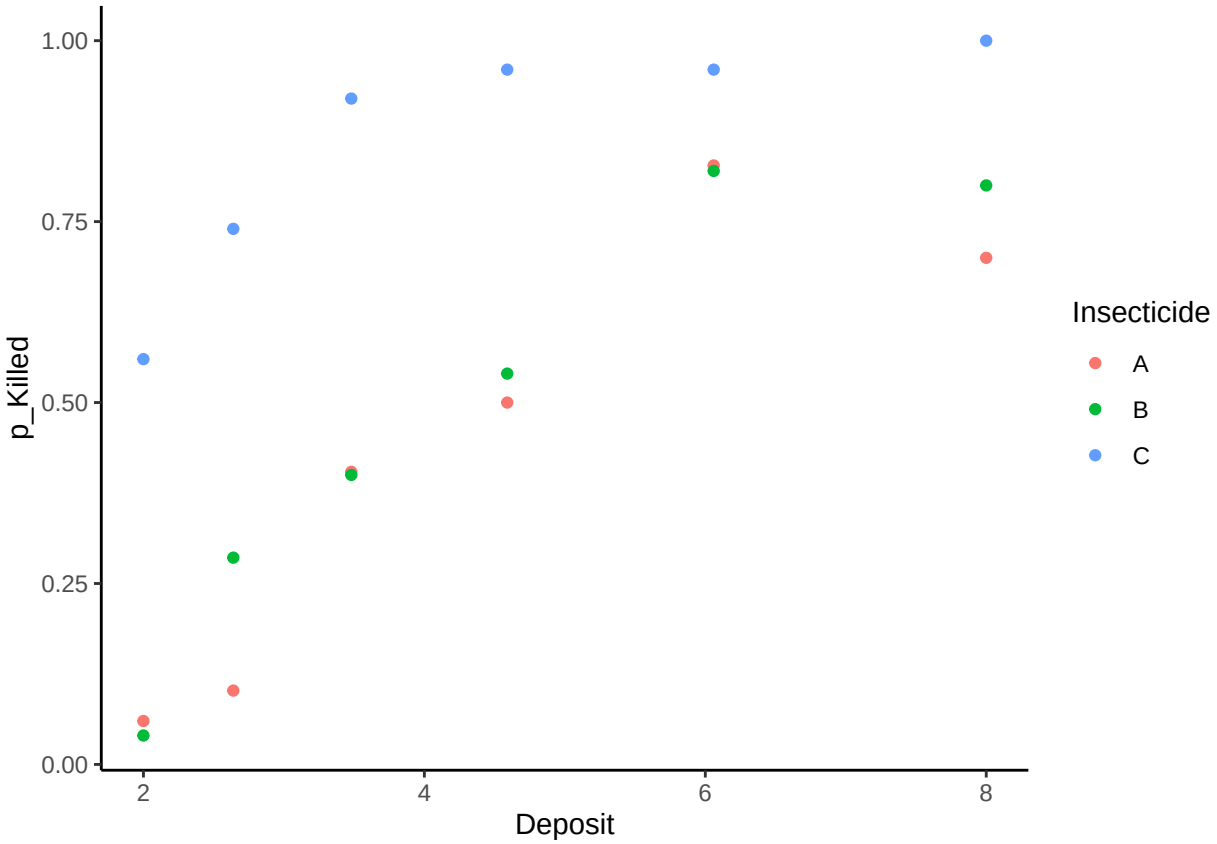
3. Modelos lineales generalizados para datos binarios

La base de datos Preg3B.csv contiene información sobre 862 insectos que fueron expuestos a diferentes dosis (Deposit en mg) 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.

Notar que aquí el evento de interés es si el insecto muere o no (died). Además, dado que se tienen varios insectos para diferentes valores de dosis e insecticida, es posible realizar gráficas que ayudan a entender lo que se está modelando; de hecho la base de datos está en un formato agregado.

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 (se generó la variable 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, para los datos desagrupados generados con las 862 filas de observaciones. 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), mientras que para el intercepto, InsecticideC y lnD 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 pos ser más simple en la interpretación.

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

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 del modelo. Sin embargo, antes de continuar, revisaremos en el siguiente inciso algunos otros modelos mejores por su AIC y simplicidad interpretativa o parsimonia.

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 las intersecciones con lnD y (lnD)^2 rechazan la 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, el siguiente Cuadro muestra los resultados.

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

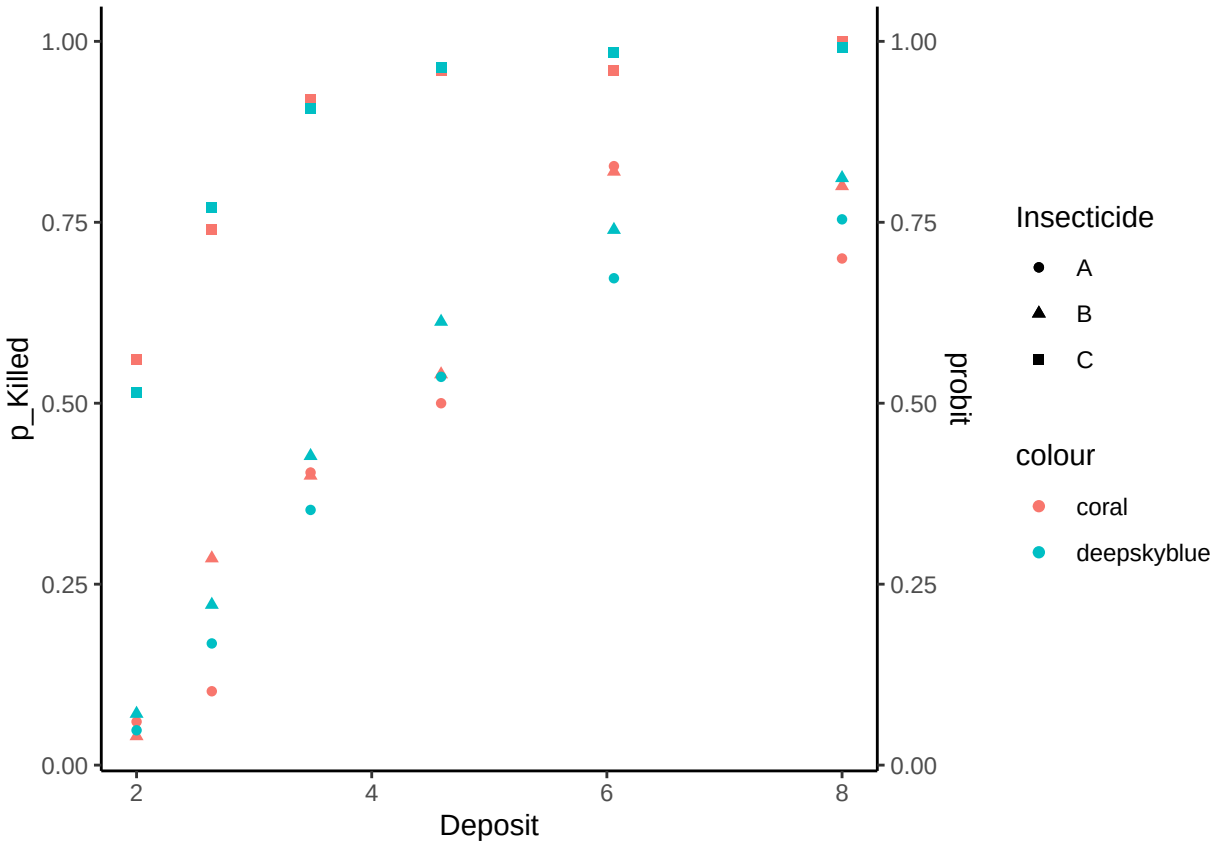
La prueba de hipótesis global con la chi-cuadrada del modelo **probit reducido** muestra un valor Chisq de 254.2325 y un p-value muy pequeño (Pr(>Chisq): 7.974736e-54), mucho menor a 0.05, es decir se rechaza la hipótesis nula, por lo que podemos proceder con el análisis de este modelo reducido más sencillo, de died en función de Insecticide (A,B y C, con A como de referencia), lnD y lnD2.

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

Las estimaciones de las probabilidades de que los insectos mueran (prob_s), dados el insecticida, lnD y lnD2, son las siguientes para el modelo probit reducido. Se puede observar que para el insecticida A, la probabilidad de muerte mayor al 70% se da hasta la dosis de entre 6 y 8 miligramos, mientras que para el insecticida B se da entre 4.59 a 6 mg, y finalmente para el insecticida C, se da entre los 2 a 2.64 miligramos.

Insecticide	lnD	lnD2	prob_s	Deposit
A	0.6931472	0.4804530	0.0481131	2.00
A	0.9707789	0.9424117	0.1680800	2.64
A	1.2470323	1.5550895	0.3525728	3.48
A	1.5238800	2.3222100	0.5364179	4.59
A	1.8017098	3.2461580	0.6726657	6.06
A	2.0794415	4.3240700	0.7540544	8.00
B	0.6931472	0.4804530	0.0709560	2.00
B	0.9707789	0.9424117	0.2215236	2.64
B	1.2470323	1.5550895	0.4271397	3.48
B	1.5238800	2.3222100	0.6126130	4.59
B	1.8017098	3.2461580	0.7395662	6.06
B	2.0794415	4.3240700	0.8111188	8.00
C	0.6931472	0.4804530	0.5149154	2.00
C	0.9707789	0.9424117	0.7700607	2.64
C	1.2470323	1.5550895	0.9069897	3.48
C	1.5238800	2.3222100	0.9634528	4.59
C	1.8017098	3.2461580	0.9841476	6.06
C	2.0794415	4.3240700	0.9915328	8.00

La siguiente gráfica muestra en el eje x la dosis del insecticida (Deposit) y en el eje y izquierdo la proporción de insectos muertos observados (se generó la variable p_Killed) para cada combinación dosis-insecticida (Deposit-Insecticide), distinguiendo con una figura distinta el insecticida asociado. Adicionalmente se agrega un eje derecho, con las probabilidades obtenidas con el modelo probit para cada dosis e insecticida. El color **coral** indica los datos de la tasa de mortalidad (p_Killed) y el color **deepskyblue** indica los valores de las probabilidades calculadas con el modelo (probit). Con el modelo se obtuvieron probabilidades muy cercanas a las proporciones o tasas de mortalidad observadas, especialmente para el insecticida C (cuadros).


1.00
0.75
0.50
0.25
0.00
p_Killed
probit
2
4
6
8
Deposit
Insecticide
A
B
C
colour
coral
deepskyblue