Facultad de Ciencias
UNAM

Seminario de Estadística

Tarea 1B

Introducción a los modelos lineales generalizados

Enríquez Hernández Leobardo
Tlahuiz Tenorio Giovanni Saúl

7 de abril de 2024

Índice

1. Regresión lineal múltiple.

i) Modelo de RLM reducido para E(bpsystol; bmi, sex, age) con datos originales.

Para analizar si existe una asociación entre la presión arterial sistólica (bpsystol) como variable dependiente y el índice de masa corporal (bmi), ajustaremos un modelo de regresión lineal múltiple, considerando el sexo (sex: 1-hombre y 2-mujer con nivel de referencia hombre) y la edad (age) de los pacientes. Para ello usaremos la base de datos `reg1B.csv` con 295 pacientes, 142 hombres y 153 mujeres, de entre 20 y 74 años. En el cuadro de MODELOS se muestran los resultados del Modelo 1 planteado, sin pretratamiento de los datos.

La prueba global F muestra un p-value menor a 0.05, por lo que rechazamos la hipótesis nula de que los parámetros estimados son cero, es decir, podemos decir que al menos un coeficiente estimado es distinto de cero, por lo que el modelo es estadísticamente significativo al nivel de confianza del 95 %. Las pruebas individuales también rechazan la hipótesis nula con la preba $t - student$, es decir, todos los coeficientes son significativos al 5 %, pues se rechaza la hipótesis nula de que en lo individual sean iguales a cero.

En el siguiente Cuadro se pueden observar las pruebas de Shapiro-Wilk, Breusch-Pagan y Durbin-Watson para el Modelo 1, que plantean la hipótesis nulas de normalidad, homoscedasticidad y no autocorrelación, respectivamente. Se concluye que el Modelo 1 presenta no autocorrelación y homoscedasticidad, sin embargo no presenta normalidad de los errores. Por lo que tendremos que hacer algunos ajustes al modelo, con algunos tratamientos a las variables.

	1
Normality (Shapiro-Wilk)	0.001
Homoscedasticity (Breusch-Pagan)	0.095
Autocorrelation of residuals (Durbin-Watson)	0.981

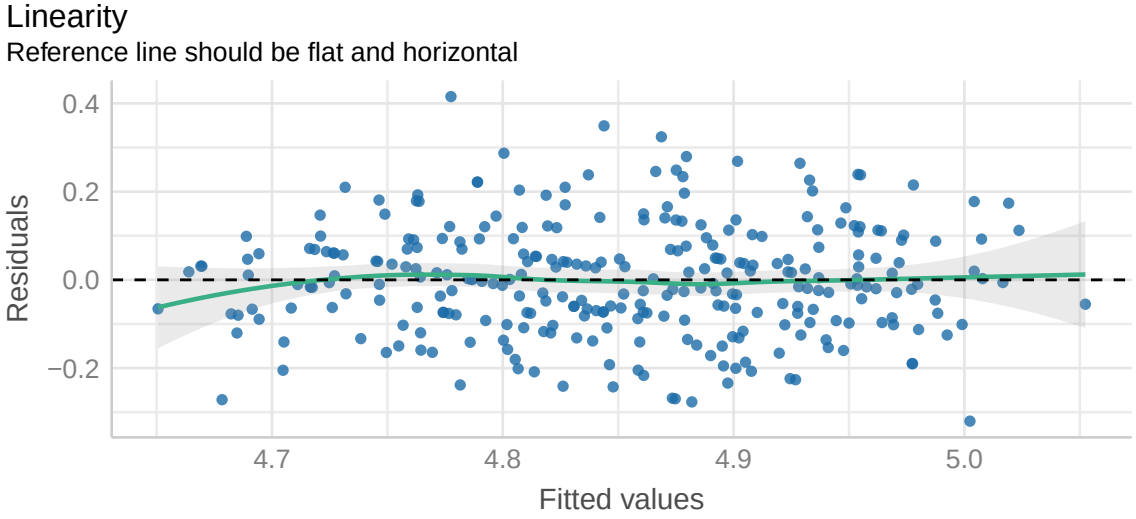
ii) Modelo adecuado con transformación de datos.

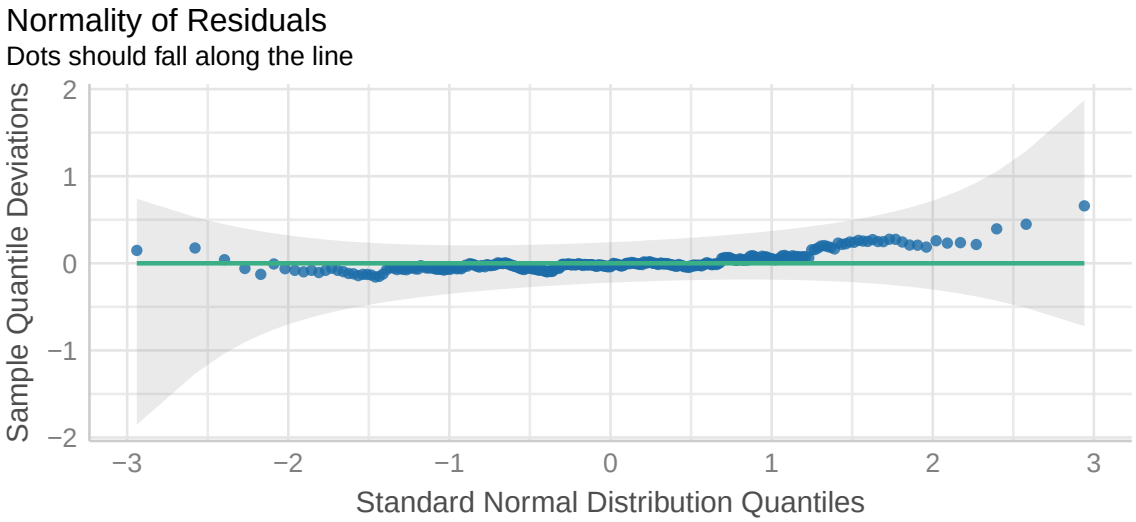
Como tenemos un problema con la normalidad, procederemos a hacer primero una transformación a la variable dependiente, probaremos con una transformación más usual que es la logarítmica, la cual se puede interpretar más fácilmente. Por simplicidad no consideraremos en el Modelo 2 interacciones entre las variables y se propone una transformación Box Cox logarítimica de la variable dependiente. Para este Modelo 2, se observa en el Cuadro de MODELOS que la prueba global F rechaza la hipótesis nula, por lo que al menos un coeficiente estimado es distinto de cero, y las pruebas $t - student$ individuales de los coeficientes estimados también rechazan las hipótesis nulas analizados individualmente. Notemos que *** implica que se rechaza la hipótesis nula incluso con un nivel de confianza del 99 %, el p-value es menor a 0.01. Además al comparar los AIC, tenemos para el Modelo 1 es de 2507,213 y para el Modelo 2, considerando que la transformación a la variable dependiente fue logarítmica, es de 2485.7472039 lo cual es menor, esto favorece la alección del Modelo 2.

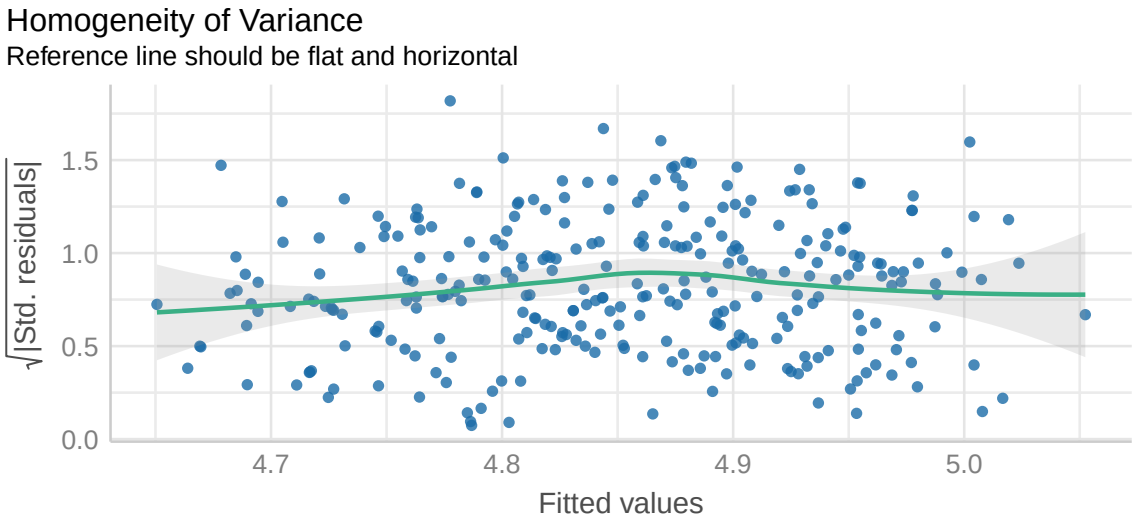
Para poder tener una interpretación válida de los coeficientes, veremos que el Modelo 2 cumple con los supuestos del modelo de regresión lineal. Primero se mostrarán algunas gráficas. La Gráfica **Residuals vs Fitted Values**, se utiliza para comprobar los supuestos de relación lineal, una línea horizontal, sin patrones distintos, es indicación de una relación lineal, lo que es bueno en nuestro caso. La Gráfica **Sample Q Deviation vs Standard Normal Distribution Q**, se utiliza para examinar si los residuos se distribuyen normalmente, es bueno que los puntos residuales sigan la línea recta, en nuestro caso parece que todo se ajusta bien, pues tenemos muchos valores que siguen la linea. La Gráfica **Scale-Location: Sqrt(|Std. Residuals|) vs Fitted values**, se utiliza para comprobar la homogeneidad de la varianza de los residuos (homoscedasticidad), la línea horizontal con puntos igualmente distribuidos es una buena indicación de homocedasticidad, este es el caso en nuestro modelo, donde no tenemos un problema de heterocedasticidad. La Gráfica **Std. Residuals vs Leverage**, se utiliza para identificar casos de valores influyentes, es decir, valores extremos que podrían influir en los resultados de la regresión cuando se incluyen o excluyen del análisis, al parecer ningún valor sale de la distancia de Cook.

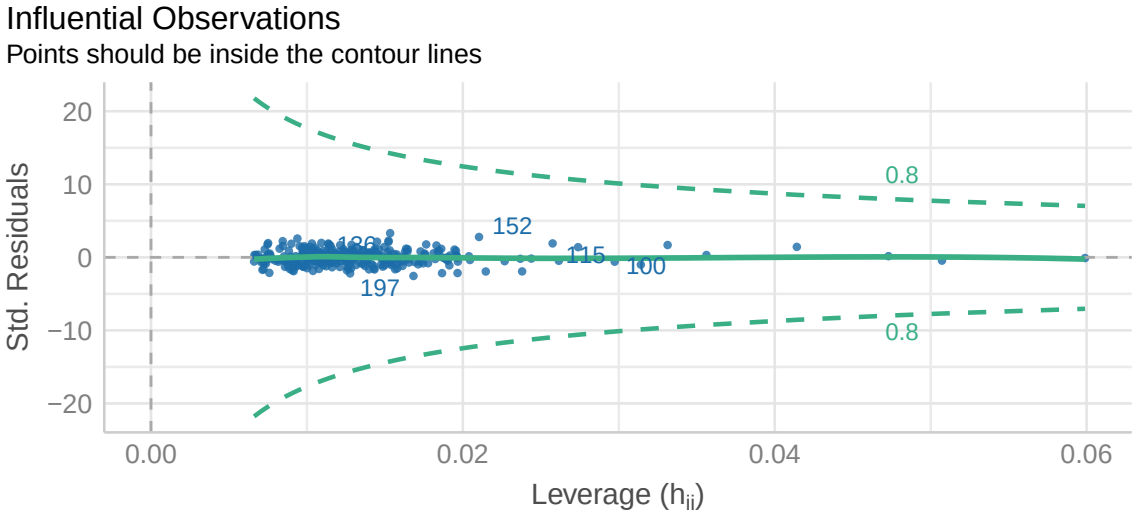
MODELOS

	Dependent variable:	
	bpsystol	I(log(bpsystol))
	(1)	(2)
bmi	1.208***	0.009***
	(0.202)	(0.002)
sex2	−5.664***	−0.049***
	(1.964)	(0.015)
age	0.484***	0.004***
	(0.059)	(0.0004)
Constant	78.496***	4.461***
	(5.510)	(0.042)
Observations	295	295
R^2	0.310	0.321
Adjusted R^2	0.302	0.314
Residual Std. Error (df = 291)	16.784	0.127
F Statistic (df = 3; 291)	43.497***	45.922***
AIC:	**2507.213**	−376,2099(**2485.747**)
Note:		*p<0.1; **p<0.05; ***p<0.01









En el siguiente Cuadro, se muestra las pruebas de Shapiro-Wilk, Breusch-Pagan y Durbin-Watson para el Modelo 2, que plantean la hipótesis nulas de normalidad, homoscedasticidad y no autocorrelación, respectivamente. En todos los casos no hay evidencia suficiente para rechazar las hipótesis nulas.

	1
Normality (Shapiro-Wilk)	0.596
Homoscedasticity (Breusch-Pagan)	0.254
Autocorrelation of residuals (Durbin-Watson)	0.975

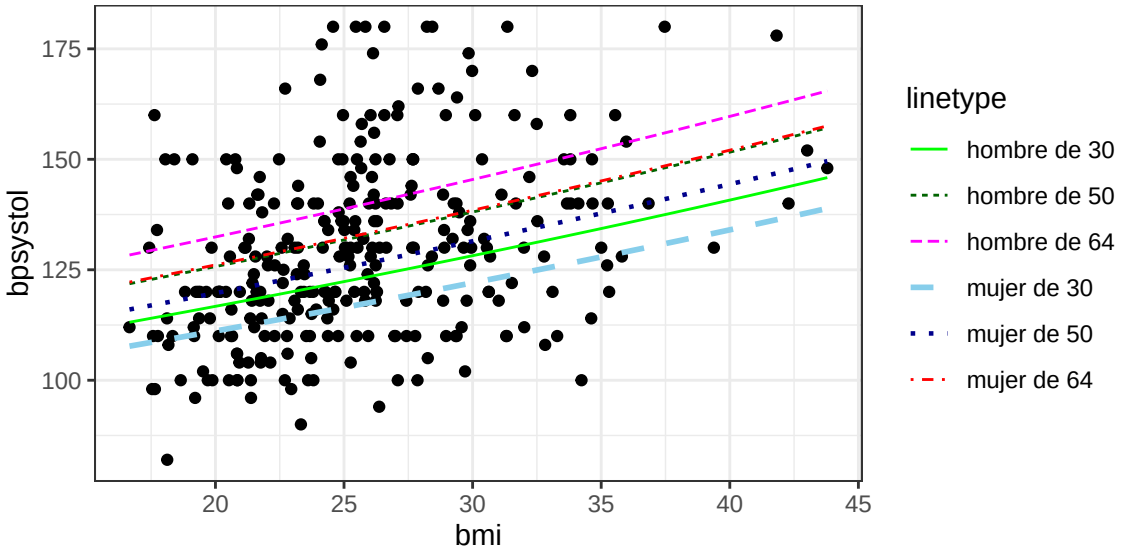
iii) Asociación entre masa corporal y presión arterial sistólica.

Se puede concluir que para una persona de cierta edad y sexo, tener un índice de masa corporal alto se asocia con una alta presión arterial sistólica. Tomando en cuanta en el Cuadro anterior que se rechaza la hipótesis nula $H_0 : \beta_1 = 0$ contra la anternativa de que $H_a : \beta_1 \neq 0$, a continuación se plantea la prueba de hipótesis con dirección, en donde la hipótesis nula es $H_0 : \beta_1 < 0$ y la alternativa $H_a : \beta_1 > 0$.

El resultado de la prueba con dirección `Simultaneous Tests for General Linear Hypotheses` con el ajuste $lm(formula = I(log(bpsystol)) \sim bmi + sex + age, data = datos)$ muestra un p-value de $1{,}2e-09$, lo cual rechaza la hipótesis nula planteada. Por lo tanto, para una persona de cierta edad y sexo, tener un índice de masa corporal alto se asocia con una alta presión arterial sistólica.

iv) Gráfica resumen con la estimación puntual de la relación bpsystol y bmi.

A continuación presentaremos una gráfica resumen con la estimación puntual asociada a la relación entre bpsystol y bmi. Para esto consideremos sólo tres posibles edades: 30, 50 y 64, así como la diferenciación entre mujeres y hombres. El comportamiento en general es que los hombres tienden a tener una mayor presión arterial sistólica, comparado con las mujeres. En todos los casos al aumentar la masa corporal, la presión arterial sistólica incrementa tanto para hombres como para mujeres. Además podemos observar que a mayor edad, es mayor la presión arterial sistólica tanto para hombres como para mujeres.



2. Modelos lineales generalizados para datos continuos

Consideraremos la base de datos `Preg1B.csv` con información sobre 295 pacientes seleccionados de forma aleatoria. Se desea analizar si existe una asociación entre la presión arterial sistólica (bpsystol) y el índice de masa corporal (bmi), considerando el sexo (sex: 1-hombre, 2-mujer, con hombre como referencia) y la edad (age) de los pacientes.

i) Explorando modelos con variable dependiente contínua.

Para presentar un modelo que parezca adecuado para modelar E(bpsystol; bmi, sex, age), exploramos una malla de los diferentes modelos lineales generalizados comúnmente usados: para el componente aleatorio cuando la variable dependiente es continua exploramos las distribuciones normal, gamma, e inversa gaussiana; empleamos distintas funciones ligas tales como la inversa, identidad, logarítmica, y 1/mu^2(solo para IG); y consideramos el componente lineal tanto de potencias (-3, -2.5, ..., 2.5, 3) como de polinomios (grado 1 al 5). Consideramos por simplicidad que no hay interacción entre las covariables del modelo. En el siguiente Cuadro se muestraa el mejor modelo, con el menor AIC de 2484.009 (que coincide con el mejor modelo por su BIC de 2502.443), con la siguiente estructura:

glm(formula = bpsystol ~ age+sex+I(bmi^(1.5)), family = inverse.gaussian(link = identity), data = datos).

Sin embargo, se elige el modelo más simple o parsimonioso sin el exponente de 1,5 para la variable bmi, pues al considerar bmi sin modificación se obtiene un AIC de 2484.1, el cual no parece ser muy diferente a 2484.009. En el siguiente Cuadro se muestra el modelo final elegido.

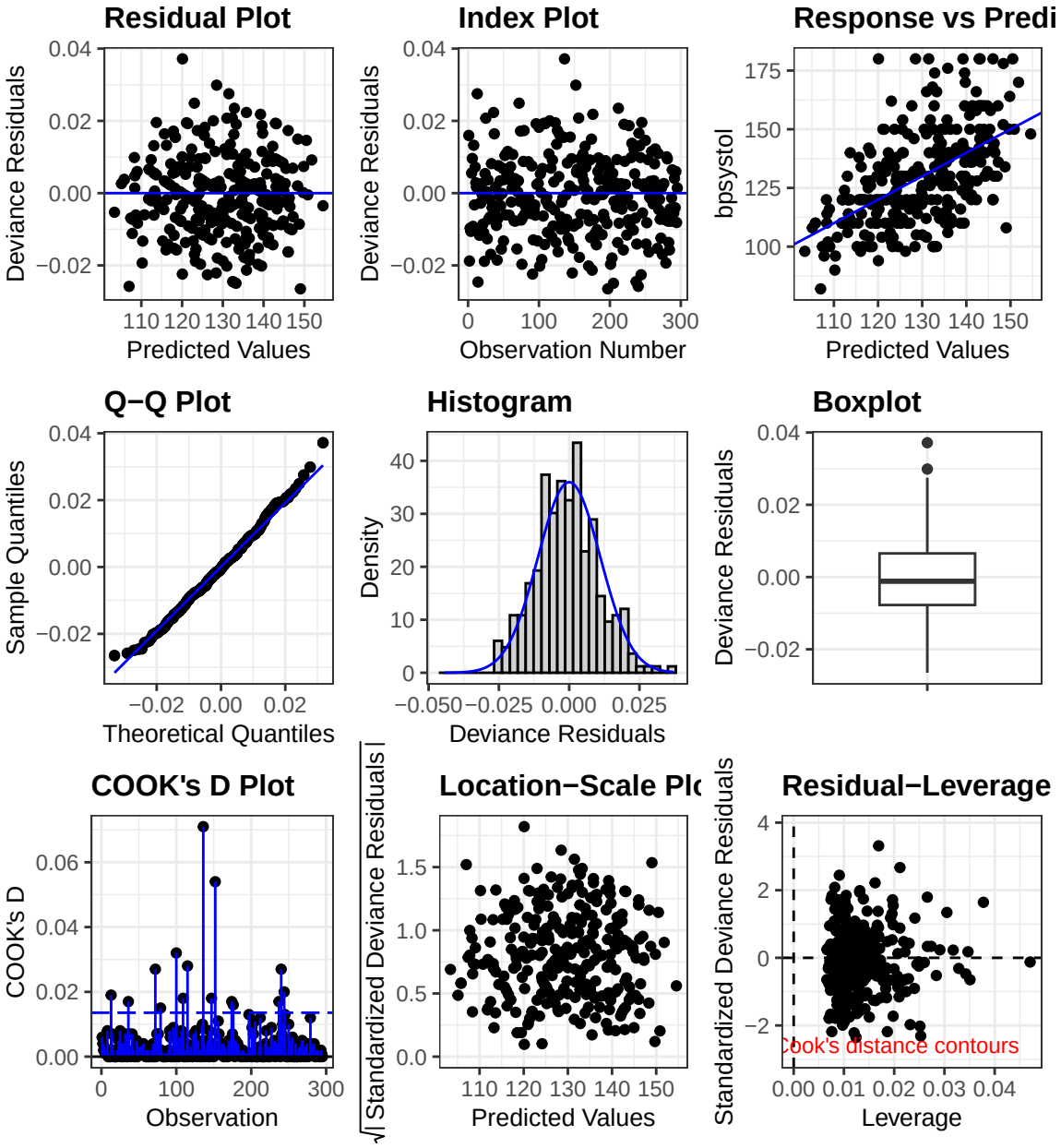
glm(formula = bpsystol ~ age+sex+bmi, family = inverse.gaussian(link = identity), data = datos).

	Dependent variable:	
	bpsystol	
	(1)	(2)
age	0.48671***	0.48269***
	s.e. (0.057)	s.e. (0.057)
	p-value: 1.06e-15	p-value: 1.95e-15
sex2	−7.05833***	−6.88649***
	s.e. (1.908)	s.e. (1.906)
	p-value: 0.000258	p-value: 0.000356
I(bmi^(1.5))	0.15131***	
	s.e. (0.026)	
	p-value: 1.95e-08	
bmi		1.17620***
		s.e. (0.203)
		p-value: 1.68e-08
Constant	90.16891***	80.02163***
	s.e. (3.902)	s.e. (5.254)
	p-value: <2e-16	p-value: <2e-16
Observations	295	295
Log Likelihood	−1,238.004	−1,238.068
Akaike Inf. Crit.	2,484.009	2,484.136
Note:		*p<0.1; **p<0.05; ***p<0.01

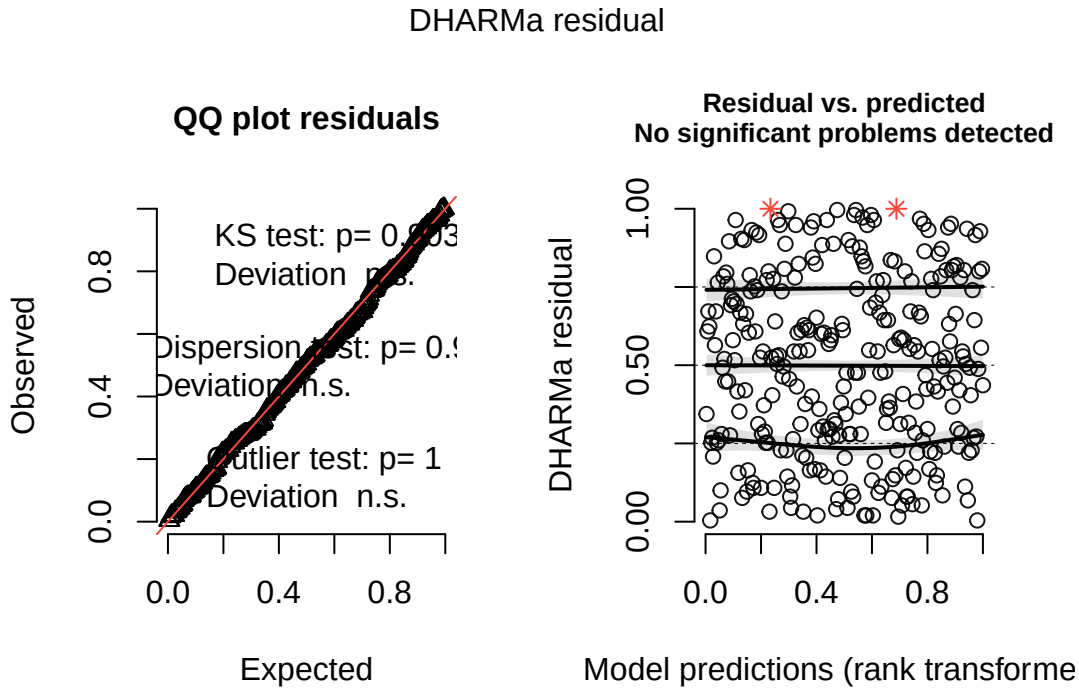
La prueba de hipótesis global con la chi-cuadrada del modelo lineal general **inversa gaussiana con liga identidad** muestra un valor Chisq de 142.2139 y un p-value muy pequeño (Pr(>Chisq): 1.259176e-30), mucho menor a 0.05, es decir se rechaza la hipótesis nula, por lo que podemos proceder con el análisis de los supuestos de este modelo reducido más sencillo.

En la prueba de normalidad `Lilliefors (Kolmogorov-Smirnov) normality test` tenemos que el p-value es de 0.560576574569643, 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 0.44368350946798, lo que también no rechaza la hipótesis nula de normalidad.

En las siguientes gráficas podemos observar en **Residual Plot** que se conserva la linealidad y varianza constante. En **Q-Q Plot** y **Histogram** se observa un buen comportamiento de la normalidad de los errores. En **Index Plot** no hay patrones relacionados con la forma en que se ordenaron los datos, lo que puede proporcionar información sobre tendencias adicionales en los datos que no se han tenido en cuenta en el modelo, no hay una tendencia obvia en el gráfico. En **Location-Scale Plot** se observa que hay homoscedasticidad. En el **Boxplot** se puden observar algunos aoutliers, sin embargo en **COOK'S D Plot** y en **Residuals-Leverage Plot** parece no haber outliers influyentes.



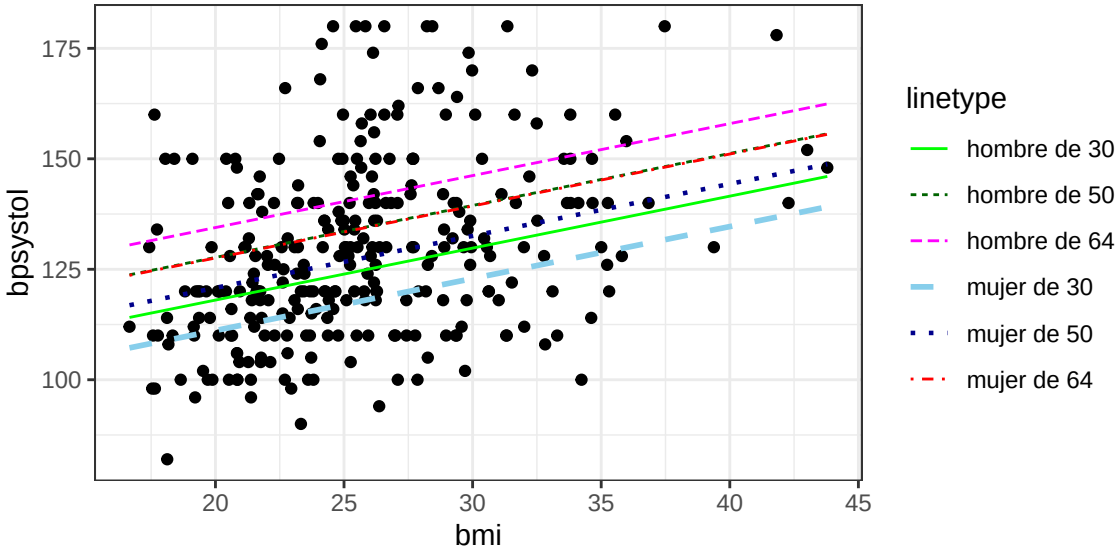
Además en las siguientes gráficas se comprueba las observaciones de las gráficas anteriores.



ii) Asociación entre masa corporal y presión arterial sistólica, y estimación puntual.

En esta sección describiremos la asociación entre masa corporal y presión arterial sistólica y la prueba de hipótesis de esta relación. Dado que lo que buscamos responder es si tener un indice de masa corporal alto se relaciona con tener una presion sistolica alta agregaremos una prueba de hipótesis con dirección, donde la hipótesis nula es $H_0 : \beta_4 < 0$ contra la alternativa $H_0 : \beta_4 > 0$. El p-value asociado a la prueba es de Pr(>z) = 3.22e-09, por lo tanto, rechazamos la hipótesis nula y por lo tanto hay relación asociación entre la masa corporal alta y la presión arterial sistólica alta para una persona de cierta edad y sexo.

Por otra presentaremos una gráfica resumen con la estimación puntual de la relación bpsystol y bmi, considerando edades de 30, 50 y 64, así como la diferenciación entre mujeres y hombres.



iii) Comparativo modelo de regresión lineal múltiple contra modelo lineal generalizado.

En esta sección compararemos el modelo de regresión lineal múltiple del ejercicio anterior (ejercicio 1) contra el modelo lineal generalizado con base en sus AIC (ejercicio 2). Además, compararemos las conclusiones e interpretaciones de ambos modelos, para indicar cuál nos parece más adecuado y fácil de interpretar. El AIC del primer modelo de regresión lineal es de $-376{,}2099$, el cual no es directamente comparable con el AIC del modelo lineal general mlg de $2,484{,}136$, pero haciendo una transformación (ya que la variable dependiente se tranformó en logaritmo) podemos notar que el modelo de mayor AIC es el modelo OLS con un AIC de 2485.7472039. Además, se observa que el modelo más sencillo de interpretar sus coeficientes de manera directa es el mlg, sin tener que hacer transformaciones adicionales a los coeficientes, por lo que nos parece adecuado elegir este como el mejor modelo.

Tomando en cuenta el modelo elegido y habiendo mostrado el cumplimiento de los supuestos del modelo, además de una prueba de hipótesis global satisfactoria, podemos concluir que la relación entre bmi es directa (positiva) con bpsystol, el incremento en una unidad en mbi, incrementa el bpsystol en 1.76 unidades. Por otra parte, el ser mujer, con respecto a ser hombre, tiene una relación inversa (negativa) con bpsystol, es decir, ser mujer disminuye en 6.886 unidades la bpsystol. Con respecto a la edad, un incremento en una unidad de edad, incrementa bpsystol en 0.483 unidades.

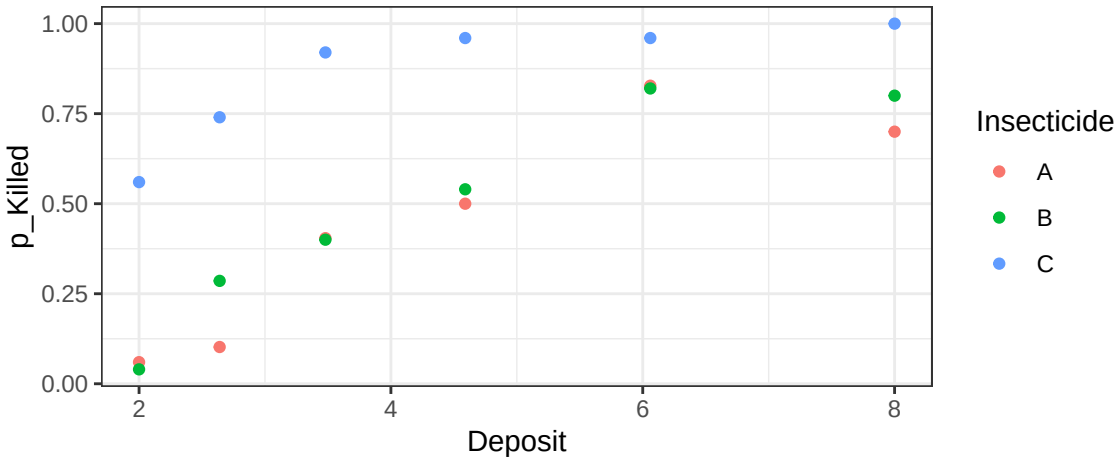
	Dependent variable:	
	I(log(bpsystol))	bpsystol
	OLS	*glm: inverse.gaussian link = identity*
	(1)	(2)
bmi	0.009***	1.176***
	(0.002)	(0.203)
sex2	−0.049***	−6.886***
	(0.015)	(1.906)
age	0.004***	0.483***
	(0.0004)	(0.057)
Constant	4.461***	80.022***
	(0.042)	(5.254)
Observations	295	295
R^2	0.321	
Adjusted R^2	0.314	
Log Likelihood		−1,238.068
Akaike Inf. Crit.	-376.2099 (**2485.747**)	**2,484.136**
Residual Std. Error	0.127 (df = 291)	
F Statistic	45.922*** (df = 3; 291)	
Note:		*p<0.1; **p<0.05; ***p<0.01

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

	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, 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 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 $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 = -3{,}920 + 0{,}195 InsecticideB + 1{,}701 InsecticideC + 3{,}775 lnD - 0{,}750 (lnD)^2$.

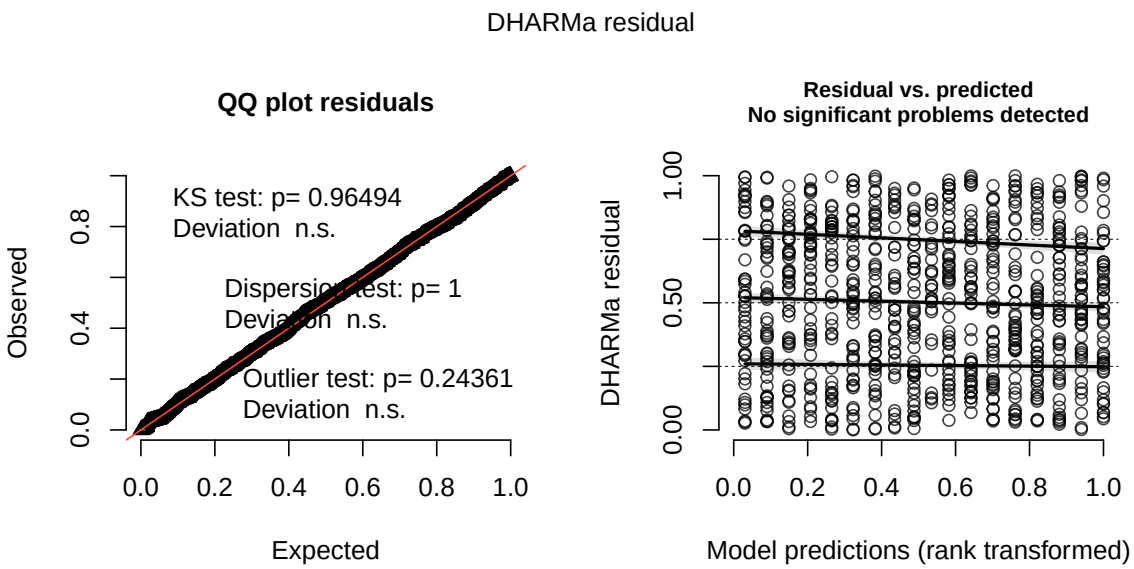
La prueba de hipótesis global con la chi-cuadrada del modelo **probit reducido** muestra un valor Chisq de 254.2325 y un p-value muy pequeño (Pr(>Chisq): 7.974736e-54), mucho menor a 0.05, es decir se rechaza la hipótesis nula, por lo que podemos proceder con el análisis de los supuestos de este modelo 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 0.616389122230475, por lo que no se rechaza la hipótesis nula de normalidad. Por otra parte, pa la

	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

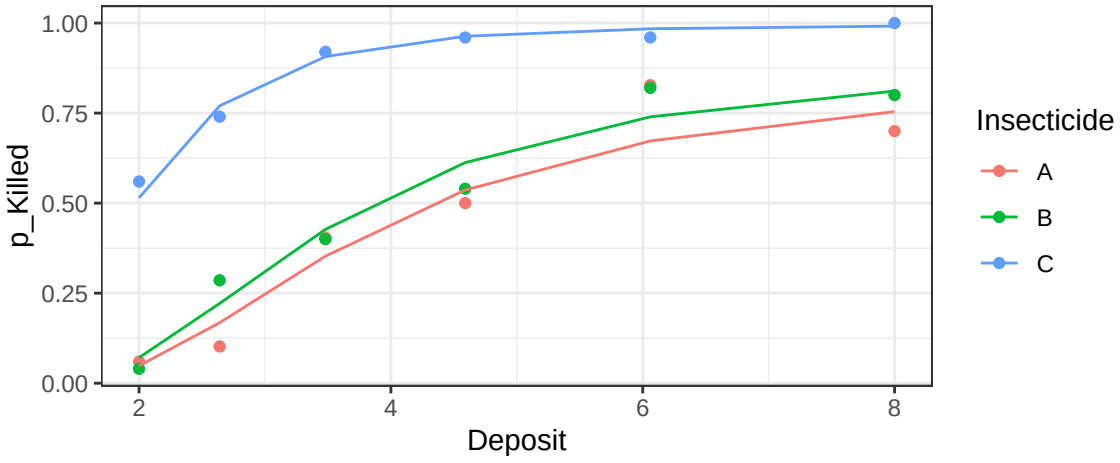
prueba de normalidad de `Shapiro-Wilk normality test` el p-value es de 0.592709734522069, lo que también no rechaza la hipótesis nula de normalidad. Esto se observa en la siguiente Gráfica.



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 $\beta_4(lnD)^2 + \beta_3 lnD + (\beta_0 + \beta_2 - \Phi^{-1}(0{,}7))$.

Insecticida	A	B	C
Dósis Mínima	6.5380798	5.4705365	2.4129177

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 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 $H_a : \beta_1 InsecticidaB \neq 0$.

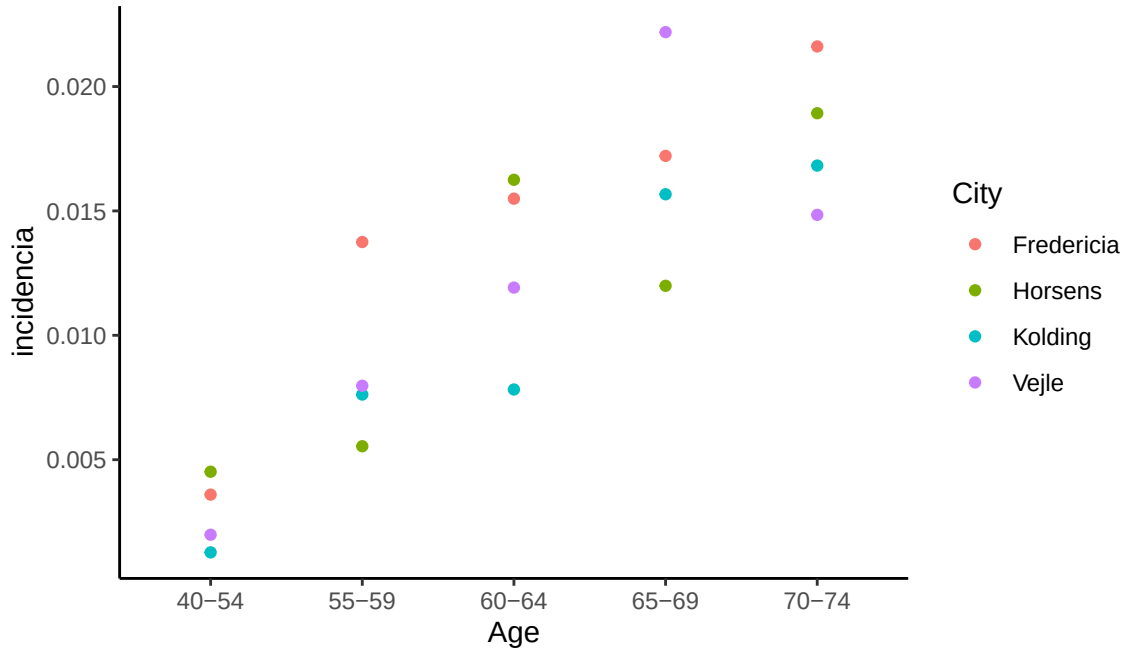
Resultado: Chisq: 2,652954 y p-value: 0,1033576. Lo que no rechaza la hipótesis nula, es decir no hay suficiente evidencia para rechazar que el insecticida A tenga el mismo desempeño que B.

4. Modelos lineales generalizados para datos de conteos

La base de datos Preg4.csv contiene información sobre el número de casos de cáncer de pulmón (Cases) registrados entre 1968 y 1971 en cuatro ciudades de Dinamarca (City). En estos casos se registró también la edad de los pacientes (Age, variable categorizada en 5 grupos). El interés del análisis es estudiar si se puede indicar que a mayor edad existe mayor incidencia de cáncer de pulmón. Notemos que para realizar el análisis la variable de conteos Cases depende de forma inherente de la población de la ciudad (Pop), pues entre más grande la ciudad es mayor el número de casos que se pueden observar; de manera que el estudio se debe enfocar en las tasas de incidencia.

i) Gráfica de dispersión de grupos de edad e incidencia

Podemos apreciar de la siguiente Gráfica presentada que por cada grupo de edad la incidencia en cada ciudad va en aumento, por ejemplo en el grupo de edad de 40-54 la incidencia de cáncer esta por debajo de 0.005 pero conforme avanzan los grupos de edad los niveles aumentan para todas las ciudades.



ii) Distribución Poisson con liga logarítmica y un segundo modelo.

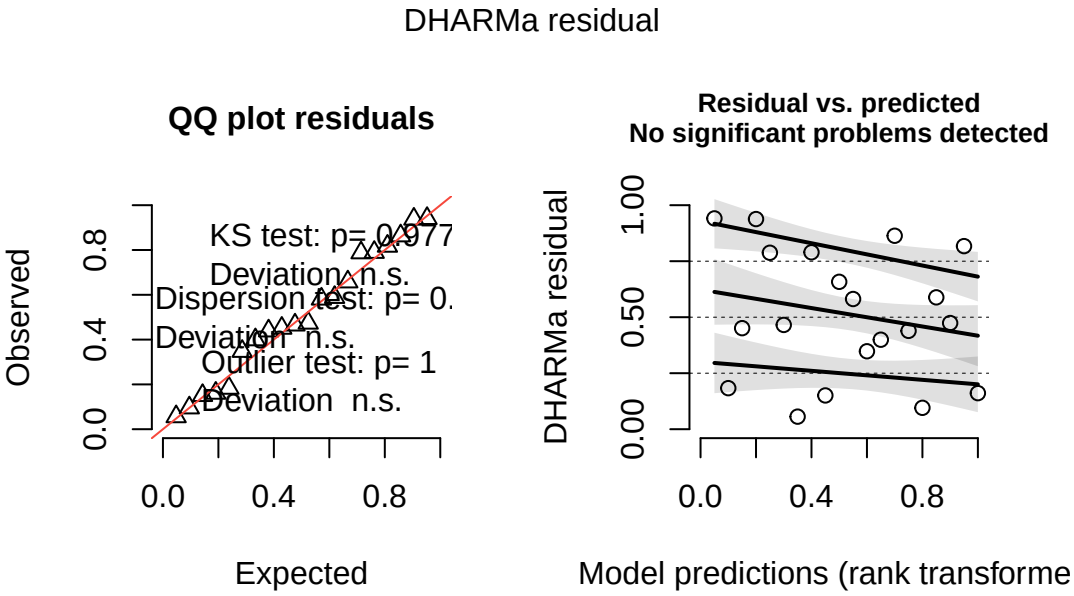
Como primer modelo consideraremos uno con distribución Poisson y función log, ademas de considerar las demás covariables de Age y City con su interacción. Aplicamos el código glm(formula = Cases $\sim$ offset(logPop) + Age * City, family = poisson(link = "log").

El AIC obtenido con este Modelo 1 es de 121.47 y de la regla del dedo para analizar si hay un problema por considerar el parametro de dispersion igual a 1 obtuvimos un valor de $-\infty$, lo cual nos dice que no es un bueno modelo, de todas formas se hizo la verificacion de los supuestos que por cuestión de espacio no se muestra, pero salio muy mal en estos por lo que se decidió no usarse.

De las verificaciones de los supuestos para el primer modelo observamos que tenemos muchos problemas por lo que optaremos por ajustar un segundo modelo, con la diferencia de que solo incluiremos a la covariable Age sin interacción. Usamos el código glm(formula = Cases $\sim$ offset(logPop) + Age, family = poisson(link = "log"), data = data4). Este Modelo 2 nos da un AIC de 108.45.

Vemos que este Modelo 2 tuvimos un valor de 1.131886, muy cercano a 1 con la regla de dedo para analizar si hay un problema por considerar el parametro de dispersion igual a 1, lo cual nos dice que es buen modelo, por lo que procedimos con la verificación de los supuestos del modelo de manera gráfica.

	Dependent variable:	
	Cases	
	(1)	(2)
Age55-59	1.341***	1.082***
	(0.426)	(0.248)
Age60-64	1.461***	1.502***
	(0.426)	(0.231)
Age65-69	1.566***	1.750***
	(0.437)	(0.229)
Age70-74	1.793***	1.847***
	(0.426)	(0.235)
CityHorsens	0.228	
	(0.410)	
CityKolding	−1.038*	
	(0.584)	
CityVejle	−0.595	
	(0.539)	
Age55-59:CityHorsens	−1.137*	
	(0.652)	
Age60-64:CityHorsens	−0.180	
	(0.570)	
Age65-69:CityHorsens	−0.589	
	(0.606)	
Age70-74:CityHorsens	−0.360	
	(0.585)	
Age55-59:CityKolding	0.448	
	(0.746)	
Age60-64:CityKolding	0.355	
	(0.758)	
Age65-69:CityKolding	0.944	
	(0.729)	
Age70-74:CityKolding	0.788	
	(0.737)	
Age55-59:CityVejle	0.050	
	(0.724)	
Age60-64:CityVejle	0.332	
	(0.694)	
Age65-69:CityVejle	0.849	
	(0.680)	
Age70-74:CityVejle	0.219	
	(0.712)	
Constant	−5.628***	−5.862***
	(0.302)	(0.174)
Observations	20	20
Log Likelihood	−40.736	−49.226
Akaike Inf. Crit.	121.473	108.451
Note:	*p<0.1; **p<0.05; ***p<0.01	



Lo que sigue sera hacer una prueba anova en la que compararemos ambos modelo y decidir si se puede usar el segundo modelo.

```
## Analysis of Deviance Table
##
## Model 1: Cases ~ offset(logPop) + Age * City
## Model 2: Cases ~ offset(logPop) + Age
##   Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
## 1         0      0.000
## 2        15     16.978 -15  -16.978   0.3202
```

Como obtuvimos un p-value mayor que 0.05 no tenemos evidencia suficiente para rechazar la hipótesis nula, por lo que concluimos que no se tiene una mejora significativa tomando mas variables y su interacción.Adicional mente tenemos que el AIC del modelo con Age como única covariable es menor que el que incluye la interacción por lo que tenemos las herramientas suficientes para descartar dicho modelo.

iii) Modelo binomial negativo, comparación e intervalo de confianza simultáneo.

Planteando un modelo binomial negativo con el código glm.nb(Cases ~ offset(logPop)+ Age , data = data4, link = "log"), tenemos el resultado del siguiente Cuadro, con un AIC de 110.45.

Como ultimo criterio para escoger modelo final compararemos los resultados tanto del AIC como del BIC para los modelos poisson con covariable edad (Modelo 2) contra el binomial negativo (Modelo 3), donde es claro que el mejor fue el poisson.

```
## [1] "AIC:(2), AIC:(3)"
```

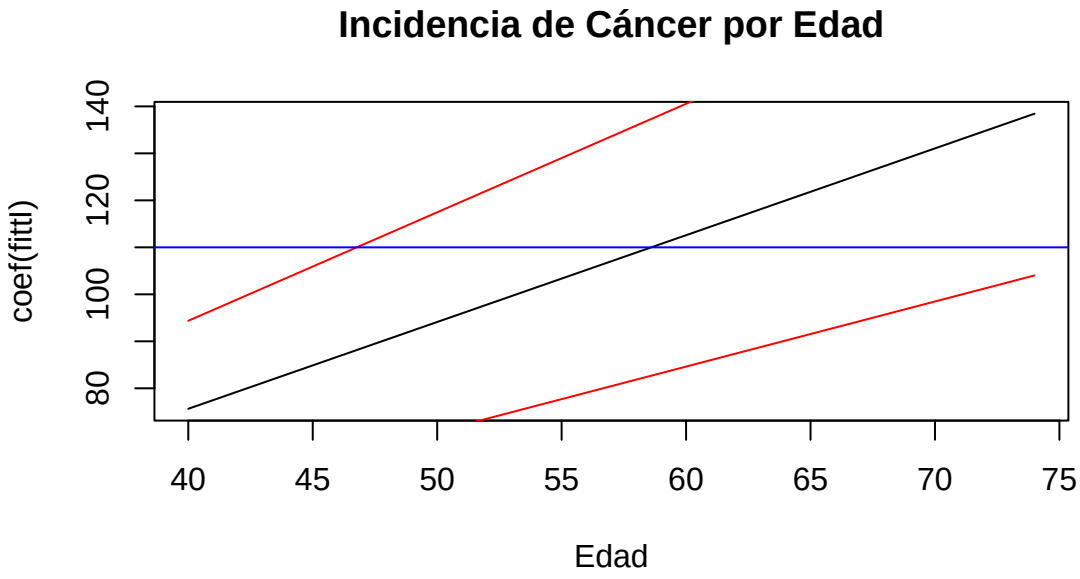
```
## [1] 108.4512 110.4515
```

```
## [1] "BIC:(2), BIC:(3)"
```

```
## [1] 113.4299 116.4259
```

Finalmente haremos intervalos de confianza simultáneos con el modelo Poisson con covariable edad, que para nosotros resultó ser el mejor de los 3.

	Dependent variable:
	Cases
Age55-59	1.082***
	(0.248)
Age60-64	1.502***
	(0.231)
Age65-69	1.750***
	(0.229)
Age70-74	1.847***
	(0.235)
Constant	−5.862***
	(0.174)
Observations	20
Log Likelihood	−50.226
θ	152,366.700 (s.e: 5,232,704.000)
Akaike Inf. Crit.	110.451
Note:	*p<0.1; **p<0.05; ***p<0.01




```
## integer(0)
```

De los resultados obtenidos en la gráfica con los intervalos podemos ver que aproximadamente a partir de los 57-58 años de edad la incidencia de cáncer de pulmón en las ciudades de Dinamarca va en aumento,cosa que ya se podía observar en la gráfica presentada en el primer punto del ejercicio.

Habiendo realizado todo el análisis, podemos decir que la edad juega un papel importante en el modelo y nos ayudó a ver de forma más clara y concisa que en efecto para las ciudades estudiadas por el equipo de investigadores en todas se incrementa la incidencia de cancer pulmonar conforme avanzan los años de edad.

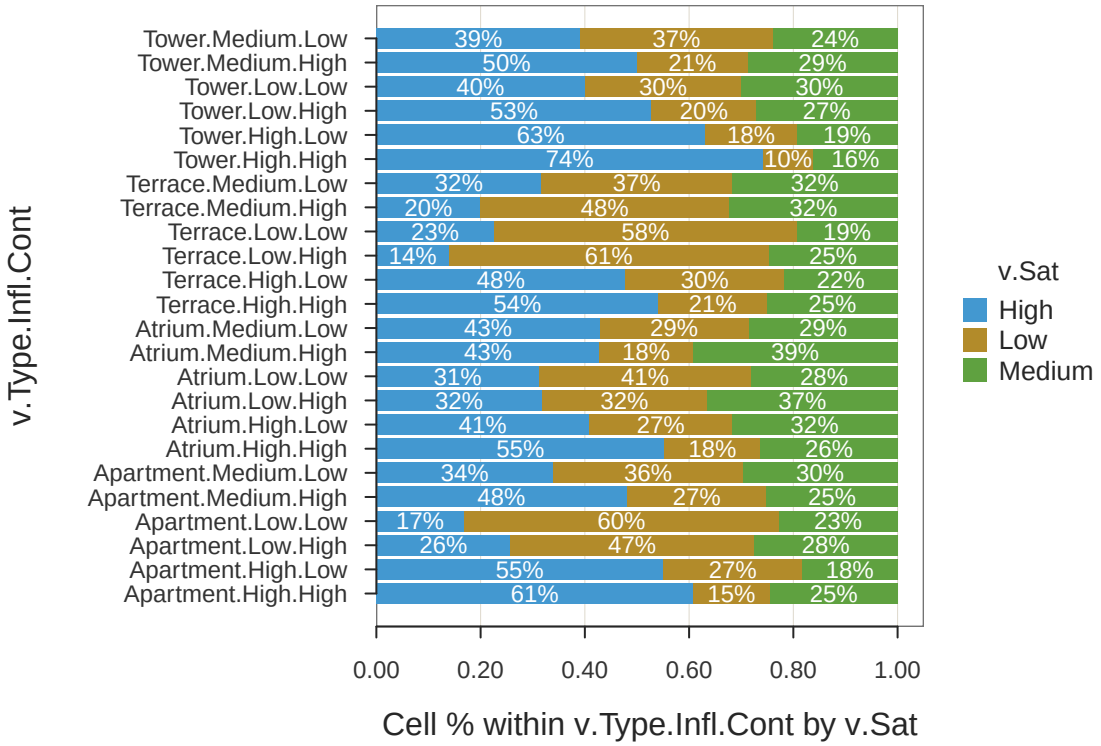
5. Modelos lineales generalizados para datos categóricos

La base de datos Preg5.csv contiene información sobre el nivel de satisfacción (Sat) de un conjunto de individuos que rentan una vivienda. El interés es identificar si entre los factores que definen este nivel están: el tipo de vivienda (Type: apartment, atrium, terrace y tower), la percepción sobre su influencia en las decisiones sobre el mantenimiento de la vivienda (Infl: high, medium y low) y el contacto que tienen con el resto de inquilinos (Cont: high y low).

i) Gráfica de frecuencias relativas

Todas las covariables son categóricas, a continuación mostramos la gráfica que describe las frecuencias relativas para los tres niveles de satisfacción considerando cada cruce Type-Infl-Cont (en ese orden). Podemos observar que la mayor satisfacción (74 %) se presenta en Tower.High.High, que se refiere a vivir en una torre, con alta influencia sobre el mantenimiento de la vivienda y con alto contacto con el resto de los inquilinos. Por otro lado, el menor nivel de satisfacción (14 %) corresponde a vivir en una terraza, con baja influencia sobre el mantenimiento de la vivienda y con alto nivel de contacto con los demás inquilinos.

```
## >>> Note: v.Type.Infl.Cont is not in a data frame (table)
## >>> Note: v.Sat is not in a data frame (table)
```




```
## >>> Suggestions
## Plot(v.Type.Infl.Cont, v.Sat)  # bubble plot
## BarChart(v.Type.Infl.Cont, by=v.Sat, horiz=TRUE)  # horizontal bar chart
## BarChart(v.Type.Infl.Cont, fill="steelblue")  # steelblue bars
##
## Cramer's V: 0.252
##
## Chi-square Test of Independence:
##      Chisq = 213.070, df = 46, p-value = 0.000
```

ii) Modelo logístico multinomial nominal

Ajustamos varios modelos para la variable dependiente de satisfacción (Sat), considerando en un modelo completo las intereacciones de la influencia sobre mantenimiento (Infl), tipo de vivienda (Type) y contacto con otros vecinos (Cont) y no considerando estas interacciones en un modelo reducido. Luego hacemos uso de la función anova que nos permite realizar análisis de varianza entre los modelos ajustados, planteando las hiótesis H_0 : Podemos utilizar el modelo reducido contra H_a : Debemos utilizar el modelo completo. El p-value asociado a la prueba es menor a 0.05, por lo que a un nivel de confianza de 95 %, no tenemos evidencia para rechazar la hipótesis nula, por lo tanto podemos usar el modelo reducido.

```
## Analysis of Deviance Table
##
## Model 1: Sat ~ Infl + Type + Cont
## Model 2: Sat ~ Infl * Type * Cont
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3348     3470.1
## 2      3314     3431.4 34   38.662   0.2671
```

Por otra parte, podemos calcular las AIC de ambos modelos, el resultado es que el AIC del modelo completo o con interacciones es de 3527.4216616 y para el modelo reducido es de 3498.0838663. Por lo tanto, se tiene una mayor evidencia de que el modelo reducido es mejor por tener un menor AIC.

```
##
## Call:
## vglm(formula = Sat ~ Infl + Type + Cont, family = multinomial(refLevel = "Low"),
##     data = Datos)
##
## Coefficients:
##               Estimate Std. Error z value             Pr(>|z|)
## (Intercept):1   1.2201     0.1585   7.699   0.0000000000000137 ***
## (Intercept):2   0.1709     0.1825   0.936             0.349200
## InflLow:1      -1.6126     0.1671  -9.649 < 0.0000000000000002 ***
## InflLow:2      -0.6649     0.1863  -3.568             0.000359 ***
## InflMedium:1   -0.8778     0.1641  -5.348   0.0000000890670893 ***
## InflMedium:2   -0.2185     0.1872  -1.167             0.243151
## TypeAtrium:1    0.3277     0.1886   1.737             0.082391 .
## TypeAtrium:2    0.5671     0.1947   2.913             0.003577 **
## TypeTerrace:1  -0.6767     0.1756  -3.854             0.000116 ***
## TypeTerrace:2  -0.2309     0.1748  -1.321             0.186653
## TypeTower:1     0.7356     0.1553   4.738   0.0000021614222276 ***
## TypeTower:2     0.4357     0.1725   2.525             0.011562 *
## ContLow:1      -0.4818     0.1241  -3.881             0.000104 ***
## ContLow:2      -0.3609     0.1324  -2.726             0.006420 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: log(mu[,1]/mu[,2]), log(mu[,3]/mu[,2])
##
## Residual deviance: 3470.084 on 3348 degrees of freedom
##
## Log-likelihood: -1735.042 on 3348 degrees of freedom
##
## Number of Fisher scoring iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
```

```
##
## Reference group is level  2  of the response
```

iii) Modelo logístico acumulativo (cumulative logit) ordinal

Considerando las covariables del modelo reducido (sin interacciones) y la variable Sat como ordinal, ajustaremos un modelo logístico acumulativo (cumulative logit) sin considerar el supuesto de proporcionalidad (parallel) y otro asumiendo este supuesto. Dado que este último está anidado en el primero, realizaremos una prueba de hipótesis con la función anova para analizar si es plausible asumir este modelo más sencillo, donde planteamos las hipótesis H_0 : Podemos utilizar el modelo reducido contra H_a : Debemos utilizar el modelo completo. El modelo reducido es aquel que tiene probabilidades proporcionales, así que nos quedaremos con ese modelo, pues no hay suficiente evidencia para rechazar la hipótesis nula.

```
## Analysis of Deviance Table
##
## Model 1: Sat ~ Infl + Type + Cont
## Model 2: Sat ~ Infl + Type + Cont
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      3354     3479.1
## 2      3348     3470.6  6   8.5706   0.1992
```

Por otro lado, el AIC del modelo logístico acumulativo sin supuesto de proporcionalidad es de 3498.5787001 y el AIC para el modelo logístico acumulativo con el supuesto de proporcionalidad es de 3495.1492991. Este menor AIC apoya la elección del modelo reducido logístico acumulativo con el supuesto de proporcionalidad.

```
##
## Call:
## vglm(formula = Sat ~ Infl + Type + Cont, family = cumulative(parallel = TRUE),
##     data = Datos)
##
## Coefficients:
##               Estimate Std. Error z value             Pr(>|z|)
## (Intercept):1 -1.57289    0.12519 -12.564 < 0.0000000000000002 ***
## (Intercept):2 -0.38604    0.11938  -3.234             0.001222 **
## InflLow        1.28882    0.12670  10.172 < 0.0000000000000002 ***
## InflMedium     0.72242    0.12372   5.839        0.00000000524 ***
## TypeAtrium    -0.20616    0.13993  -1.473             0.140675
## TypeTerrace    0.51866    0.13358   3.883             0.000103 ***
## TypeTower     -0.57235    0.11875  -4.820        0.00000143628 ***
## ContLow        0.36028    0.09536   3.778             0.000158 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])
##
## Residual deviance: 3479.149 on 3354 degrees of freedom
##
## Log-likelihood: -1739.575 on 3354 degrees of freedom
##
## Number of Fisher scoring iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Exponentiated coefficients:
##     InflLow  InflMedium  TypeAtrium TypeTerrace   TypeTower     ContLow
```

```
##    3.6284964    2.0594206    0.8137002    1.6797833    0.5641981    1.4337373
```

iv) Selección del modelo e interpretación de resultados

Comparando los AIC de todos los modelos, elegimos el modelo logístico multinomial ordinal reducido acumulativo con proporcionalidad, que tiene el menor AIC de 3495.1492991. A continuación se presenta en una gráfica las probabilidades estimadas para cada nivel de satisfacción (Sat: low, medium y high) al considerar la variable de influencia sobre el mantenimiento (Infl) y el nivel de contacto con otros inquilinos (Cont: high y low), cuando se asume que la persona renta una vivienda tipo Apartment.

Podemos observar que la Gráfica de la columna izquierda (Cont=low), cuando se tiene contacto bajo con el resto de inquilinos, muestra las probabilidades de baja, media y alta satisfacción (Sat), considerando únicamente Apartment, y la influencia sobre el mantenimiento (Infl) bajo, medio y alto. En este caso, la probabilidad de baja satisfacción se asocia con mayor probailidad (52 %) a Apartments con baja influencia sobre el mantenimiento y bajo contacto con los otras personas que habitan el lugar, y la mayor probabilidad (51 %) se asocia con Apartments con alta influencia sobre el mantenimiento, a pesar del bajo contacto. Por otra parte, el la Gráfica de la columna derecha, se observa que hay una probabilidad de satisfacción muy alta (60 %) asociada a Apartments donde hay una alta influencia en mantenimiento y alto contacto con los demás inquilinos del lugar, la probabilidad que la satisfacción sea baja en un Apartment de estas características es baja (17 %).