Tarea Examen 1

Versión B

Alejandro Bonilla Alarcón & Angel Josué Mejía Nájera

02-03-2023

Índice

1. Ejercicio 1

1.1. Datos

La base de datos *Preg1B.csv* contiene información sobre 400 pacientes seleccionados de forma aleatoria. Se desea analizar si existe una asociación entre la presión arterial diastólica (bpdiast) y el índice de masa corporal (bmi), en particular, si es posible observar que tener un índice de masa corporal alto se asocia con una alta presión arterial diastólica. Para realizar este análisis se considera el sexo (sex: 1-hombre, 2-mujer) y la edad (age) de los pacientes, pues la presión arterial diastólica podría variar de acuerdo con estos factores.

1.2. Análisis Descriptivo

1.2.1. Estadísticas básicas

Cuadro 1: Principales estadísticas de Presión diastólica por Sexo

sex	Observaciones	Media	Mediana	Varianza
hombre	185	84.47	85	128.9
mujer	215	81.72	80	155.3

A partir del Cuadro 1 observamos lo siguiente:

- El número de observaciones es disparejo ($n = 185$ para hombres y $n = 215$ para mujeres) para cada Sexo.
- En promedio, la Presión Diastólica en Hombres (84.5) es mayor a la de las Mujeres (81.7).
- Notamos mayor variabilidad de la presión para las mujeres y menor variabilidad para los hombres.

1.2.2. Análisis gráfico

Para analizar el efecto del BMI y el Sexo en el valor de la Presión Diastólica, se realizó una gráfica tipo Scatter Plot y un BoxPlot (Figura 1) diferenciando cada observación por medio del Sexo mediante color y forma.

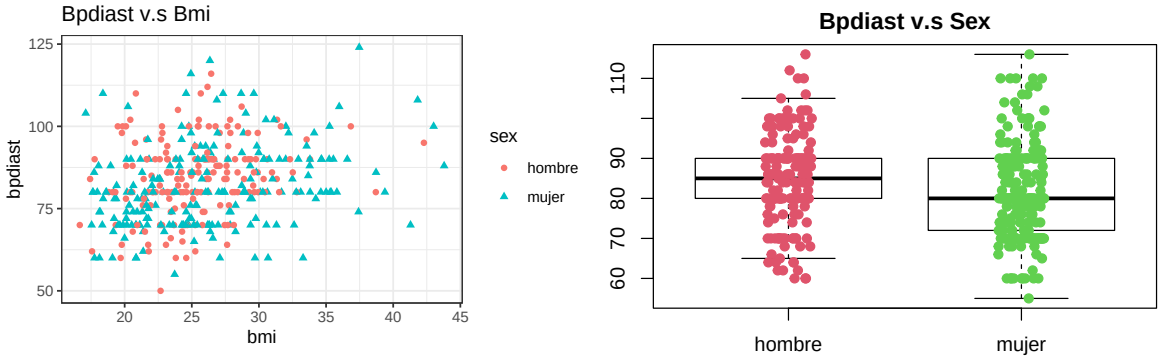


Figura 1: Diagrama de dispersión y boxplot considerando como variable de interés a la Presión Diastólica. El valor de BMI se considera como una variable de ajuste y cada paciente observado se represente a través de un color y forma diferente, dependiendo su sexo.

A partir de la Figura 1 observamos lo siguiente:

- Existe una correlación positiva entre BMI y Presión Diastólica en valores del BMI menores a 33, considerando tanto a hombres como mujeres.
- Para valores del BMI mayores a 33, hay una predominancia de mujeres con elevados niveles de presión diastólica.
- Para valores de BMI menores a 33, los hombres parecen poseer mayores niveles de presión diastólica que mujeres.
- Los valores de la Presión Diastólica se concentran en 80-90 para hombres mientras que para mujeres están más dispersos.

1.3. Ajuste de Modelo

Comenzamos ajustando un modelo de regresión lineal múltiple para $E(Y_i; \text{bmi}, \text{sex}, \text{age})$ considerando la variable bmi como el principal ajuste y sin tomar en cuenta interacciones, notando que el nivel de referencia para la variable categórica `sex` es `hombres`. De tal modo que, la parametrización queda como sigue:

$$\text{E}(Y_i; x_{i1} = \text{bmi}, x_{i2} = \text{sex}, x_{i3} = \text{age}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i\text{M}} + \beta_3 x_{i3}$$

Posteriarmente, se evaluó si este modelo es adecuado para modelar la esperanza. De no ser el caso, sería necesario realizar cambios. En este caso, se observó que el modelo pasó la prueba de significancia de la regresión (se rechazó que los coeficientes B_i en conjunto sean cero); es decir, las covariables utilizadas, en conjunto, aportan información suficiente para modelar $E(Y_i = bpdiast)$.

Sin embargo, como se puede apreciar en la Figura 2, se rechaza que exista linealidad entre la covariable Edad con respecto a la Presión Diastólica. Y en general, el modelo ajustado falla dicho supuesto con una significancia del 5 %. Así mismo, no se cumple el supuesto de normalidad aunque sí el de homocedasticidad.

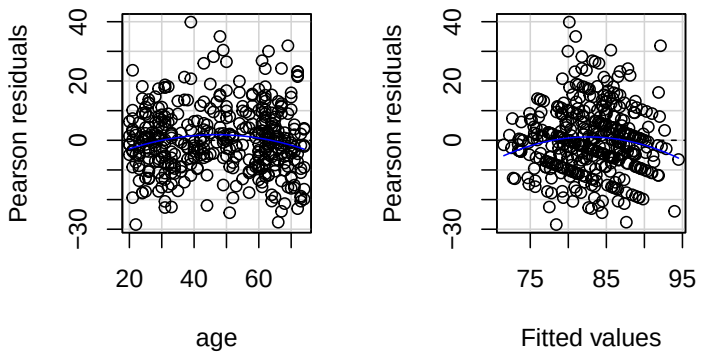


Figura 2: Gráficos de residuales de Pearson vs Edad y Valores ajustados. Notar que la línea azul es lo suficientemente curva para detectar un problema de linealidad. Anexo se presenta la prueba de hipótesis de linealidad, donde H_0 = Cumple Linealidad y se rechaza para ambas pruebas.

```
##            Test stat Pr(>|Test stat|)
## age            -2.57            0.010 *
## Tukey test     -2.42            0.016 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ante la insuficiencia del primer modelo planteado, se emplearon otros 2 modelos potenciales; uno en el que le aplicó la transformación BoxTidwell para tratar la no linealidad, y otro modelo al cual se le aplicó la transformación Box-Cox a la variable `bpdiast` y donde se minimizó el AIC a través de una búsqueda exhaustiva de transformaciones a las covariables contínuas. A continuación, se presentan los 2 modelos resultantes:

Modelo con transformación Box-Tidwell

$$\mathrm{E}(Y_i; x_{i1} = \mathrm{bmi}, x_{i2} = \mathrm{sex}, x_{i3} = \mathrm{age}^{-1}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i\mathrm{M}} + \beta_3 x_{i3}^{-1}$$

Modelo con transformación Box-Cox y transformación de Covariables con AIC mínimo

$$\mathrm{E}(ln(Y_i); x_{i1} = \mathrm{bmi}^{1.7}, x_{i2} = \mathrm{sex}, x_{i3} = \mathrm{age}^{-0.8}) = \beta_0 + \beta_1 x_{i1}^{1.7} + \beta_2 x_{i\mathrm{M}} + \beta_3 x_{i3}^{-0.8}$$

El primer modelo cumplió el supuesto de linealidad más no el de normalidad. En cambio, el 2° modelo cumplió todos los supuestos y obtuvo el menor AIC de entre todos los modelos, como se puede apreciar en el Cuadro 2 . Por lo que optamos por elegir al segundo modelo como el indicado para contestar la problemática planteada.

Es preciso notar que, la transformación $ln()$ sobre la variable dependiente Y, conlleva una pérdida en la interpretabilidad de las B_i, aunado al hecho de que ya no se puede modelar directamente la Esperanza de la Presión Diastólica ($E(Y_i = bpdiast)$) en la escala original, sino su Mediana ($Med(Y_i = bpdiast)$) a través de la transformación $e^{E(ln(Y_i);X)}$. Esto se cumple bajo el supuesto de normalidad con $ln(y_i) = y_i^* \sim N\left(\mathrm{E}(y^*; x), \sigma^2\right)$, donde resulta que $E(y_i^*; X) = Med(y_i^*; X)$.

Note que ambos modelos pasan la prueba de significancia de la regresión.

Cuadro 2: AIC por cada modelo ajustado

	AIC
RLM sin interacciones	3068
RLM + Box-Tidwell	3064
RLM + Box-Cox + Potencias	3055

Entonces, las expresiones del modelo ajustado para el logaritmo de la Presión Diastólica por cada sexo se describen en la expresión 1.

$$\begin{aligned}\widehat{\mathrm{E}}(ln(Y_i); x_{i1}^{1.7}, \mathrm{sex}, x_{i3}^{-0.8}) &= \begin{cases} \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1}^{1.7} + \widehat{\beta}_3 x_{i3}^{-0.8}, & \text{Si el individuo } i \text{ es hombre} \\ (\widehat{\beta}_0 + \widehat{\beta}_2) + \widehat{\beta}_1 x_{i1}^{1.7} + \widehat{\beta}_3 x_{i3}^{-0.8}, & \text{Si el individuo } i \text{ es mujer} \end{cases} \\ &= \begin{cases} 4.459 + 0.0003 x_{i1}^{1.7} - 2.478 x_{i3}^{-0.8}, & \text{Si el individuo } i \text{ es hombre} \\ 4.412 + 0.0003 x_{i1}^{1.7} - 2.478 x_{i3}^{-0.8}, & \text{Si el individuo } i \text{ es mujer} \end{cases} \end{aligned} \tag{1}$$

Se observa que las pendientes de las rectas, en la expresión 1, son iguales entre sí y solo difieren en su ordenada al origen, del cual se puede percibir que, ante una transformación exponencial la esperanza estimada, la mediana de la Presión Diastólica para hombres será mayor que para mujeres.

Con un análisis del cuadro 3, se aprecia una pequeña influencia en el cambio de las Betas estimadas; sin embargo, al eliminar los Outiliers se afecta el tamaño de la muestra, lo cual puede conllevar a problemas en la comparación del AIC con respecto a otros modelos. Además, dado a que se está modelando, finalmente,

la mediana de la Presión Diastólica y no la media, los efectos de estos outliers son menos significativos. Por tanto, se consideró no eliminarlos.

Cuadro 3: Comparación de Betas entre modelo con Outliers vs sin Outliers

Betas	Ajuste.con.outliers	Ajuste.sin.outliers
Intercept	4.4589	4.4648
I(bmi^1.7)	0.0004	0.0004
sexmujer	-0.0468	-0.0502
I(age^-0.8)	-2.4785	-2.4648

1.4. ¿Se puede indicar que para una persona de cierta edad y sexo, tener un índice de masa corporal alto se asocia con una alta presión arterial diastólica?

Para responder la problemática, es necesario realizar la siguiente prueba de hipótesis de una sola dirección:

$$H_0 : \beta_1 <= 0 \quad \text{v.s} \quad H_a : \beta_1 > 0$$

La prueba se rechazó con un $p_{value} < 0.05$, es decir, hay evidencia significativa en contra de que el valor de $B_1 <= 0$ y por lo cual resulta pausible considerar que $B_1 > 0$.

Esta nos permite indicar el coeficiente asociado a la variable bmi es positivo ($B_1 > 0$) dado que los demás factores son constantes; lo que indicaría que, ante valores positivos del BMI, el valor de la Presión Diastólica tendrá un aumento, especialmente para valores altos.

Notemos que la transformación $bmi^{1.7}$ es creciente y la transformación $e^{E(ln(Y_i);X)}$ igual, por lo que la correlación se mantendría en el sentido de, a mayor valor del BMI, la Mediana de la Presión Arterial Diastólica aumentará, con base en los resultados obtenidos en la prueba de hipótesis 1.4.

1.5. Gráfica con curvas ajustadas

Finalmente, podemos notar esta correlación positiva en la Figura 3, igualmente, la Presión Diastólica es mayor para hombres que para mujeres e incluso, a mayor Edad mayor Presión.

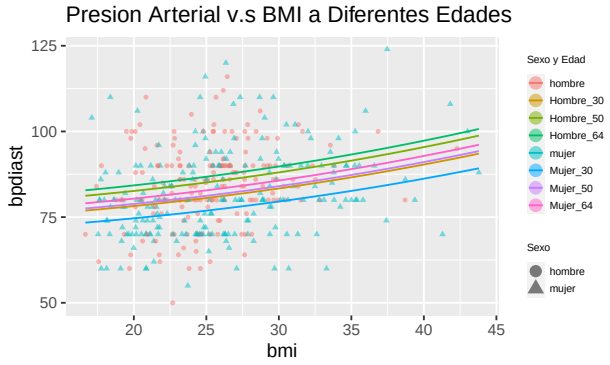


Figura 3: Diagrama de dispersión de la Presión Arterial Diastólica vs BMI, ajustando las curvas enunciadas en la expresión 1 para las edades 30, 50 y 64. Cada observación y curva se representa a través de un color diferente, dependiendo su sexo.

Por lo tanto, podemos concluir que, para una persona de cierta edad y sexo, tener un índice de masa corporal alto se asocia con una alta presión arterial diastólica con una confianza del 95 %.

2. Ejercicio 2

Usando los datos del Ejercicio 1 se buscó entre un conjunto de modelos previamente determinados aquellos que tuvieran el menor AIC y BIC.

2.1. Selección de modelo

Se consideraron varias opciones para el componente lineal, las distribuciones y la función liga:

- Distribuciones: normal, gamma, inversa gaussiana.
- Liga: inversa, identidad, logaritmo, $1/x^2$ (solo Inv.Gauss).
- Componente lineal: polinimios hasta grado 5 y potencias entre -3 y 3 en intervalos de 0.5 (para la potencia 0 se considera la transformación log de la covariable).

Se ajustó un modelo para cada posible combinación del componente lineal, distribución y función liga, dando un total de 6,480 modelos.

Se guardó cada componente de la combinación en una lista, así como el respectivo AIC y BIC para posteriormente acceder a todos los modelos ajustados.

El modelo con menor AIC fue el siguiente:

$$\mathrm{E}(bpdiast) = \beta_0 + \beta_1 x_{Mujer} + \beta_2 x_{BMI}^2 + \beta_3 x_{Age} + \beta_4 x_{Age}^2.$$

Mientras que el modelo con menor BIC fue el siguiente:

$$\mathrm{E}(bpdiast) = \beta_0 + \beta_1 x_{Mujer} + \beta_2 x_{BMI}^2 + \beta_3 x_{Age}^{-0.5}.$$

Ambos modelos toman una distribución Gamma con función liga identidad y tienen la misma forma funcional para el sexo y el BMI.

El AIC y BIC para ambos modelos se presenta en el Cuadro 4.

Cuadro 4: AIC y BIC para cada modelo

Modelo	AIC	BIC
Mínimo AIC	3053	3077
Mínimo BIC	3054	3074

Se observa que el AIC del modelo con menor BIC no difiere mucho al del modelo con menor AIC. Además, el modelo con menor BIC tiene menos parámetros para la covariable *Age*, por lo que este parece un mejor modelo con el que trabajar.

2.2. ¿Se puede indicar que para una persona de cierta edad y sexo, tener un índice de masa corporal alto se asocia con una alta presión arterial diastólica?

El modelo seleccionado se presenta en la expresión 2.2.

$$\mathrm{E}(bpdiast) = \beta_0 + \beta_1 x_{Mujer} + \beta_2 x_{BMI}^2 + \beta_3 x_{Age}^{-0.5}.$$

Note que al modelo 2.2 se le asoció una distribución Gamma y función liga identidad.

De acuerdo al criterio de la primera derivada, el valor esperado de la presión arterial diastólica (bpdiast) es creciente con respecto al BMI, tomando al sexo y edad fijos, si $2\beta_2 x_{BMI} > 0$, esto es, si $\beta_2 > 0$ ya que x_{BMI} siempre es mayor a 0.

Por lo que se realizó la siguente prueba de hipótesis:

$$H_0 : \beta_2 \leq 0 \qquad H_1 : \beta_2 > 0.$$

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: glm(formula = form, family = Dist(link = FunLigas[l]), data = datos)
##
## Linear Hypotheses:
##        Estimate Std. Error z value  Pr(>z)
## 1 <= 0  0.01035    0.00211     4.9 4.7e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Se obtuvo un $p - value < 0.05$, por lo que rechazamos H_0 en favor a H_1. Así que se puede indicar que para una persona de cierta edad y sexo, tener un índice de masa corporal alto se asocia con una alta presión arterial diastólica.

2.3. Gráficas con curvas ajustadas

Finalmente, podemos notar esta correlación positiva en la Figura 4, al igual que Presión Diastólica es mayor para hombres que para mujeres e, incluso, a mayor Edad mayor Presión.

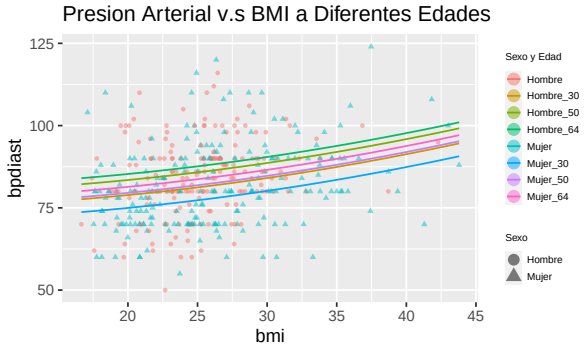


Figura 4: Curvas del modelo ajustado con menor BIC para las edades 30, 50 y 64, y cada sexo. Además se grafican los datos observados.

2.4. Comparación de modelos ajustados

Se ajustaron dos modelos para poder explicar la presión arterial diastólica con las covariables BMI, sex y Age.

En el modelo 1 se consideró un modelo lineal normal y se recurrió a una transformación Box-Cox y transformaciones tipo potencia (x^λ):

$$\mathrm{E}[log(bpdiast)] = \beta_0 + \beta_1 x_{BMI}^{1.7} + \beta_2 x_{Mujer} + \beta_3 x_{Age}^{-0.8}.$$

Para el modelo 2 se recurrió a un modelo lineal generalizado con distribución Gamma, función liga identidad y transformaciones tipo potencia (x^λ):

$$\mathrm{E}(bpdiast) = \beta_0 + \beta_1 x_{Mujer} + \beta_2 x_{BMI}^{2} + \beta_3 x_{Age}^{-0.5}.$$

Por lo que es de interes conocer cuál de los dos enfoques resulta mejor en este caso.

Realizando una comprobación de supuestos a ambos modelos no se observó ningún problema en ningún modelo.

En términos del criterio AIC, el modelo 2 parece ser mejor que el modelo 1 como se observa en el Cuadro 5. Es necesario recordar que hay que realizar ciertos ajustes para obtener el AIC del modelo 1, ya que se está trabajando con una transformación de la variable dependiente Y-bpdiast.

Cuadro 5: AIC para cada modelo

Modelo	AIC
Modelo 1	3055
Modelo 2	3054

En cuanto a la interpretabilidad, es pertinente considerar lo siguiente:

1. Las observaciones son datos positivos por lo que tiene sentido modelarlos con una distrubición que solo tome valores positivos como la Gamma en el modelo 2. El modelo 1 hace uso de la distribución normal que toma valores positivos y negativos.
2. En el modelo 1 modelamos el valor esperado de una transformación de los datos (log), por lo que no es posible hacer conclusiones directas sobre las observaciones. Para el modelo 2 modelamos el valor esperado de los datos y es posible hacer conclusiones más directas sobre el valor esperado de los datos.

Dadas las consideraciones de la distribución, el criterio AIC y la interpretabilidad, es preferible usar el modelo 2, presentado en la expresión 2.4.

3. Ejercicio 3

3.1. Datos

Se desea analizar la información sobre 862 insectos que fueron expuestos a diferentes dosis (Deposit, mg) de tres diferentes 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 también se cuenta con el registro del número de insectos muertos (Killed) y el número total de insectos expuestos (Number) por cada dosis e insecticida.

Las variables de estudio son:

- Deposit = Dosis de exposición al insecticida en mg
- Insecticide = Insecticida (A, B, C) - El nivel de referencia es **A**
- Y = Si el insecto está muerto o no (1-Sí, 0-No)

Entonces, el objetivo del estudio es analizar la probabilidad de que un insecto muera dada una cierta dosis por cada tipo de insecticida, y posteriormente contestar a las siguientes problemáticas:

- ¿Se puede indicar que los insecticidas A y B tienen un desempeño similar?
- ¿Cuál es la dosis mínima para la cual se puede indicar que el 75
- ¿Se puede indiciar que algún insecticida es mejor?

3.2. Análisis Descriptivo

Los resultados del estudio se pueden visualizar en la Figura 5; además, se aprecia lo siguiente:

- Los insectos afectados por el Insecticida C poseen una mayor probabilidad de muerte que los otros 2 insecticidas, alcanzando, para valores de exposición mayores a 3, una tasa de mortalidad de casi 1.
- Los insectos afectados por los Insecticidas A y B poseen un rendimiento similar.
- La probabilidad de muerte aumenta conforme se da un incremento en la exposición a los insecticidas.

```
## NULL
```

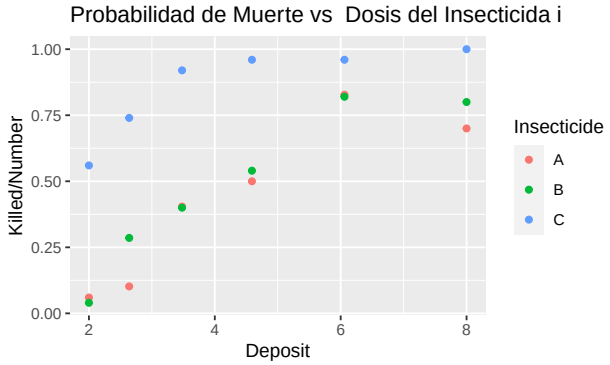


Figura 5: Diagrama de dispersión considerando como variable de interés a la probabilidad de que un insecto esté muerto. La dosis de exposición se considera como una variable de ajuste. Cada proporción se representa a través de un color diferente, dependiendo el insecticida utilizado (A, B o C)

3.3. Ajuste de Modelo

Para poder modelar la probabilidad de muerte de los insectos, se consideró inicialmente un modelo de regresión múltiple (ANCOVA) que incluye la información de las variables Deposit (Dosis) e Insectice (Tipo de Insecticida) y su interacción para poder analizar si la exposición a las dosis actúa de la misma forma para cada insecticida. Además, se considera una transformación $ln()$ sobre la covariable Deposit.

El ajuste se realizó sobre datos desagregados con respecto a 3 diferentes ligas (**probit**,**logit**,**cloglog**).

Posteriormente, se realizó la misma comparación pero ahora incluyendo la covariable ln(Deposit)ˆ2 y su interacción con la covariable Insecticide.

Es necesario notar que se decidió utilizar como criterio de comparación el AIC pues este no penaliza tanto la inclusión de nuevas covariables, lo que nos permite centrarnos en la suficiencia que aporta una covariable a la hora de modelar la probabilidad.

Como resultados del análisis, se encontró que los 6 modelos ajustados cumplieron la verificación de supuestos. Para esto, se utilizó la paquetería DHARMa.

La inclusión del término cuadrático mejoró la bondad de ajuste de los residuales y la linealidad, a cambio de una pequeña pérdida en la correcta modelación del parámetro de dispersión. Y, de entre los 6 modelos ajustados, los que obtuvieron menor AIC fueron los modelos con adición del término cuadrático, como se puede apreciar en la tabla 6 (esto posiblemente debido a la menor penalización del número de covariables que aporta el AIC).

Cuadro 6: Comparación de AICs entre modelos con distintas ligas, y considerando si incluye o no el término cuadrático de la variable ln(Deposit)

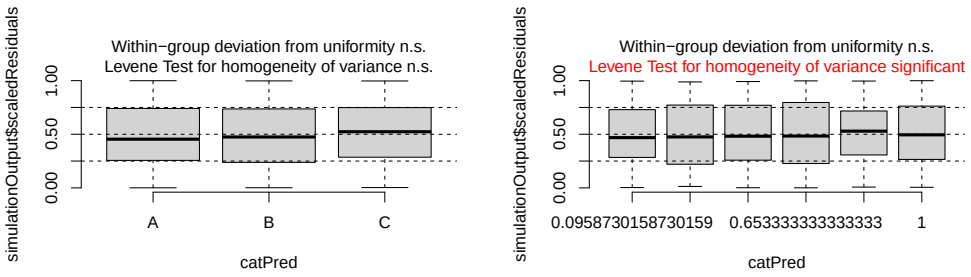
Ligas	AIC.s..Término.Cuadrático	AIC.c..Término.Cuadrático
Probit	789.4	786.6
Logit	789.3	786.9
Cloglog	800.5	786.1

Dado que la diferencia entre AIC de los modelos que incluyen término cuadrática es mínima y considerando la mejor interpretabilidad que aporta el modelo con liga logit, se decidió utilizar este último para el modelado de la probabilidad.

El modelo seleccionado con liga logit y término cuadrático se visualiza en la expresión 2. Considerando $\pi_1 = P(Y = InsectoMuere; X)$

$$\text{Eta}(Y_i; x_{i1} = \text{Insecticide}, x_{i2} = \ln(\text{Deposit}), x_{i3} = \ln(\text{Deposit})^2) = \ln\left(\frac{\pi_1}{1-\pi_1}\right) = \\ \beta_0 + \beta_1 x_{i\text{B}} + \beta_2 x_{i\text{C}} + \beta_3 x_{i2} + \beta_4 x_{i3} + \beta_5 x_{i\text{B}} x_{i2} + \beta_6 x_{i\text{C}} x_{i2} + \beta_7 x_{i\text{B}} x_{i3} + \beta_8 x_{i\text{C}} x_{i3} \tag{2}$$

A continuación se visualiza el cumplimiento de los supuestos con la paquetería DHARMa:



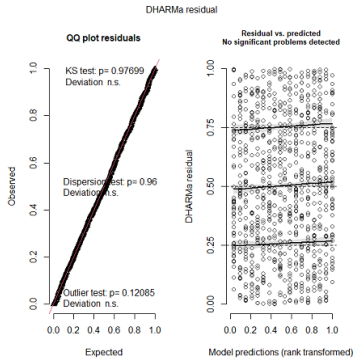


Figura 6: (1° fila) Gráficos de residuales vs $covariable_i$, i en (Insecticide, Deposit). Se espera que cada el rango intercuantil quede encerrado entre las bandas punteadas.(2° fila) Gráfico de bondad ajuste de los residuales y gráfico de Residuales $DHARMa$ vs valores ajustados

En la Figura 6 no se aprecian problemas con los supuestos, la bondad de ajuste de los residuales no se rechaza. Además, no se rechaza que se esté modelando correctamente la varianza dado el parametro de dispersion. En cuanto al componente lineal y las covariables, no parece evidente que haya un problema con la linealidad. Por lo que se concluye que no se encuentra evidencia en contra de los supuestos.

Para la prueba de significancia de la regresión, se tiene evidencia a favor de que al menos una covariable, en conjunto a las demás, aporta información suficiente para modelar la probabilidad.

3.4. ¿Se puede indicar que los insecticidas A y B tienen un desempeño similar?

Primeramente, se encontró evidencia de que, con una confianza del 95 %, el $ln(Dosis)^2$ actúa igual para los 3 tipos de insecticidas, es decir, dado x_3 las 3 curvas ajustadas poseen la misma pendiente. Del mismo modo se concluyó para $ln(Dosis)$ una vez reducido el modelo. Por lo que finalmente, es plausible considerar un modelo sin interacciones.

Luego, se observó que el insecticida B no aporta mayor información dada la inclusión de las demás covariables al modelo (se empleó una prueba tipo t). Posteriormente, se verificó a través de una prueba $anova$, que el modelo reducido es suficiente para modelar la probabilidad (No se rechazó que $B_j = 0, j \notin \{B_2, B_3, B_4\}$). Del mismo modo, se verificó que el modelo reducido cumple con los supuestos.

Por lo tanto, se concluye que es suficiente considerar a la Probabilidad estimada de que un Insecto muera, dado que se utilice el Insecticida A o B, iguales. En otras palabras,los insecticidas A y B tienen un desempeño similar con un 95 % de confianza.

La expresión del modelo reducido se aprecia en 3

$$\widehat{\text{Eta}}(Y_i; \text{Insect}, \ln(Deposit), \ln(Deposit)^2) = \begin{cases} \widehat{\beta}_0 + \widehat{\beta}_2 x_{i2} + \widehat{\beta}_3 x_{i3}, & \text{Si el Insecticida } i \text{ es A o B} \\ (\widehat{\beta}_0 + \widehat{\beta}_1) + \widehat{\beta}_2 x_{i2} + \widehat{\beta}_3 x_{i3}, & \text{Si el Insecticida } i \text{ es C} \end{cases} \tag{3}$$

Con $B_0 = -6.820, B_1 = 2.804, B_2 = 6.891, B_3 = -1.434$

Dado que el coeficiente B1 es positivo (única diferencia entre las curvas), es de esperar que haya un crecimiento en la probabilidad de que un insecto muera al aplicar el insecticida C a que si se aplica alguno de los otros 2.

Notar que el AIC del modelo reducido es: **780.7**, resultando menor que los modelos presentados en 6 y menor al AIC del modelo reducido **sin** considerar término cuadrático, lo demuestra lo favorable que fue la inclusión del término cuadrático, pues a pesar de adicionar una covariable más, esto mejoró el AIC en vez de penalizarlo.

3.5. Gráfico de Estimación Puntual (Modelo Reducido)

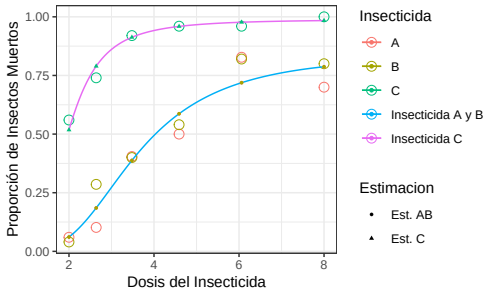


Figura 7: Diagrama de dispersión, de las proporciones observadas de insectos muertos por cada insecticida (A, B o C), y curvas ajustadas del modelo descrito en 3 diferenciadas por color, además de estimaciones puntuales dados los valores de dosis observados.

3.6. Dosis mínima p/Insecticida con la que el 75 % de los insectos se muere.

Primeramente, es necesario notar que se utilizó el modelo reducido descrito en 3 para obtener el valor solicitado. Además, se emplearon métodos numéricos para hallar la solución la cual se describe en la tabla 7. Para encontrar dicha dosis, se buscó el valor $x \;\; t.q \;\; \pi_1 = 0.75 = \left(\frac{e^{\beta_0+\beta_1 1_C+\beta_2 ln(x)+\beta_3 ln(x)^2}}{1+e^{\beta_0+\beta_1 1_C+\beta_2 ln(x)+\beta_3 ln(x)^2}}\right)$.

Cuadro 7: Dosis por Insecticida t.q el 75 % de los insectos muere

Insecticida	Dosis
A	6.701
B	6.701
C	2.503

3.6.1. ¿Se puede indicar que un insecticida es el mejor? (Considerando la menor dosis)

Considerando que la menor de las dosis encontradas es: 2.503, podemos realizar la siguiente prueba de hipótesis:

$$H_0 : \text{Eta}(Y_i; AvB, \ln(2.503), \ln(2.503)^2) <= \text{Eta}(Y_i; C, \ln(2.503), \ln(2.503)^2) \quad \text{v.s}$$
$$H_a : \text{Eta}(Y_i; AvB, \ln(2.503), \ln(2.503)^2) > \text{Eta}(Y_i; C, \ln(2.503), \ln(2.503)^2)$$
$$\iff H_0 : 0 >= \beta_1 \quad \text{v.s} \quad H_a : 0 < \beta_1$$

Se rechaza H_0 con un $p_{value} < 0.05$. Entonces, bajo un nivel de significancia del 5 %, la probabilidad de matar un insecto con el insecticida C es mayor que al usar el insecticida A o B (ver Figura 8). Observe que la inversa de la función liga es creciente, por lo cual se mantiene la relación de orden.

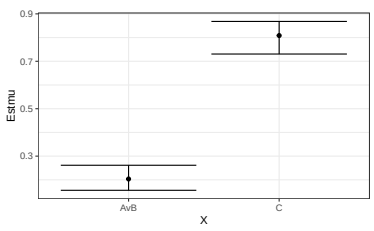


Figura 8: Estimación Puntual al 0.95 de confianza.

4. Ejercicio 4

4.1. Datos

La base de datos `Preg4.csv` contiene información sobre el número de casos de cáncer de pulmón registrados entre 1968 y 1971 en cuatro cuidades 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.

4.2. Análisis gráfico

Se realizó un diagrama de dispersión para analizar el efecto de la edad y la ciudad de residencia en la incidencia de casos de cáncer de pulmón, considerando que la población total en cada categoría afecta al número de casos registrado (en una población más grande se esperaría observar un mayor número de casos).

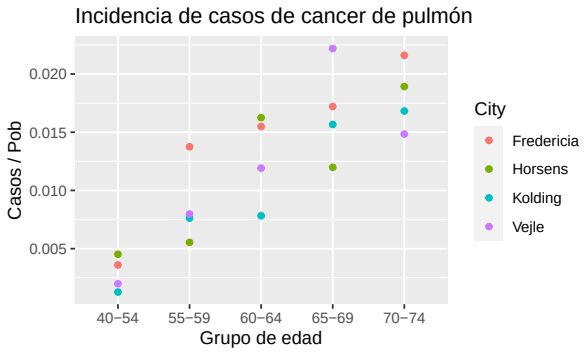


Figura 9: Diagrama de dispersión considerando como variable de interés a la incidencia de casos (Casos / Población) diferenciando por grupo de edad y ciudad.

A partir de la Figura 9 se observó lo siguiente:

- Sin difenciar por ciudad se observa una relación creciente entre el grupo de edad y la incidencia de casos.
- Esta relación se mantiene en cada ciudad.

De lo anterior, se tienen dos consideraciones al momento de seleccinar un modelo. Primero, dado que el número de casos es una variable de conteo es natural pensar en un modelo con distribución Poisson. Segundo, parece posible trabajar con un modelo con única covariable *Age*.

4.3. Selección de modelo

Se consideraron dos posibles modelos para modelar la incidencia de casos, esto es,

$$\mathrm{E}(Y)/Pop$$

Donde Y es el número de casos y Pop es la población total de cada categoría.
Por un lado, se consideró un modelo con covariables *Age* y *City*, así como sus interacciones.

$$Incidencia = \mathrm{e}^{\beta_0+\beta_{Age}+\beta_{City}+\beta_{Age:City}}$$

Por otro lado, un modelo sólamente con la covariable *Age*.

$$Incidencia = \mathrm{e}^{\beta_0+\beta_{Age}}$$

Donde:

β_{Age} es el coeficiente de acuerdo al grupo de edad.

β_{City} es el coeficiente de acuerdo a la ciudad.

$\beta_{Age:City}$ es el coeficiente de la interacción del grupo de edad y la ciudad.

Es de interés trabajar con un modelo que tenga el menor número de parámetros posibles. Por lo que se realizó la siguiente prueba ANOVA para comparar entre los dos modelos anteriores:

$$H_0 : \beta_{City} = 0, \beta_{Age:City} = 0 \qquad H_1 : \beta_{City} \neq 0, \beta_{Age:City} \neq 0.$$

Se obtuvo como resultado un $p - value = 0.3202$, mayor a 0.05, por lo que no se encuentra evidencia en contra de H_0 a un nivel de significancia del 5 %.

```
## Analysis of Deviance Table
##
## Model 1: Cases ~ (Age + City)^2
## Model 2: Cases ~ Age
##   Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
## 1         0          0
## 2        15         17 -15      -17     0.32
```

De igual forma se compararon ambos modelos usando los criterios AIC y BIC como se observa en el Cuadro 8.

Cuadro 8: AIC y BIC para cada modelo

Modelo	AIC	BIC
Poisson(Age, City, Age:City)	121.5	141.4
Poisson(Age)	108.5	113.4

Por lo anterior y dado lo observado en la Figura 9, parece plausible trabajar con el modelo Poisson con una única covariable *Age*.

4.4. Generalización del modelo

Dado que un modelo Binomial Negativo puede verse como una generalización del modelo Poisson para tomar en cuenta la sobre/sub dispersión en los datos, se ajustó un modelo Binomial Negativo con única covariable *Age*.

Primero se realizó una comprobación de supuestos para el modelo Poisson y el modelo Binomial Negativo en donde no se observó algún problema en los supuestos.

Luego se compararon ambos modelos usando los criterios AIC y BIC como se observa en el Cuadro 9. Se puede ver que el modelo Poisson es mejor que el modelo Binimial Negativo tanto en el criterio AIC como en el criterio BIC.

Cuadro 9: AIC y BIC para cada modelo

Modelo	AIC	BIC
Poisson(Age)	108.5	113.4
BN(Age)	110.5	116.4

Ahora, con el modelo Poisson se calcularon intervalos de confianza simúltaneos para cada grupo de edad como se observa en la Figura 10.

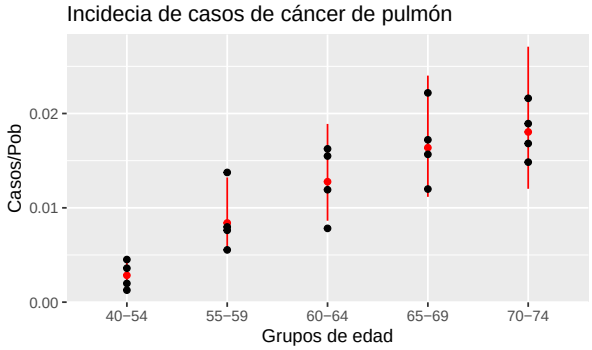


Figura 10: Datos observados (puntos negros) con los respectivos intervalos de confianza por grupo de edad (líneas rojas) y las estimaciones del modelo Poisson ajustado (puntos rojos).

Observando las estimaciones del modelo parece ser que a mayor grupo de edad, existe una mayor incidencia de casos. Sin embargo, dado que los intervalos de confianza se translapan y no estan unos por arriba de otros completamente, no es posible afirmar lo anterior con seguridad.

4.5. Modificación de las covariables

Dado que la edad es naturalmente una variable numérica se decidió sustituir a los grupos de edad por su respectivo valor medio para que el modelo considerara a la variable edad como númerica.
Se ajustaron cuatro modelos considerando las distribuciones Poisson y Binomial Negativa y los componentes lineales Age y Age, Age^2.

En cuanto a la comprobación de supuestos, se presentaron problemas con el componente lineal para los modelos con unica covariable Age, en tanto que los modelos con covariables Age y Age^2 no presentaron ningún problema.

Posteriormente se compararon los dos modelos que no presentaron problemas con los supuestos. De acuerdo al Cuadro 10 vemos que el modelo Poisson con covariables Age y Age^2 tiene mejores valores en ambos criterios.

Cuadro 10: AIC y BIC para cada modelo sin problemas en los supuestos

Modelo	AIC	BIC
Poisson(Age, Age^2)	104.5	107.5
BN(Age, Age^2)	106.5	110.5

Trabajando con el modelo Poisson

$$Incidencia = \mathrm{e}^{\beta_0+\beta_1 x_{Age}+\beta_2 x_{Age}^2},$$

Es de interés conocer si la incidencia de casos es creciente con la edad en el rango de 40 y 74 años.
De acuerdo al criterio de la primera derivada, la incidencia de casos es creciente con la edad si se cumple que $\beta_1 + 2x_{Age}\beta_2 > 0$. Por lo que se realizó la siguiente prueba de hipótesis:

$$H_0 : \beta_1 + 2x_{Age}\beta_2 \leq 0 \qquad H_1 : \beta_1 + 2x_{Age}\beta_2 > 0.$$

,
Donde se obtuvo un $p - value$ menor a 0.05, así que hay evidencia en contra de H_0 en favor de H_1.
Por lo tanto, es posible afirmar que el promedio de incidencia de casos de cáncer de pulmón es creciente con respecto a la edad en el rango de 40 a 74 años.

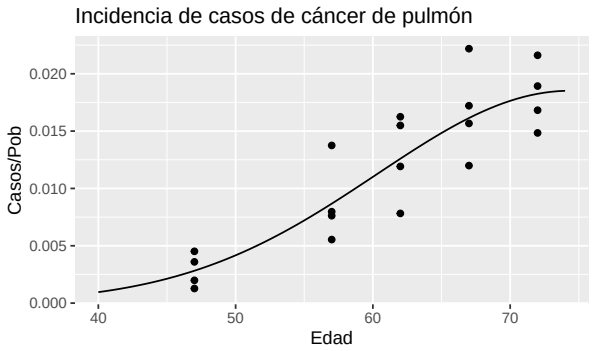


Figura 11: Curva estimada con el modelo Poisson seleccionado junto con los datos observados

En la Figura 11 se puede ver cómo la curva estimada de la incidencia de casos es creciente conforme la edad aumenta.

5. Ejercicio 5

5.1. Datos

Se tiene información de una encuesta sobre el nivel de satisfacción (Sat) de un conjunto de 1681 individuos que rentan una vivienda. El interés es identificar si entre los factores que definen este nivel están: el tipo de vivienda (Type), la percepción sobre su influencia en las decisiones sobre el mantenimiento de la vivienda (Infl) y el contacto que tienen con el resto de inquilinos (Cont), haciendo, en conjunto, un total de 72 categorías posibles.
Entonces, el objetivo es determinar si la frecuencia relativa de cada nivel de satisfacción está influenciado por las 3 categorías principales mostradas. Es importante notar que la variable dependiente de interés $Y = Sat$, es una variable ordinal de 3 categorías.
Las variables de estudio son:

- Type = Tipo de Vivienda (Apartment, Atrium, Terrace, Tower) - El nivel de referencia es **Apartment**
- Infl = Influencia (High, Low) - El nivel de referencia es **High**
- Cont = Contacto (High, Low) - El nivel de referencia es **High**
- Y = Satisfacción (1-High, 0-Low)

5.2. Análisis Descriptivo

Los resultados del estudio se pueden visualizar en la Figura 12; además, se observa lo siguiente:

- La satisfacción de los inquilinos varía ligeramente según el tipo de vivienda Los inquilinos de las torres son los generalmente más satisfechos, seguidos por los de los atrios, luego apartamentos y, finalmente, terrazas.
- Dentro de cada tipo de vivienda, los grupos con la mayor satisfacción generalmente son aquellos que tienen una alta percepción de influencia en las decisiones de mantenimiento (High) y que tienen un contacto frecuente con otros inquilinos (High).
- En general, los grupos con menor satisfacción son aquellos con baja influencia (Low) y aquellos con poco contacto con otros inquilinos (Low).
- En algunos casos, la influencia y el contacto parecen tener un efecto opuesto en la satisfacción. Por ejemplo, para los inquilinos en Terrazas, el grupo con la menor satisfacción es aquel con poca influencia pero alto contacto (frecuencia relativa de 0.61 para Sat-Low). Lo mismo ocurre para los inquilinos de los apartamentos y los atrios.
- Los apartamentos Tower tienen la mayor frecuencia de tener una alta satisfacción en comparación con otros tipos de apartamentos.
- A mayor nfluencia que posean los inquilinos, mayor frecuencia tendrán de estar satisfechos con su vivienda.

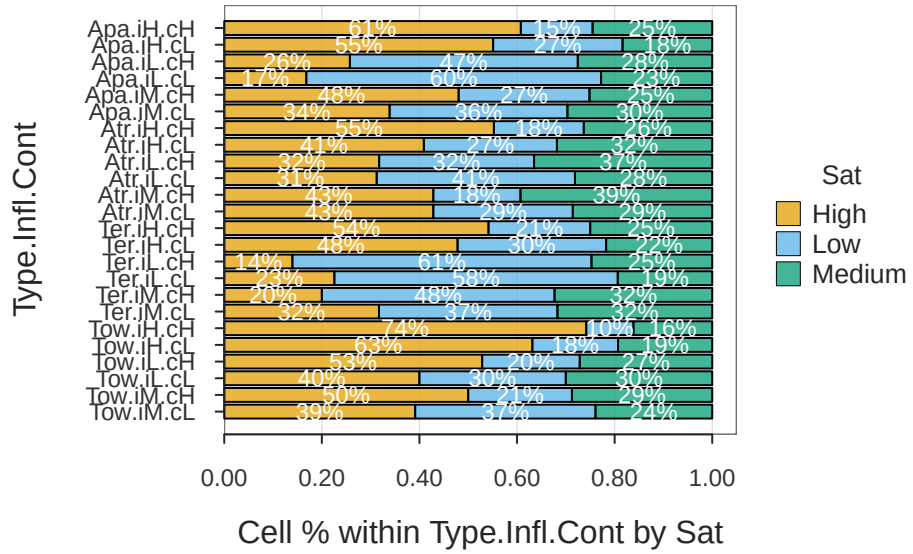


Figura 12: Frecuencias relativas de satisfacción para cada posible individuo definido por el tipo de apartamento, influencia en el mantenimiento y contacto con vecinos.
(Apa=Apartment, Tow=Tower, Ter=Terrace, Atr=Atrium, i=Influence, c=cont, H=High, M=Medium, L=Low)

5.3. Ajuste de Modelo

Primero se decidió ajustar un modelo de regresión múltiple multinomial con liga *logit* a través de la paquetería *vglm*, en donde se consideró a todas las posibles interacciones entre las variables $Type, Infl$ y $Cont$, dando un total de 48 covariables posibles en el modelo lo cual puede provocar un aumento drástico en el error tipo II (No Rechazar H_0 cuando H_0 es falsa). Por lo que resulta pertinente encontrar un modelo reducido.

Para proceder a reducir el modelo, se realizó una prueba de significancia de la regresión, comparando si todas las covariables no eran significativas contra si al menos una covariable, dada las demás, sí era significativa. La prueba se rechazó a un nivel de confianza del 95 %, lo que significa que el conjunto de covariables seleccionadas es apto para modelar la satisfacción. Luego, se procedió a considerar un modelo reducido sin interacciones.

Se realizó una prueba anova para discernir si el modelo reducido era suficiente para modelar las probabilidades. Observe que, el conjunto de covariables del modelo reducido, es un subconjunto del completo (es un modelo anidado). La salida se puede apreciar en la sección 5.3.1.

5.3.1. Prueba-Anova (Modelo c/ Interacciones vs Modelo s/ Interacciones)

```
## 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      3314       3431
## 2      3348       3470 -34    -38.7     0.27
```

No se rechaza la hipotesis nula, es decir, no hay evidencia de que las betas que solo se encuentran presentes en el modelo completo sean diferenres de cero; es decir, no hay correlacion entre el tipo de departamento, influencia y contacto. Por lo tanto, podemos quedarnos con el modelo reducido.

En el cuadro 11, es notable la mejora en los índices del modelo completo con respecto al modelo reducido, especialmente en el BIC que da mayor penalizacion a un mayor numero de covariables. Con esto, es concluyente que es óptimo usar el modelo reducido.

Cuadro 11: Comparación de AIC y BIC entre modelos logísticos

Modelo	AIC	BIC
Modelo Completo	3527	3788
Modelo Reducido	3498	3574

Además, en la tabla 12 se pueden observar los coeficientes ajustados del modelo reducido.

Cuadro 12: Coeficientes ajustados de MLG logístico Reducido

log(pi_2/pi_1)	log(pi_3/pi_1)	Betas
-1.2201	-1.0492	(Intercept)
1.6126	0.9477	InflLow
0.8778	0.6592	InflMedium
-0.3277	0.2394	TypeAtrium
0.6767	0.4458	TypeTerrace
-0.7356	-0.2999	TypeTower
0.4818	0.1210	ContLow

5.3.2. Ajuste modelo cuando se considera Ordinal a la variable categórica

Ahora bien, dado que la variable categórica es ordinal (está ordenada), se puede utilizar el hecho presentado en la expresión 5.3.2. Y por lo visto en la sección 5.3.1, se usó el modelo reducido para emplear este supuesto de probabilidades acumulativas.

$$\text{logit}(P(Y \leq j)) = \log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = \log\left(\frac{\pi_1 + \ldots + \pi_j}{\pi_{j+1} + \ldots + \pi_c}\right) \tag{4}$$

De tal modo, que esperaríamos obtener el mismo número de parámetros pero quizás un mejor AIC o BIC. En el cuadro 13 se presentan los coeficientes del modelo ajustado.

Cuadro 13: Coeficientes ajustados de MLG logístico Acumulativo

P[Y<=1]	P[Y<=2]	Betas
0.4339	1.3003	(Intercept)
0.0031	-0.3426	TypeAtrium
-0.5765	-0.0289	TypeTerrace
0.4867	0.0389	TypeTower
-1.2662	-0.3093	InflLow
-0.7630	-0.3901	InflMedium
-0.3144	0.1054	ContLow

5.3.3. Ajuste modelo cuando se considera supuesto de curvas paralelas

Aparte de considerar a la variable categórica de interés como una variable ordinal, se puede agregar el supuesto de curvas paralelas que reduciría el número de covariables a la mitad, dicho supuesto se puede visualizar en la expresión 5.3.3.

$$\log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = \beta_0^{(j)} + \beta_1 x_1 + \ldots + \beta_p x_p \tag{5}$$

Y dado que se puede considerar a este modelo de probabilidades proporcionales como un modelo anidado del logístico acumulativo, procedemos a realizar la prueba anova de significancia, de donde se obtuvo el resultado de que, con un 95 % de confianza, se rechazaba H_0, es decir, hay evidencia suficiente para decir que el modelo mas completo aporta informacion necesaria para el modelado de la probabilidad.

Ante esta situacion, si se elige el modelo con curvas paralelas, se podrían perder efectos importantes aportados por los coeficientes adjuntos al componente lineal de [P(Y <= 2)], en el modelado. Por lo que se decidió, finalmente, comparar mediante el AIC (dado que ambos modelos poseen el mismo número de covariables) el modelo reducido 12 contra el modelo acumulativo.

La comparación se puede apreciar en el cuadro 14.

Cuadro 14: Comparación de AIC entre modelo reducido y acumulativo

Modelo	AIC
Modelo Reducido	3498
Modelo Acumulativo	3504

Dado que el modelo multinomial reducido posee menor AIC, obtamos por usar este para modelar las probabilidades de satisfacción dado las covariables propuestas.

5.4. Gráfico de Estimación Puntual (Modelo Reducido)

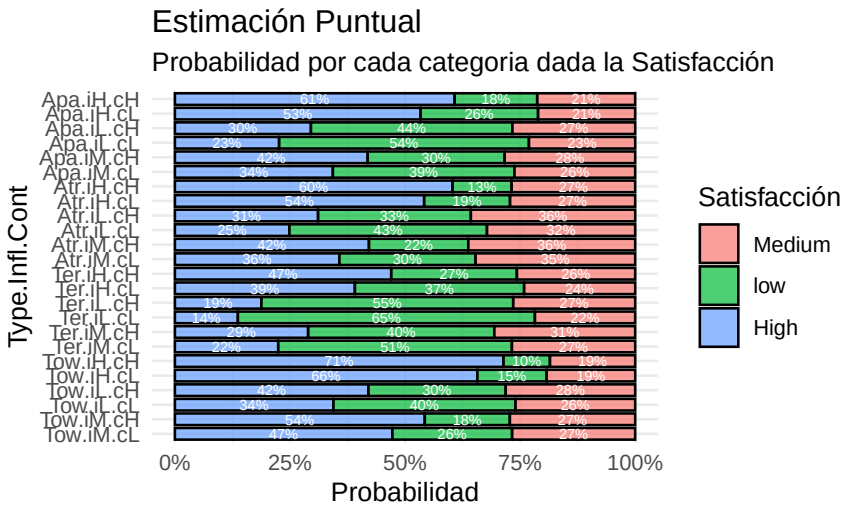


Figura 13: Probabilidades estimadas, relativas a la satisfacción para cada posible individuo definido por el tipo de apartamento, influencia en el mantenimiento y contacto con vecinos (72 posibles categorias). (Apa=Apartment, Tow=Tower, Ter=Terrace, Atr=Atrium, i=Influence, c=cont, H=High, M=Medium, L=Low)

Las estimaciones puntuales se presentan en la Figura 13, de la cual, a grandes rasgos se pueden obtener las mismas conclusiones que en el apartado 5.2, pero ahora en el sentido de probabilidades estimadas. Finalmente, si consideramos puntualmente, el efecto sobre la satisfacción cuando observamos variable Infl, asumientdo que el inquilino renta una vivienda tipo Apartment y tiene un nivel de contacto con otros inquilinos como High. Se puede apreciar que, a medida que aumenta la influencia percibida, se puede estimar un aumento en la probabilidad de satisfacción alta, una disminución en la probabilidad de satisfacción baja y una estabilización en torno al 30 % de la satisfacción media.