1.408)	(0.782)	(0.857)
lnD2	−2.167**	−1.066**	−2.198***	−1.407***	−0.750***	−0.844***
	(0.918)	(0.499)	(0.691)	(0.491)	(0.276)	(0.295)
InsecticideB:lnD	−2.479	−0.773	−2.376			
	(3.663)	(1.982)	(2.790)			
InsecticideC:lnD	−5.238	−1.872	−5.572**			
	(4.300)	(2.198)	(2.530)			
InsecticideB:lnD2	0.839	0.277	0.748			
	(1.231)	(0.678)	(0.895)			
InsecticideC:lnD2	1.971	0.628	1.558*			
	(1.656)	(0.817)	(0.843)			
Constant	−8.512***	−4.560***	−8.123***	−6.946***	−3.920***	−4.647***
	(1.993)	(1.022)	(1.637)	(0.967)	(0.524)	(0.594)
Observations	862	862	862	862	862	862
Log Likelihood	−384.307	−384.460	−384.028	−385.098	−385.006	−391.604
Akaike Inf. Crit.	786.613	786.919	786.055	780.196	780.011	793.208
Note:						*p<0.1; **p<0.05; ***p<0.01



La regla de dedo para verificar el **parámetro de dispersión** de 1, con la devianza de residuales entre grados de libertad, muestra un valor de 0.8984962, lo cual se acerca a 1.

iv) Modelo adecuado. Comparaciones, probabilidades y prueba de hipótesis.

La siguiente gráfica de dispersión muestra en el eje x la dosis del insecticida (Deposit) y en el eje y la proporción de insectos muertos observados (se generó la variable p_Killed) para cada combinación dosis-insecticida (Deposit-Insecticide), distinguiendo el insecticida asociado por colores. Adicionalmente se agregaron las curvas con las probabilidades obtenidas con el modelo probit para cada dosis e insecticida. Con el modelo se obtuvieron probabilidades muy cercanas a las proporciones o tasas de mortalidad observadas, especialmente para el insecticida C.



A continuación se muestra un cuadro de la dosis mínima para cada insecticida con la que se puede indicar que el 70% de los insectos se muere. Para ello recordemos que $\Phi^{-1}(0.7) = \beta_0 + \beta_3 lnD + \beta_4 (lnD)^2$, $\Phi^{-1}(0.7) = \beta_0 + \beta_1 InsecticidaB + \beta_3 lnD + \beta_4 (lnD)^2$ y $\Phi^{-1}(0.7) = \beta_0 + \beta_2 InsecticidaC + \beta_3 lnD + \beta_4 (lnD)^2$, por lo que resolviendo para cada insecticida, se obtienen los respectivos valores de lnD y por lo tanto de D que es la dósis en mg (Deposit). Es decir, para A, resolveremos $\beta_4 (lnD)^2 + \beta_3 lnD + (\beta_0 - \Phi^{-1}(0.7))$, para B $\beta_4 (lnD)^2 + \beta_3 lnD + (\beta_0 + \beta_1 - \Phi^{-1}(0.7))$ y para C $\beta_4 (lnD)^2 + \beta_3 lnD + (\beta_0 + \beta_2 - \Phi^{-1}(0.7))$.

Como se observa en la Gráfica anterior, el insecticida C es mejor, pues con menores dosis se tienen mayor probabilidad de muerte que A y B según el modelo probit. Además, como se mostró en el cuadro anterior, se encontró que la menor dósis mínima con la que el 70% se muere es para el insecticida C. A continuación mostramos una prueba de hipótesis

Insecticida	A	B	C
Dósis Mínima	6.5380798	5.4705365	2.4129177

Table 3:

que comprueba esto. Planteamos entonces que $\beta_0 + \beta_2 InsecticidaC + \beta_3 lnD + \beta_4 (lnD)^2 > \beta_0 + \beta_3 lnD + \beta_4 (lnD)^2$ y $\beta_0 + \beta_2 InsecticidaC + \beta_3 lnD + \beta_4 (lnD)^2 > \beta_0 + \beta_1 InsecticidaB + \beta_3 lnD + \beta_4 (lnD)^2$, de donde obtenemos la hipótesis nula $H_0 : \beta_2 InsecticidaC < 0$ o $\beta_2 InsecticidaC < \beta_1 InsecticidaB$ y la hipótesis alternativa $H_a : \beta_2 InsecticidaC > 0$ y $\beta_2 InsecticidaC > \beta_1 InsecticidaB$.

Resultado: Chisq: 152.8355 y p-value: 6.489137e-34. Lo que rechaza la hipótesis nula, es decir no hay suficiente evidencia para asegurar que el insecticida C tenga menor efectividad que A y B.

A continuación se muestra la prueba de hipótesis que muestra si A y B tienen un desempeño similar. En este caso planteamos que $\beta_0 + \beta_3 lnD + \beta_4 (lnD)^2 = \beta_0 + \beta_1 InsecticidaB + \beta_3 lnD + \beta_4 (lnD)^2$ de donde tenemos la prueba de hipótesis $H_0 : \beta_1 InsecticidaB = 0$ y la alternativa $H_a : \beta_1 InsecticidaB \neq 0$.

Resultado: Chisq: 2.652954 y p-value: 0.1033576. Lo que no rechaza la hipótesis nula, es decir no hay suficiente evidencia para rechazar que el insecticida A tenga el mismo desempeño que B.