



Universidad Nacional Autónoma de México
Facultad de Ciencias

# Examen 1 A

Realizado por

Cícero Álvarez Alicia Guadalupe 318010807
Hernández Alva Luis Ángel 315251674
Isunza Alvarez Marcos Guillermo 419002921
Regalado Urbina Brandon Imanol 317312878

Profesores

Gonzalo Pérez de la Cruz
Noe Eusebio Amador González
César Humberto Valle Márquez

Asignatura

**Seminario de Estadística I:**

**Aprendizaje Estadístico Automatizado**

29 de octubre de 2023

# 1.- Inferencia en la presión diastólica a partir del BMI, el sexo y la edad usando regresión lineal múltiple

Se nos proporcionó la información de 400 pacientes, en ella se incluye el índice de masa corporal (bmi), el sexo y la edad. Es de nuestro interés coroborar si un bmi alto se asocia con una presión diastólica alta para cierta edad y sexo.

Comenzamos el análisis presentando los datos, teniendo en cuenta el sexo se muestra la relación entre presión vs bmi y presión vs edad. Como dato extra los indiviudos muestreados para este anális tienen en promedio 48 años, un bmi de 26 y una presión de 83 (Tabla 1).

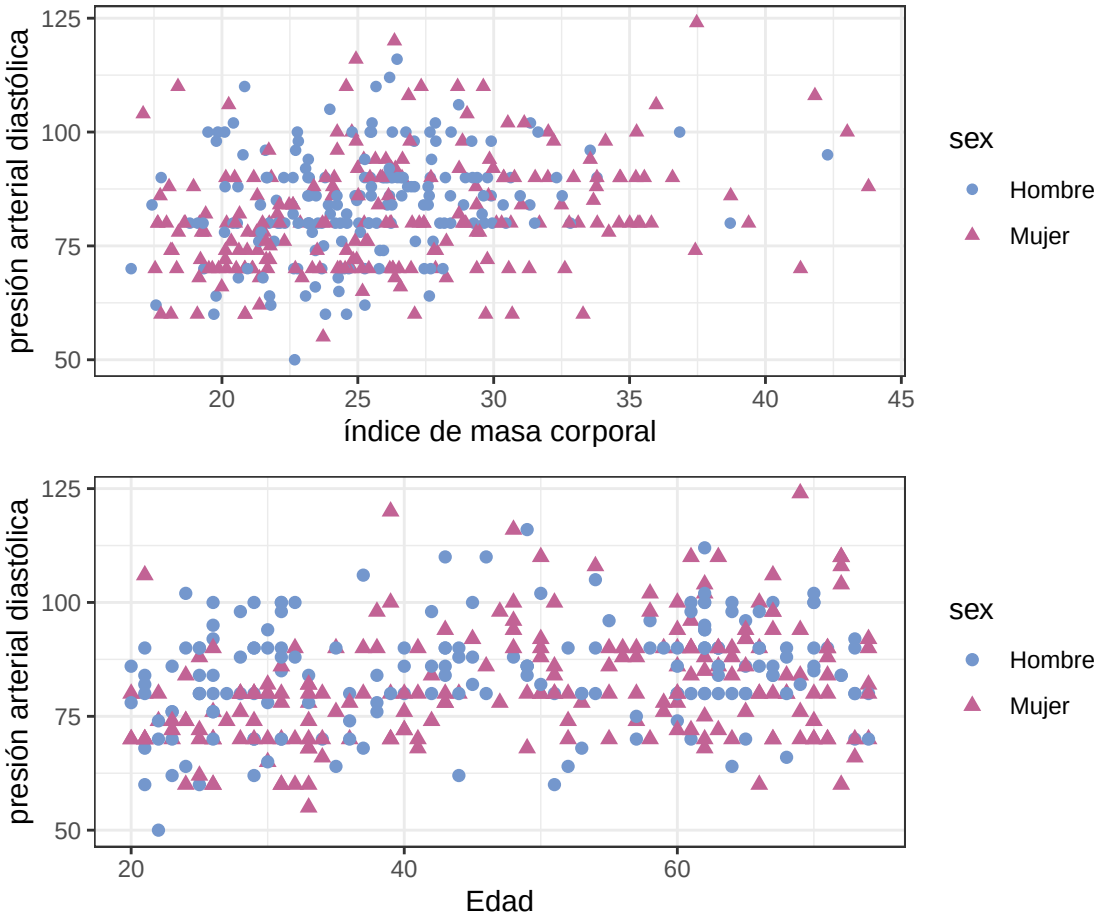



Figura 1: Presión diastólica contra el índice de masa corporal y la edad

Cuadro 1: Medidas de tendencia

| Variable | Promedio | Mediana |
|---|---|---|
| Presión | 83 | 80 |
| BMI | 26 | 25 |
| Edad | 48 | 49 |

Según la Figura 1 en ambos casos parece haber una relación creciente, con forme crece el bmi y la edad, crece la presión además se nota una variabilidad de presión más o menos constante. Derivado de esto, una regresión lineal múltiple puede ayudarnos a modelar nuestros datos de una manera adecuada.

Tras ajustar un primer modelo sin transformar notamos problemas con Linealidad y Normalidad, para linealidad, con un p-value de 0.01553 se rechazó Tukey test y para normalidad, se usaron los errores estándarizados $e_{st}$, con un p-value de 0.01192 y 0.003654 se rechazarón las pruebas Shapiro y Kolgomorov respectivamente. Para solucionar estos problemas hacemos uso de una transformación tipo Box-Cox y Box-Tidwell las cuales suguieren una transformación logarítmica y un exponente cercano a menos uno respectivamente, así optamos por ajustar un segundo modelo:

$$\mathbb{E}[\log(bpdiast); bmi, sex, age] = \beta_0 + \beta_1 bmi + \beta_2 sex + \beta_3 age^{-1}$$

Que también puede verse como:

$$\mathbb{E}[\text{bpdiast}] = e^{\beta_0+\beta_1 bmi+\beta_2 sex+\beta_3 age^{-1}+\frac{\sigma^2}{2}}$$

De manera rápida la Figura 2 no parece mostrar evidencia estadística en contra de nuestros supuestos, para un análisis más completo se realizarón las pruebas (y no se rechazaron) Tukey-test para linealidad, la bptest (lmtest) y ncvTest (car) para Homocedasticidad, la Shapiro y Kolgomorov (nortest) para Normalidad y la bgtests y dwtest (lmtest) para Independencia, no tenemos valores influyentes que afecten nuestro análisis y la aleatoriedad fue garantizada por el investigador.

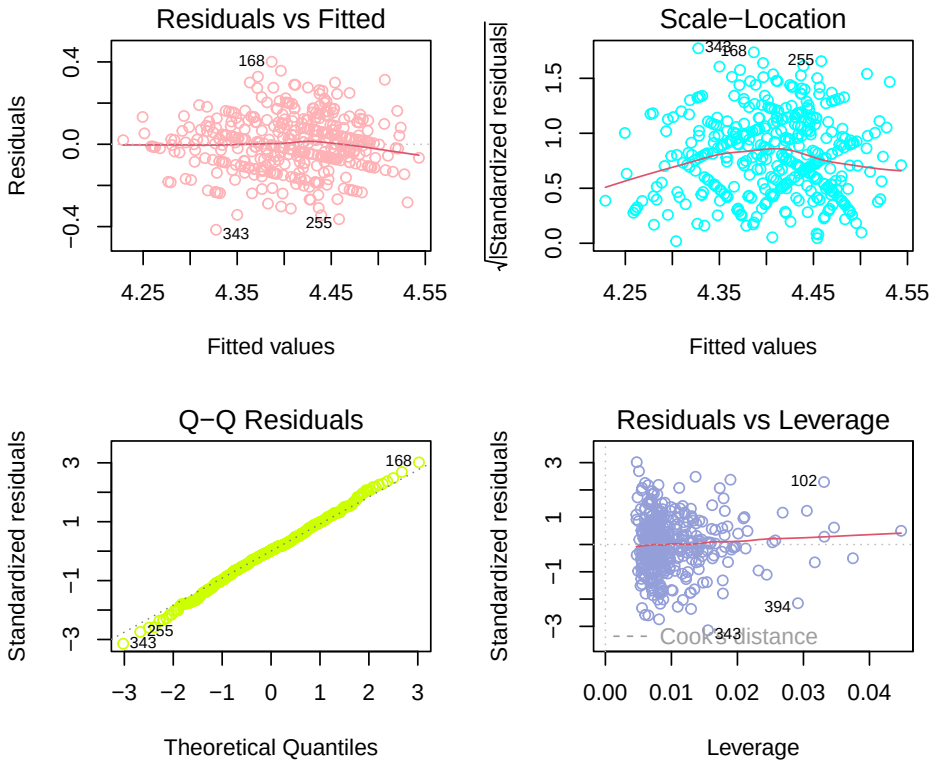



Figura 2: Supuestos modelo transformado

Ya que no se encontró evidencia estadística de que nuestro modelo no es adecuado, procedemos a trabajar con él. Tras revisar el summary con un p-value de $1.328e-15$ rechazamos la prueba F asociada a la tabla Anova por lo que nuestras variables parecen ser significativas, para los p-values individuales todos son menores a la significancia por lo no hay razón para considerar un modelo reducido.

¿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 ello notemos nos preguntan si las variables sexo y edad son inlfuyentes en la presión y al mismo tiempo la relación con el indice de masa corporal es creciente, esto para nuestro modelo se traduce a $\beta_2 \neq 0$, $\beta_3 \neq 0$ y $\beta_1 > 0$ todos al mismo tiempo, pero dado que la prueba F ya se rechazó nuestros $\beta_s$ son significativos, entonces nos basta con presentar la prueba:

$$H_0 : \beta_1 \leq 0 \ vs \ H_a : \beta_1 > 0$$

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = I(log(bpdiast)) ~ bmi + sex + I(age^-1), data = Datos1)
##
## Linear Hypotheses:
##        Estimate Std. Error t value   Pr(>t)
## 1 <= 0 0.006778   0.001381   4.908 6.73e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Con una confianza del 95 % y un p-value de $6.73e-07$, encontramos evidencia en los datos en contra de que $\beta_1 \leq 0$ por lo que es plausible que para cierto sexo y edad, en promedio una presión diastólica alta está asociada con un índice de masa corporal alto.

A fin de complementar los interpretación consideramos sólo tres edades; 30 años, 50 años y 64 años, en la Figura 3 se muestran los datos con esta nueva consideración.

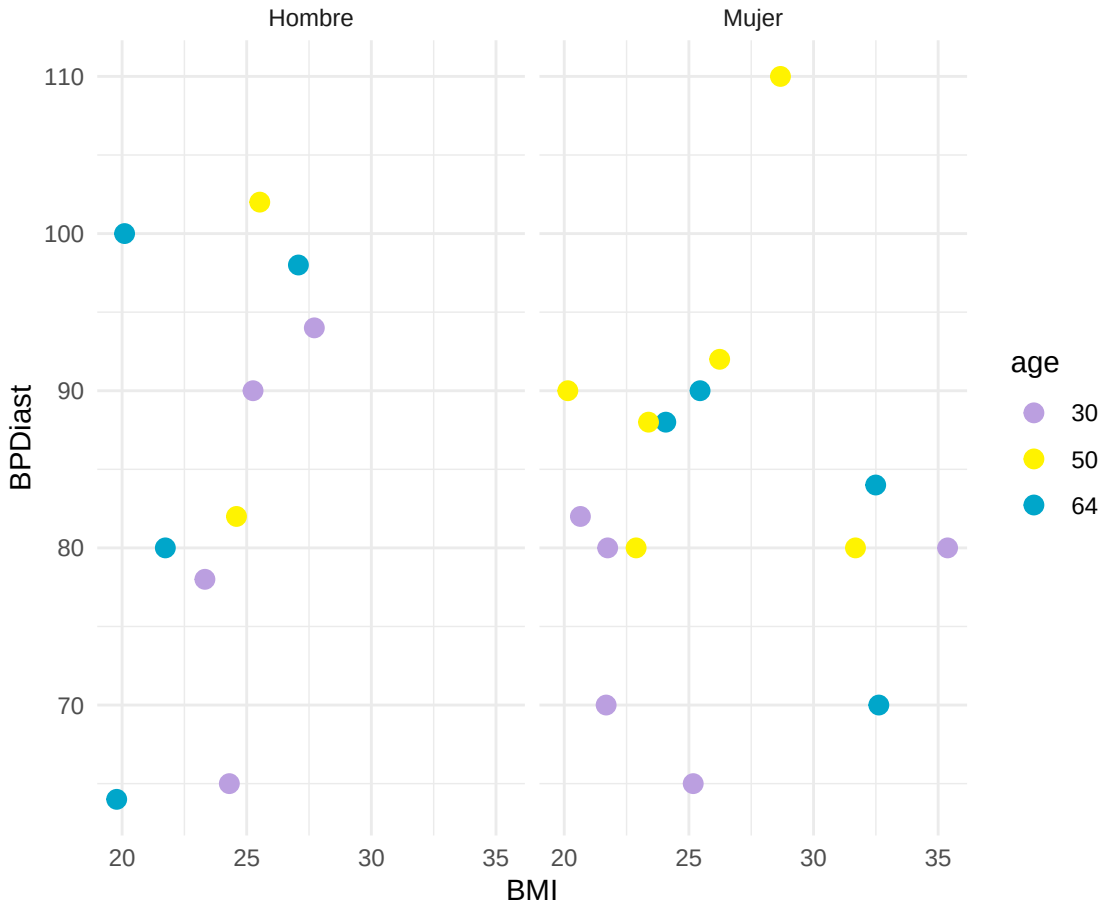



Figura 3: Relación entre BMI y presión diastólica por edad restringida y sexo

Según la gráfica anterior se pueden apreciar varias cosas; la presión diastólica de las mujeres tiende a ser más estable pues ronda entre 75 y 95 mientras que la de los hombres va desde 65 a 100, en las mujeres parece influir más la edad pues en las de 30 años la presión se mantuvo cercana a 80, en cambio las de 50 y 64 tuvieron una presión más alta y similar; esto podría deberse a varios factores como el estilo de vida y las enfermedades crónicas, aún así podemos notar la asociación entre el bmi y la presión parece ser más fuerte para los hombres que para las mujeres y en ambos grupos las personas jovenes no tienen tan marcada esta relación creciente.

Cuadro 2: Estimación de los betas por MV

| Betas | Estimados |
|---|---|
| b0 | 4.3570219 |
| b1 | 0.0067782 |
| b2 | -0.0456210 |
| b3 | -4.0296930 |
| Sigma^2 | 0.0177556 |

Para concluir, en la Tabla 2 tenemos las estimaciones para los betas a partir de los cuales podemos concluir estimaciones puntuales, por ejemplo un hombre de 46 años con un bmi de 26 tendrá en promedio 85 de presión diastólica, lo cuál concuerda con los datos de la Tabla 1. Más detalles en el chunk "Estimación puntual".

```
## [1] "Para un hombre de 46 años con bmi de 26.00 se estima que en promedio tendrá 85.09 de presión diastólica.
```

## 2. Inferencia sobre la presión arterial diastólica, a partir del índice de masa corporal usando modelos lineales generalizados para datos continuos

En la sección anterior se nos proporcionaron datos de índice de masa corporal y presión arterial diastólica sobre 400 pacientes seleccionados de forma aleatoria. Se busca determinar si hay suficiente evidencia para afirmar que tener un índice de masa corporal alto se asocia con una alta presión arterial diastólica. En esta ocasión haremos el análisis buscando presentar un modelo que parezca adecuado explorando los diferentes modelos lineales generalizados comúnmente usados cuando la variable dependiente es continua (normal, gamma, inversa gaussiana).

Comenzamos el análisis presentando los datos a continuación.

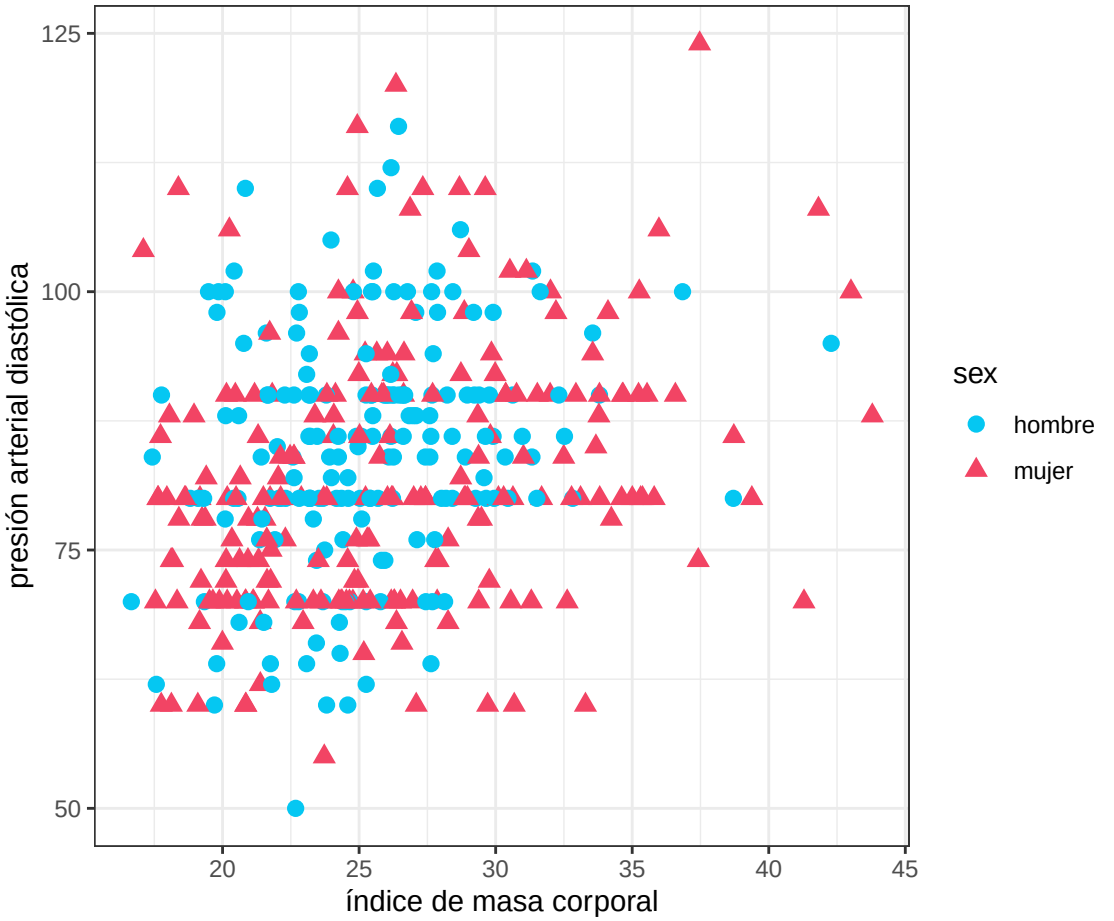



Figura 1: Observamos los datos de la presión arterial diastólica contra el índice de masa corporal

A partir de la Figura 1, podemos observar que parece existir una tendencia a que la presión arterial diastólica crezca conforme aumenta el índice de masa corporal. Vemos que la linealidad no es algo tan preciso, además de que la varianza no parece ser constante, pues los puntos parecen ir aumentando la dispersión conforme aumenta el índice de masa corporal. Derivado de lo anterior, analizaremos los casos normal, gamma e inversa gaussiana con modelos lineales generalizados, usando mallas para elegir el mejor modelo para los datos con la mejor distribución y liga.

Bajo el criterio AIC observamos los 3 mejores modelos para trabajar.

Cuadro 1: Mejores 3 modelos de la esperanza de la presión arterial diastólica

| Tipo | Info.Adicional1 | Info.Adicional2 | Fórmula | AIC |
|---|---|---|---|---|
| GLM | Familia:Gamma | Liga:Identity | bpdiast=beta_0+beta_1bmi^(1.5)+beta_2sex+beta_3age | 3057.721 |
| GLM | Familia:Gamma | Liga:Identity | bpdiast=beta_0+beta_1bmi^(2)+beta_2sex+beta_3age | 3057.761 |
| GLM | Familia:Gamma | Liga:Identity | bpdiast=beta_0+beta_1bmi+beta_2sex+beta_3age | 3057.932 |

Por una mejor interpretación, en la que podemos decir que usamos la variable del indice de masa corporal al cuadrado, la expresión matemática para modelar la esperanza de los valores de presión arterial que elegimos es la siguiente:

$$E[bpdiast|bmi, age, sex] = \beta_0 + \beta_1 bmi^2 + \beta_2 age + \beta_3 sex$$

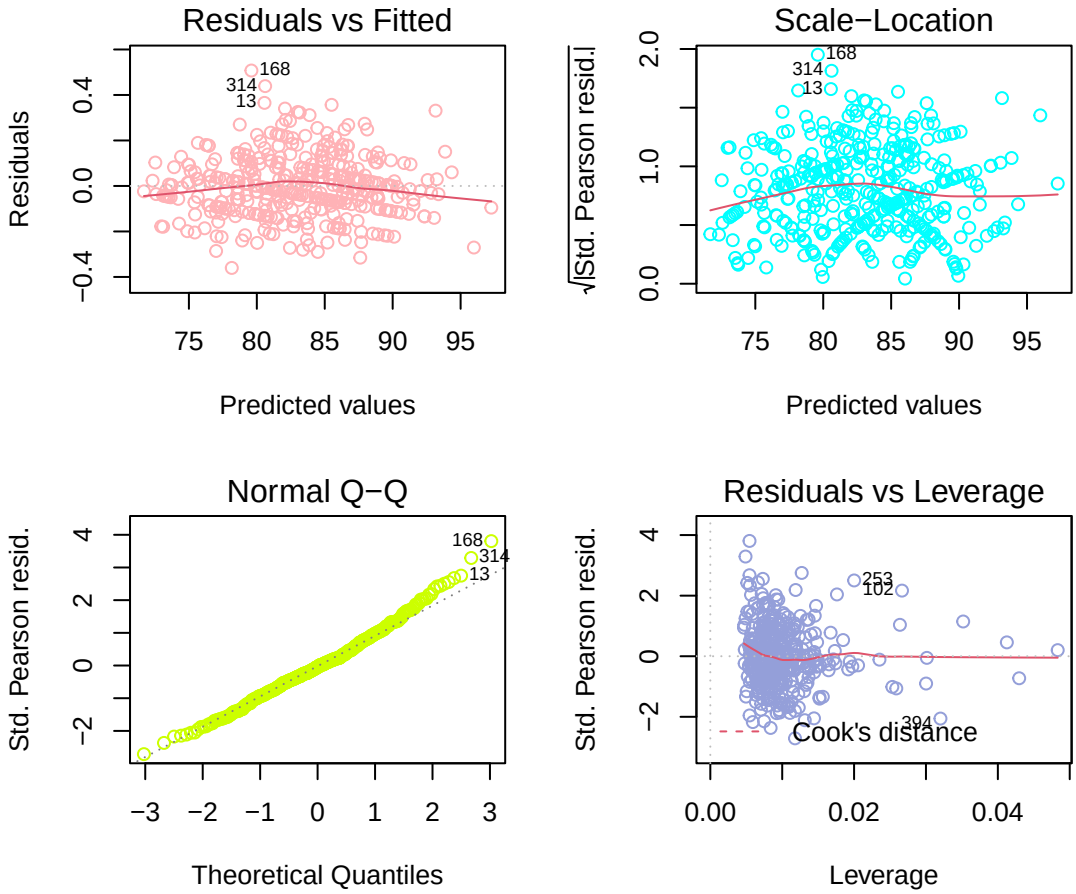



Figura 2: Verificación de supuestos ajuste

Cuadro 2: Pruebas de hipótesis en la verificación de supuestos

| Supuestos | Test | p.value |
|---|---|---|
| Linealidad | Pr(>\|Test stat\|) | I(bmi^2):0.9634 y age:0.7323 |
| Homocedasticidad | studentized Breusch-Pagan | 0.3822 |
| Normalidad | Shapiro-Wilk normality | 0.7348 |
| Normalidad | Lilliefors (Kolmogorov-Smirnov) normality | 0.1247 |

A partir de la anterior, y de la verificación supuestos que podemos apreciar en la Figura 2, decidimos utilizar el modelo antes mencionado, ya que no se encontró evidencia fuerte en contra de los mismos, pues los p-value obtenidos en las pruebas rechazan las hipotesis de que no se cumplan los supuestos, podemos ver estas pruebas usadas y sus respectivos p-value en la Tabla 2.

También este modelo es el segundo con el menor AIC. Además, gracias a la prueba F asociada a la tabla ANOVA, vista en el chunk "PruebaF" de nuestro R Markdown, se puede verificar que tener un indice de masa corporal alto sí afecta presión arterial diastólica.

**Preguntas del Investigador**

Queremos saber si a mayor indice de masa corporal entonces mayores niveles de presión arterial diastólica, y con lo anterior, se procedió a realizar una prueba de hipótesis para determinar si en efecto la presión arterial diastólica aumenta con un indice de masa corporal alto. Como nuestro modelo es sencillo de interpretar, la prueba es más directa. En ecuaciones se ve así: $E[bpdiast|bmi+1, age^*, sex^*] > E[bpdiast|bmi, age^*, sex^*]$ si y sólo si $\beta_0+\beta_1(bmi+1)^2+\beta_2 age+\beta_3 sex > \beta_0+\beta_1 bmi^2+\beta_2 age+\beta_3 sex$ si y sólo si $\beta_1((bmi+1)^2-bmi^2) > 0$ si y sólo si $\beta_1 > 0$, pues $bmi \geq 0$ Por lo tanto contrastamos $H_0 : \beta_1 \leq 0$ *vs* $H_a : \beta_1 > 0$

Y se realizó la prueba de hipótesis correspondiente, la cual se puede encontrar en el chunk “Prueba de Hipótesis”, y cuyos resultados se encuentran a continuación

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: glm(formula = bpdiast ~ I(bmi^2) + age + sex, family = Gamma(link = "identity"),
##     data = datos1)
##
## Linear Hypotheses:
##        Estimate Std. Error z value  Pr(>z)
## 1 <= 0 0.010693   0.002117   5.051 2.2e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Y a partir de los resultados obtenidos, con una significancia de .05, podemos afirmar que la presión arterial diastólica sí aumenta con el indice de masa corporal alto.

**Gráficas**

Vamos a ver gráficas de nuestro modelo ajustado contemplando solo las edades 30, 50 y 64, así como la diferenciación entre mujeres y hombres.

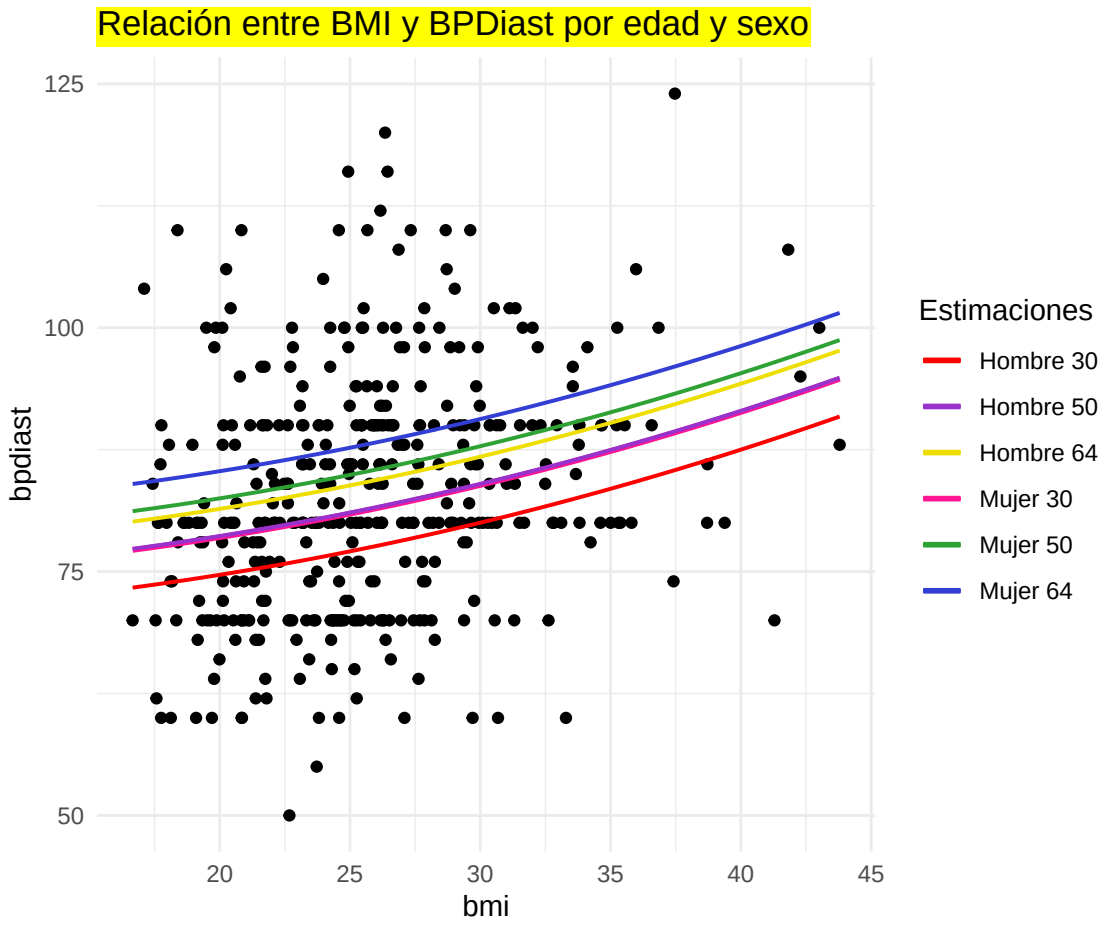



Figura 3: Relación entre el indice de masa corporal y la presión arterial diastólica por edad y sexo

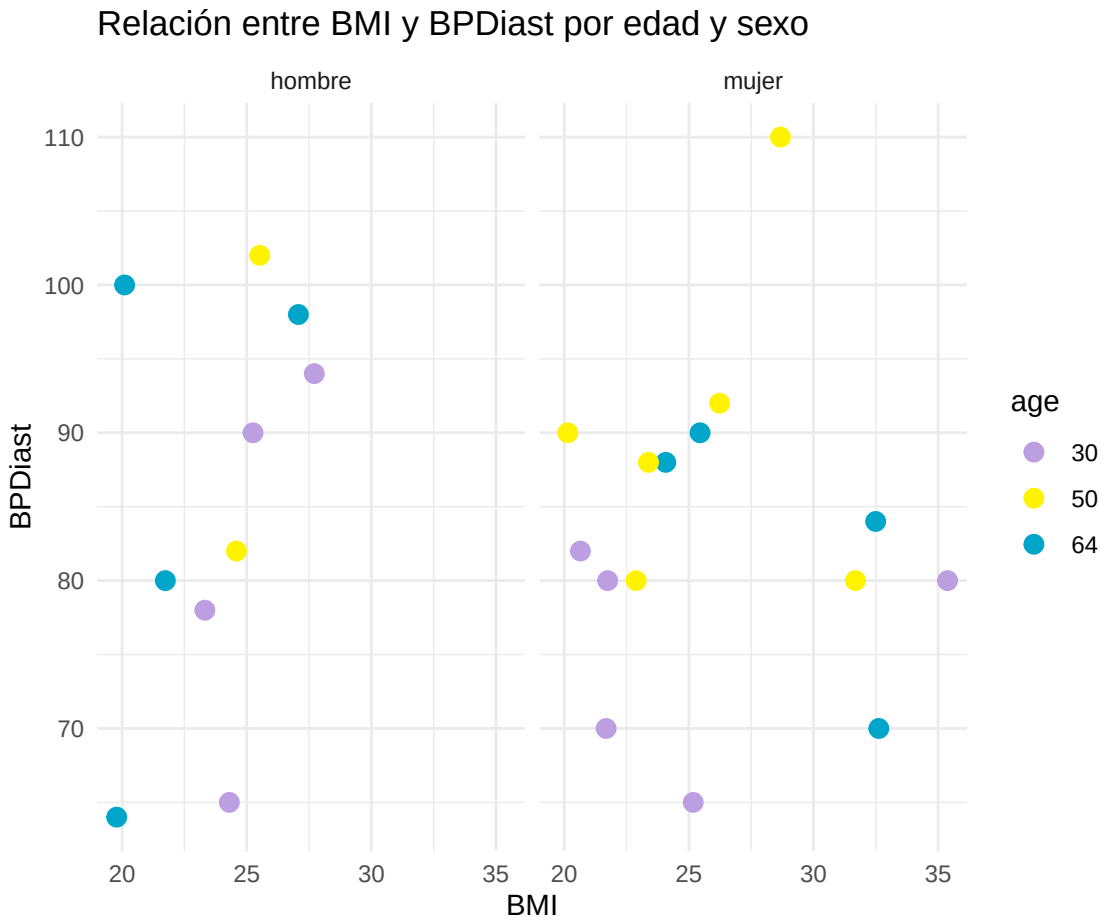



Figura 4: Relación entre el indice de masa corporal y la presión arterial diastólica por edad y sexo

En ambas Figuras 3 y 4 podemos ver los mismos resultados que en la sección anterior: la presión diastólica de las mujeres tiende a ser más estable pues ronda entre 75 y 95 mientras que la de los hombres va desde 65 a 100, en las mujeres parece influir más la edad pues en las de 30 años la presión se mantuvo cercana a 80, en cambio las de 50 y 64 tuvieron una presión más alta y similar; esto podría deberse a varios factores como el estilo de vida y las enfermedades crónicas, aún así podemos notar la asociación entre el bmi y la presión parece ser más fuerte para los hombres que para las mujeres y en ambos grupos las personas jovenes no tienen tan marcada esta relación creciente.

**Elección de un Modelo Definitivo**

Hacemos cambio de variable para tener ambos modelos transformados a mismas escalas y poder comparar de manera efectiva ambos modelos.

Cuadro 3: Modelo de RLM contra GLM

| Tipo | Modelo | AIC |
|---|---|---|
| RLM | E[log(bpdiast);bmi,sex,age]=beta_0+beta_1bmi+beta_2sex+beta_3age^{-1} | 3055.34 |
| GLM | E[bpdiast\| bmi, age, sex] = beta_0 + beta_1bmi^2 + beta_2age + beta_3sex | 3057.761 |

El modelo lineal generalizado es una interpretación directa sobre la variable bpdiast, esto facilita el entender cómo afecta el cambiar una variable respecto a la otra. También, si queremos hacer predicciones, o tener mayor precisión en la estimación, nos quedamos con el modelo de regresión lineal múltiple pues parece que su crecimiento exponencial con valores altos de indice de masa corporal es más acertado que el de la suposición de que las observaciones provienen de una Gamma. El problema con éste, es que la interpretación es muy complicada por las transformaciones hechas a la variable bpdiast.

Nosotros nos guiaremos más por el criterio de AIC con los resultados vistos en la Tabla 3, y por ello elegimos el primer modelo de RLM $\mathbb{E}[\log(bpdiast); bmi, sex, age] = \beta_0 + \beta_1 bmi + \beta_2 sex + \beta_3 age^{-1}$ sobre el segundo modelo de GLM $E[bpdiast|bmi, age, sex] = \beta_0 + \beta_1 bmi^2 + \beta_2 age + \beta_3 sex$, con familia Gamma y función liga identidad.

# Inferencia sobre la eficacia de tres insecticidas usando modelos lineales generalizados para datos binarios

Se registro información sobre 862 insectos expuestos a diferentes dosis de tres insecticidas distintos, además se registro el número de insectos muertos y el número total de insectos expuestos. Es de interés identificar para cada insecticidad la dosis mínima con la que muere el 70 % de los insectos y determinar cual es el mejor insecticida de los tres.

I) Presente una gráfica de dispersión

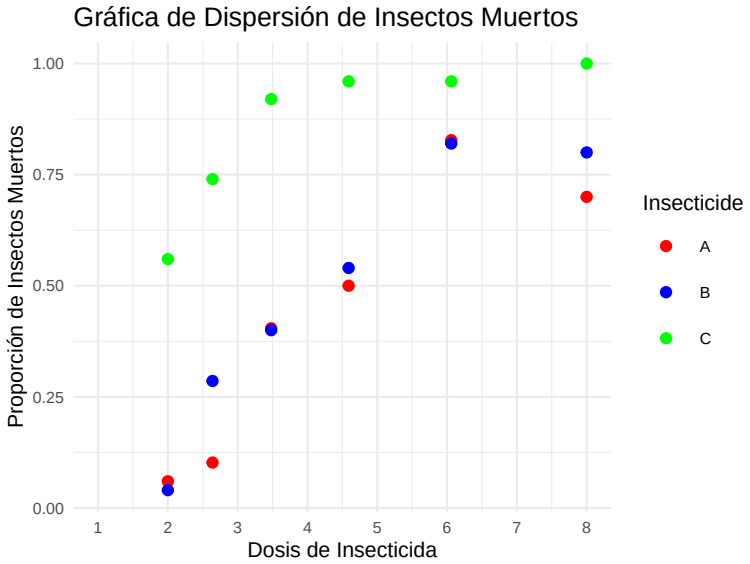



Figura 1: DosisvsProporcion

La gráfica **Dosis vs Proporcion** muestra la proporción de insectos muertos según la dosis de insecticida para los tipos de insecticida A, B y C. Todas presentan efectos a partir de 2 mg. Para los insecticidas A y B, una dosis de 2mg genera menos del 25 % de mortalidad, mientras que con el insecticida C supera el 50 %. Se puede notar que a medida que se incrementa la dosis la proporción de insectos muertos crece. Esta tendencia es uniforme para el insecticida C, pero para los insecticidas A y B varía tras superar los 6 mg. En cada dosis, el insecticida C demuestra mayor mortalidad que A y B, lo cual es un indicador positivo a favor su efectividad.

II) Ajusta modelos para datos binarios (ligas: logit, probit, cloglog) en donde incluya como covariables a Insecticide y Deposit, así como su interacción. Describa las expresiones del componente lineal o sistemático para cada insecticida como función de la dosis. Indique si alguno de los modelos parece adecuado para realizar el análisis deseado.

## Descripción del componente lineal para cada tipo de insecticida

A continuación se muestran las ecuaciones que nos dan la descripción del componente lineal para cada tipo de insecticida:

$$\begin{aligned}\eta(Y;\text{Deposit, Insecticide}) = b_0 &+ b_1 \cdot I(\text{Insecticide} = \text{B}) + b_2 \cdot I(\text{Insecticide} = \text{C}) \\ &+ b_3 \cdot \text{Deposit} + b_4 \cdot \text{Deposit} : I(\text{Insecticide} = \text{B}) \\ &+ b_5 \cdot \text{Deposit} : I(\text{Insecticide} = \text{C}),\end{aligned} \tag{1}$$

## Tabla con modelos ajustados

Lós códigos para cada nivel de referencia son: *** : 0.001, ** : 0.01, * : 0.05, . : 0.1, : 1 Tras conocer el significado de los códigos de significancia, interpretamos los coeficientes del modelo. En el primer modelo, los coeficientes del Intercepto, InsecticidaC, Deposit e interacción entre InsecticidaC y Deposit son significativos con niveles de 0.001, 0.1, 0.001 y 0.05. En los modelos dos y tres, los coeficientes significativos son el Intercepto, insecticida C y Deposit. En los tres modelos, el coeficiente del insecticida C es positivo y mayor a 1, indicando que su efecto es mayor que el insecticida A, referencia en los modelos. Esta observación concuerda con la gráfica previa. El modelo con menor AIC es el primero con un AIC de 116.97.

| Variable | Estimate (logit) | Estimate (probit) | Estimate (cloglog) |
|---|---|---|---|
| (Intercept) | -2.90282*** | -1.76573*** | -2.359889*** |
| datos3$InsecticideB | 0.09191 | 0.08449 | 0.229117 |
| datos3$InsecticideC | 1.31766 . | 1.11190** | 1.719295*** |
| datos3$Deposit | 0.55965*** | 0.33647 | 0.357678*** |
| datos3$InsecticideB:datos3$Deposit | 0.06241 | 0.03072*** | 0.008812 |
| datos3$InsecticideC:datos3$Deposit | 0.43252* | 0.14821 | -0.005942 |
| AIC | 116.97 | 119.1 | 134.93 |

Cuadro 1: Estimación de coeficientes para cada modelo ajustado

- Ajuste modelos para datos binarios (ligas: logit, probit, cloglog) en donde adicional a las covariables incluidas en ii), también incluya a la interacción de Insecticide con Deposit2. Describa las expresiones del componente lineal o sistemático para cada insecticida como función de la dosis. Indique si alguno de los modelos parece adecuado para realizar el análisis deseado y si tiene alguna ventaja la inclusión de los términos cuadráticos en el modelo.

$$\begin{aligned}\eta(Y;\text{Deposit, Insecticide}) = b_0 &+ b_1 \cdot I(\text{Insecticide} = \text{B}) + b_2 \cdot I(\text{Insecticide} = \text{C}) \\ &+ b_3 \cdot \text{Deposit} + b_4 \cdot I(\text{Deposit}^2) : I(\text{Insecticide} = \text{A}) \\ &+ b_5 \cdot I(\text{Deposit}^2) : I(\text{Insecticide} = \text{B}) \\ &+ b_6 \cdot I(\text{Deposit}^2) : I(\text{Insecticide} = \text{C}),\end{aligned} \tag{2}$$

| Variable | Estimate (logit) | Estimate (probit) | Estimate (cloglog) |
|---|---|---|---|
| (Intercept) | -6.23563*** | -3.66805*** | -4.56019*** |
| datos3$InsecticideB | 0.24578 | 0.14632 | 0.18296 |
| datos3$InsecticideC | 2.78197*** | 1.69567*** | 1.98451*** |
| datos3$Deposit | 2.10891*** | 1.23397*** | 1.35895*** |
| datos3$InsecticideA:I(datos3$$Deposit^2$) | -0.15109*** | -0.08753*** | -0.09410*** |
| datos3$InsecticideB:I(datos3$$Deposit^2$) | -0.14880*** | -0.08608*** | -0.09344*** |
| datos3$InsecticideC:I(datos3$$Deposit^2$) | -0.14087*** | -0.08905*** | -0.10986*** |
| AIC | 90.407 | 90.346 | 99.019 |

Cuadro 2: Estimación de los coeficientes para cada modelo ajustado

La tabla 2 muestra que, con una significancia del 0.001, todos los estimadores son significativos, excepto el coeficiente de la variable Insecticide B. El coeficiente de Insecticide C es positivo y supera 1,indicando que su efecto sobre la variable respuesta es mayor que el de Insecticide A, el insecticida del nivel de referencia. Los coeficientes de interacción entre Insecticide y Deposit al cuadrado son negativos en los tres modelos. Aunque pequeños en valor, son estadísticamente significativos al nivel 0.001. Esto sugiere que los modelos que incluyen la interacción de niveles de insecticida y dosis al cuadrado mejoran considerablemente respecto a los modelos previamente ajustados en el inciso ii. Esta mejora se refleja en índices AIC's más bajos en comparación con los modelos ajustados en el inciso anterior.

## Cuarto inciso: Verificación de supuestos

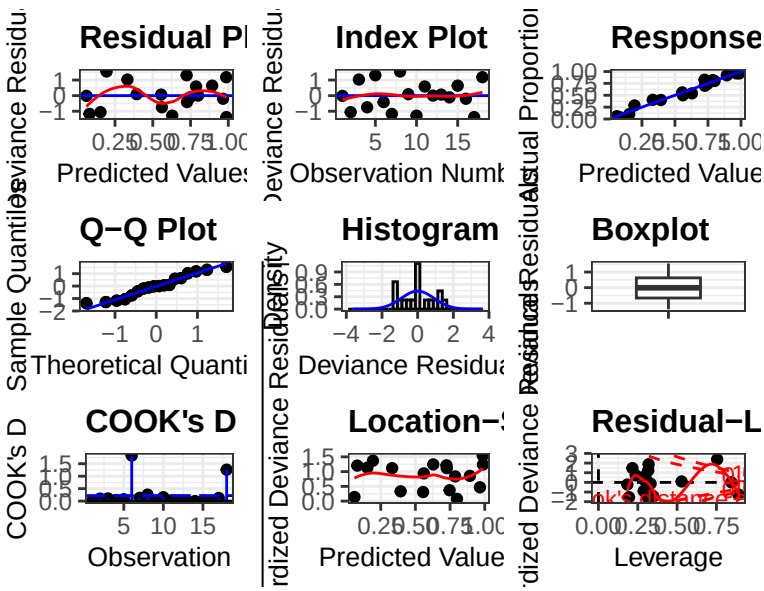



Figura 2: Verificacion

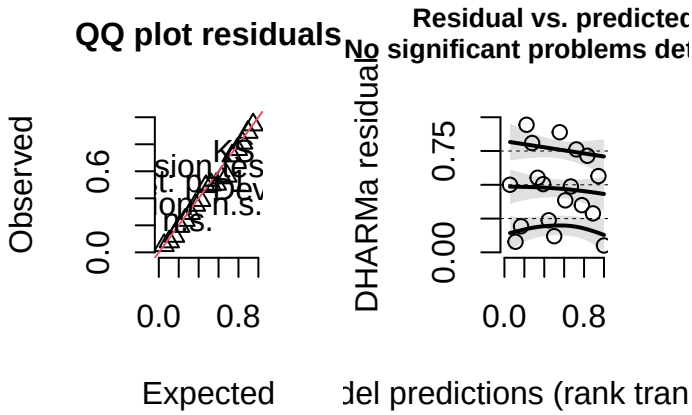


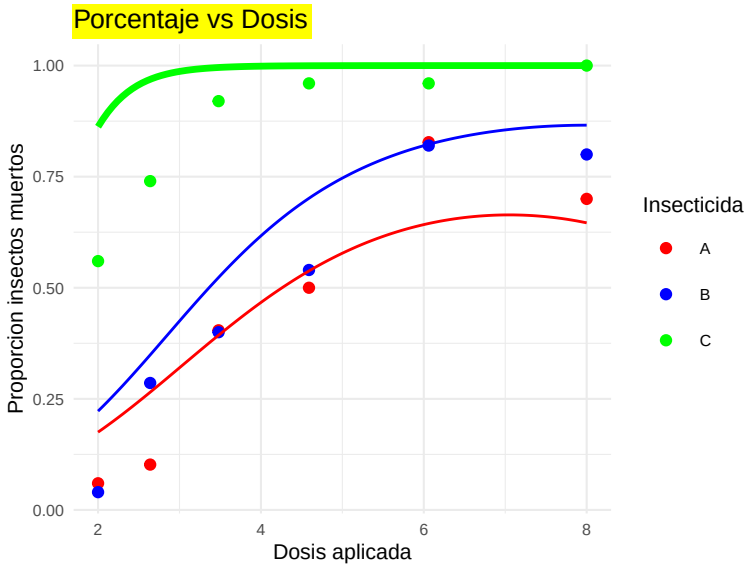

Figura 3: Residuales DHARMA

## Verificación de supuestos

Las gráficas que se muestran en la figuras 2 y 3 muestran en **Residual Plot** e **Index Plot** una nube de puntos alrededor de la línea del 0, sugiriendo que no hay evidencia en contra de los supuestos de linealidad y aleatoriedad. La gráfica **Response vs Predicted** presenta puntos formando una recta, indicando un buen ajuste del modelo. En **Cook's D Plot**, solo una observación supera la distancia de Cook de 1.5. El **Location-Scale plot** sugiere una estimación adecuada de la varianza y en **Residual-Leverage Plot**, casi todas las observaciones, salvo una, están dentro de las líneas de contorno de Cook. Considerando los residuales simulados, el modelo respeta el supuesto de linealidad: en los tests QQ-plot residuals, no se rechaza la hipótesis nula, y no hay evidencia contra los supuestos de linealidad y aleatoriedad. La gráfica **Residual vs Predicted** de los residuales DHARMA confirma que no hay evidencia en contra de los supuestos. En conclusión, no hay información contraria a los supuestos, validando el modelo para el análisis.

Sólo con el modelo que considere más adecuado entre los que se ajustaron en ii) y iii) a) Presente en la misma gráfica generada en i) los resultados de la estimación puntual para el valor esperado de la variable binaria (probabilidad de que un insecto muera).

A continuación se presenta la misma gráfica generada en el inciso i) pero con los resultados de la estimación puntual para el valor esperado de la probabilidad de que un insecto que ha sido expuesto a algún insecticida muera pero con la misma dosis.




- b) Calcule la dosis mínima para cada insecticida con la que se puede indicar que el 70 se muere.

Las funciones encontradas para cada uno de los insecticidads de acuerdo al ajuste que elegimos, son las siguientes: La encontrada para el inciso A es:

$$\phi^{-1}(x) = -0.09410x^2 + 1.35895x - 4.56019 \tag{3}$$

con

$$\phi^{-1}(0.7) = 0.5244005 \tag{4}$$

El valor de x que resuelve esta ecuación cuadrática es: 7.22077577045696+1.37633674207882i. Notemos que este valor pertenece a los números complejos, por tanto esto nos indica que para el insecticida A no existe un valor de la dosis para la cual la proporcion de insectos muertos sea mayor al 70 %. Esto es consistente con lo que se observa en la gráfica anterior.

La función encontrada para el insecticida B es:

$$\phi^{-1}(x) = -0.09344x^2 + (0.18296 + 1.35895)x - 4.56019 \tag{5}$$

con

$$\phi^{-1}(0.7) = 0.5244005 \tag{6}$$

El valor de x que resuelve esta ecuación cuadrática es: 4.5548348292549 Por lo tanto,la dosis a partir de la cual mueren el 70 % de los insectos al aplicarles el insecticida B es: 4.554 mg

La encontrada para el inciso C es:

$$\phi^{-1}(x) = -0.10986x^2 + (1.98451 + 1.35895)x - 4.56019 \tag{7}$$

con

$$\phi^{-1}(0.7) = 0.5244005 \tag{8}$$

El valor de x que resuelve esta ecuación cuadrática es: 1.60544781898825 Por lo tanto,la dosis a partir de la cual mueren el 70 % de los insectos al aplicarles el insecticida C es: 1.605 mg

Se puede observar que, de entre los tres insecticidas, la dosis mínima para la cual se cumple que muere el 70 % de los insectos corresponde al insecticida C.

c) Considerando la menor de las dosis encontradas en b), ¿se puede indicar que un insecticida es el mejor? Realice una prueba de hipótesis para argumentar en favor o en contra.

Se realizó la siguiente prueba de hipótesis simultánea para determinar cual es el mejor insecticida:

$$\begin{aligned} &H_0 : \beta_1 \leq \beta_0 \quad vs. \quad H_a : \beta_1 > \beta_0 \\ &H_0 : \beta_1 \leq \beta_0 \quad vs. \quad H_a : \beta_2 > \beta_0 \\ &H_0 : \beta_1 \leq \beta_0 \quad vs. \quad H_a : \beta_2 > \beta_1 \end{aligned} \tag{9}$$

En la primera prueba de hipótesis se falla al rechazar la hipótesis nula mientras que para las hipótesis dos y tres se logra rechazar la hipótesis nula con una significancia del 0.001. De esta manera se concluye que el efecto del insecticida C sobre la variable respuesta es mayor al efecto del insecticida A o el insecticida B, por lo tanto es posible concluir que el insecticida C es mas efectivo que el insecticida A y el insecticida B. Esto es consistente con lo que observamos en la gráfica del inciso a) y con lo concluido en el inciso b)

d) En general ¿se puede indicar que los insecticidas A y B tienen un desempeño similar? Realice una prueba de hipótesis para argumentar en favor o en contra.

Tras realizar la siguiente prueba de hopótesis $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$ observamos que se rechaza la hipótesis nula, además cuando observamos el resultado del summary del modelo ajustado, Ajuste22, podemos observar que el estimador del coeficiente asociado a la variable InsecticideB no es estadísticamente significativo por lo tanto concluimos que el insecticida A y el insecticida B tienen un desempeño similar.

## 4.- Inferencia sobre el número de casos de cáncer de pulmón en cuatro ciudades de Dinamárca usando modelos lineales generalizados para datos de conteos.

Se registró el número de casos de cáncer de pulmón entre 1968 y 1971 en cuatro ciudades de Dinamarca. También se registró la edad de los pacientes, para propósitos de este análisis se trata como variable categórica de 5 niveles. Es de interés conocer si a mayor edad existe mayor incidencia de cáncer de pulmón.

I. Presente una gráfica de dispersión en donde en el eje x se incluyan los grupos de edad (ordenados de menor edad a mayor) y en el eje y la tasa de incidencia (Cases/Pop) por cada cruce Age-City, distinguiendo con un color la Ciudad. Describa lo que se observa.

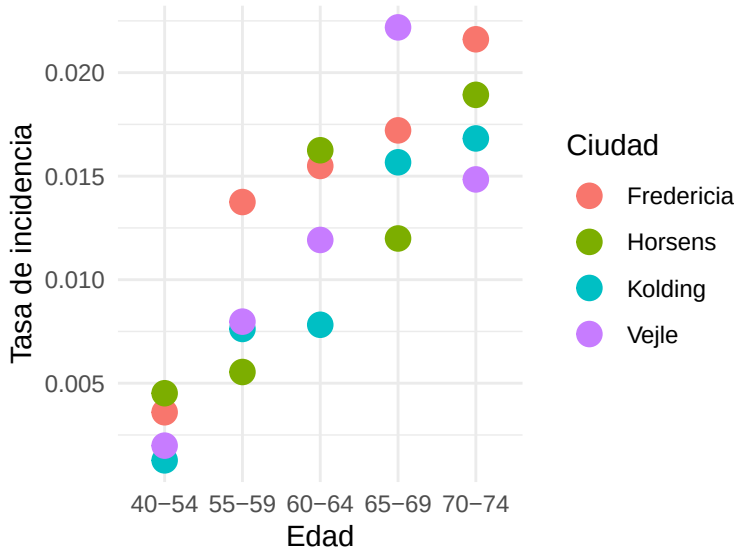



Figura 1: GraficaCiudades

En la primera gráfica se ilustra la correlación entre la edad y la tasa de incidencia, refiriéndose esta última a la proporción de casos de cáncer de pulmón por ciudad. Se puede apreciar que en todas las ciudades, a medida que aumenta la edad de los individuos, se registra una tasa de incidencia más alta. En otras palabras, en todas las ciudades existe una relación directa entre la edad avanzada de los ciudadanos y una mayor tasa de incidencia de casos de cáncer de pulmón.

II. Como un primer modelo considere la distribución Poisson con liga logarítmica y las covariables Age y City, así como su interacción. Dado que las dos covariables son categóricas, este modelo con interacciones tiene muchos parámetros y es deseable trabajar con uno más simple. Para esto considere un segundo modelo donde sólo se usa como covariable a Age. Realice una prueba de hipótesis para argumentar si es posible considerar el segundo modelo [recuerde que dado que los modelos son anidados, podría usar la función anova(mod1, mod2, test = "Chisq"), también puede usar multcomp, pero hay muchos parámetros y podría ser tedioso]. Complemente su decisión con lo que se observa en la gráfica en i) y con medidas como AIC o BIC.

Se ajustaron tres modelos lineales generalizados con distribución Poisson y liga logaritmo, además en cada uno de los modelos se agrego un término offset lo cual involucra incluir una variable adicional igual al logarítmo de la variable población; el primer modelo ajustado toma en cuenta la variable Age, City así como la interacción entre estas dos variables, el segundo modelo ajustado únicamente toma en cuenta la variable Age y el tercer modelo ajustado toma en cuenta las variables Age y City. Posteriormente, al incluir los términos offset en los tres modelos ajustados, se llevó a cabo tres pruebas de hipótesis F asociada a la tabla ANOVA entre pares de modelos, es decir se comparan los modelos 1 contra 2, 1 contra 3 y 2 contra 3. Tras realizar estas prueba, concluimos que el segundo modelo ajustado se presenta como el más adecuado de todos para proceder con el análisis.

III. Considerando el modelo seleccionado en ii), ajuste un modelo binomial negativo. Compare ambos modelos e indique cuál podría ser adecuado para realizar el análisis deseado. Con el modelo seleccionado, calcule intervalos de confianza simultáneos de las tasas de incidencia para cada grupo de edad, incluya estos en la gráfica presentada en i). Comente los resultados, en particular si se puede indicar que a mayor edad existe mayor incidencia de cáncer de pulmón.

Inicialmente, se ajustó un modelo binomial negativo para contrastarlo con el segundo modelo ajustado en el inciso anterior, el cual fue seleccionado entre tres modelos ajustados. Al analizar los valores de los criterios AIC y BIC y comparar estos entre ambos modelos, se concluyó que el modelo binomial negativo ajustado es el más adecuado para realizar el análisis, dado que presenta el menor valor en el criterio BIC.

El modelo se ve de la siguiente manera:

$$\log(E[\text{Casos}]) = \beta_0 + \beta_1 I(55-59) + \beta_2 I(60-64) + \beta_3 I(65-69) + \beta_4 I(70-74) + \text{offset} \tag{1}$$

La interpretación de los coeficientes correspondientes al grupo de edad en el modelo seleccionado es la siguiente: en el grupo de 55 a 59 años, la tasa de casos de cáncer se incrementa en un 8 %; en el grupo de 60 a 64 años, los casos de cáncer aumentan en un 50 %; y para el grupo de 65 a 69 años, la tasa de cáncer se eleva en un 84 %.

De esta manera, en todas las ciudades se observa la misma tendencia: a medida que aumenta la edad de los grupos, la tasa de casos de cáncer de pulmón también se eleva. Por lo tanto, podemos deducir que hay una mayor incidencia de casos de cáncer de pulmón conforme incrementa la edad en todas las ciudades analizadas. Una vez elegido el modelo binomial negativo procedemos a verificar que se cumplan los supuestos de linealidad, normalidad y homocedasticidad. Se verificaron usando por el método de residuales simulados y no se encontro evidencia en contra de ninguno de nuestros supuestos.

IV. Los incisos anteriores usaron a la variable Age como categórica, sin embargo, eso dificulta un poco la interpretación, además de que por su naturaleza esa variable se podría haber registrado sin categorizar. Con los datos actuales, una aproximación sería usar el punto medio de cada intervalo de edad que define las categorías de Age y usar la variable resultante como una variable continua, llámela Ageprima. Ajuste modelos usando la distribución Poisson y Binomial Negativa con la covariable Ageprima, también considere la opción de incluir a Ageprima2 . Entre esos 4 modelos indique cuál podría ser adecuado para realizar el análisis. Con ese modelo indique si a mayor edad existe mayor incidencia de cáncer de pulmón, por ejemplo, analizando si la función es creciente considerando que el intervalo de edad que es de interés es entre 40 y 74 años. Presente una gráfica que complemente su análisis.

Se procedió al ajuste de dos modelos lineales generalizados, uno con distribución Poisson y otro Binomial Negativa, incorporando una variable continua llamada AgePrima. Para seleccionar un modelo, se recurrió a los criterios AIC y BIC, optando por el modelo que incluye la variable AgePrima junto con el offset para proseguir con el análisis. Posteriormente, se ejecutó una prueba lineal y se calcularon intervalos de confianza con el objetivo de determinar si existe una correlación entre la mayor edad y una mayor incidencia de cáncer de pulmón. Se descubrió que al pasar de los 40 a los 70 años, la tasa de incidencia se incrementa de 0.000951 a 0.000969, lo que, aunque representa un aumento mínimo, confirma la correlación deseada a través de este análisis.

# 5.-Inferencia en la satisfacción de habitantes con modelos ligit-multinomial.

Se registró el nivel de satisfacción (Sat) para 1681 personas, también se nos proporcionó el tipo de vivienda (Type), el nivel de influencia (Infl) en las decisiones para el mantenimiento y el nivel de contacto (Cont) con los otros inquilinos. Es de interés saber si los factores antes mencionados influyen en el nivel de satisfacción.

Comenzamos el análisis presentando en la Gráfica 1 la proporción de satisfacción por cada categoria en nuestros datos.

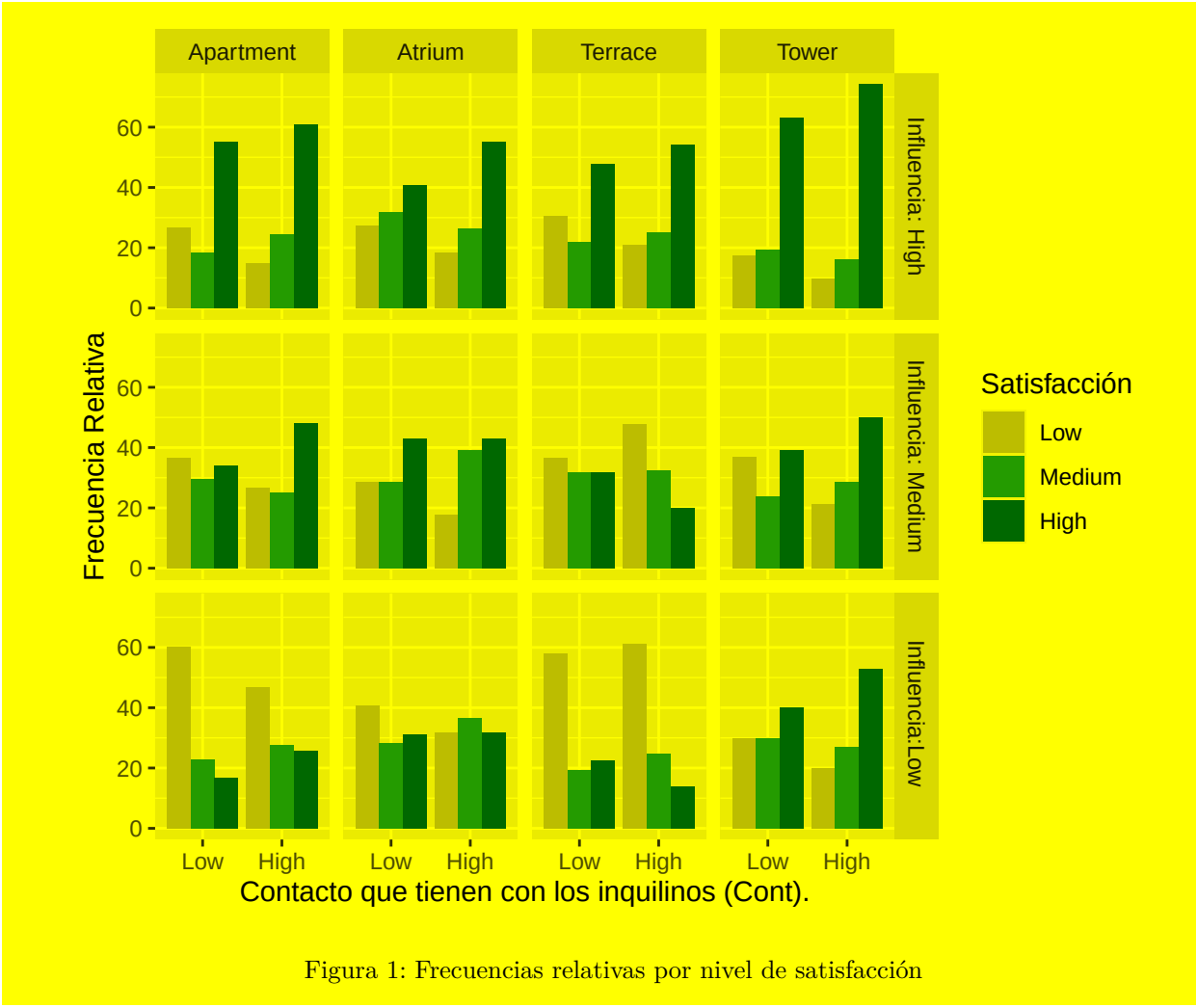



Figura 1: Frecuencias relativas por nivel de satisfacción

Podemos intuir que el tipo de vivienda más favorable es Tower mientras que la Terrace parece ser la menos querida, el que tanto se sienten incluidos en las decisiones de mantenimiento (Infl) también tiene un impacto significativo pues a mayor influencia parece ser mayor el nivel de sastisfacción aunque el contacto que tienen con otros inquilinos también muestra una ligera tendencia creciente en casi todos los lugares. Una baja influencia y vivir en la Terrace o Apartment muestran descontentos similares pero curiosamente la Terrace tiene más descontento cuando se tiene más contacto con otros inquilinos, esto probablemente se deba a que el contacto no suele ser tan profundo. No hay tanta variabilidad en la satisfacción cuando se tiene una influencia mediana independientemente del lugar y el contacto. De todos los factores el sentido de pertenencia (Infl) parece ser el más significativo, posiblemente habrá que destinar recursos a esta área.

Dadas las características del problema (ANOVA y variable de respuesta con más de dos categorías) es buena idea trabajar con un modelo logístico multinomial, se consideró la variable de respuesta como nominal y ordinal. Ajustamos para cada caso con vglm de VGAM un modelo, se verificó que tuvieran sentido por medio de modelos nulos, para el caso nominal se contrastó uno con todas las interacciones y otro que únicamente es de efectos principales, para el caso ordinal se contrastó el uso del supuesto de probabilidad. Todas la pruebas se pudieron realizar con anova() y lrtest() (Tabla 1) pues los modelos son anidados.

De acuerdo a la Tabla 1 no se encontró evidencia en contra para usar el modelo de efecto principales nominal y el ordinal bajo el supuesto de probabilidad, además que ambos son significativos pues se rechaza el modelo nulo. Por cuestiones de interpretabilidad y criterios como AIC y BIC (Tabla 2) optamos por trabajar con el modelo ordinal de odds proporcionales.

Cuadro 1: Pruebas Anova

| CONTRASTE | P-value |
|---|---|
| Nominal-Nulo-vs-Simple | 2.2e-16 |
| Interacciones-vs-Simple | 0.2671 |
| Ordinal-Nulo-vs-Simple | 2.2e-16 |
| Probabilidad | 0.1992 |

Cuadro 2: Comparación de AIC y BIC

| Modelo | AIC | BIC |
|---|---|---|
| Nominal_interacciones | 3527.422 | 3787.925 |
| Nominal_SINinteracciones | 3498.084 | 3574.064 |
| Ordinal_NP | 3498.579 | 3574.559 |
| Ordinal_P | 3495.149 | 3538.566 |

La categoría de referencia para las variables explicativas es Apartment en la variable Type y Low en las demás. Se tomó como categoría de referencia en la variable de respuesta a $\pi_c = High$. El modelo se ve como sigue:

$$\log(\frac{\mathbb{P}[Sat \leq j]}{1-\mathbb{P}[Sat \leq j]}) = \beta_0^j + \beta_1 \text{Atrium} + \beta_2 \text{Terrace} + \beta_3 \text{Tower} + \beta_4 \text{Infl:Medium} + \beta_5 \text{Infl:High} + \beta_6 \text{Cont:High} \quad j = 1, 2$$

En la Tabla 3 se tienen las estimaciones de los $\beta_i$ exponenciados, que indican cómo cambia la probabilidad de pertenecer a una categoría igual o más alta en relación con las categorías inferiores, dado un cambio de categoría en alguna covariable explicativa y manteniendo constantes las otras variables en el modelo. Por ejemplo, los que viven en Terrace presentan 67 % más ventaja comparativa de nivel de satisfacción inferior frente a una categoría de nivel superior que los que viven en Aparment, caso contrario a cuando se vive en Tower pues ahí tiene 44 % menos ventaja comparativa.

Para continuar con nuestro análisis consideramos unicamente los habitantes que viven en Tower y tienen poco contacto con otros inquilinos, con ayuda de predict calculamos las probabilidades a cada cruce y realizamos un diagrama de barras (Figura 2) de dónde podemos concluir que dada una vivienda de tipo Tower y bajo contacto con otros inquilinos:
La probailidad de ser un cliente muy satisfecho crece cuanta más influencia en las decisiones del mantenimiento tenemos, cosa contraria cuando somos clientes poco o medianamente satisfechos.
Cuando la influencia es muy alta es muy probable ser un cliente Muy satisfecho pero si la influencia es baja no hay mucha diferencia pues las probas son muy cercanas, una influencia media unicamente parece distinguir unicamente entre satisfecho o no sastifecho.

Concluimos el análisis dejando en la Gráfica 3 la probabilidad de tener cierto nivel de satisfacción de a cuerdo a la categoría donde nos encontramos.

Cuadro 3: Exponenciación de parámetros beta

| PARAMETRO | ESTIMACIÓN |
|---|---|
| Intercepto 1 | 1.0791948 |
| Intercepto 2 | 3.5362210 |
| Beta 1 | 0.8137002 |
| Beta 2 | 1.6797833 |
| Beta 3 | 0.5641981 |
| Beta 4 | 0.6974778 |
| Beta 5 | 0.5675686 |
| Beta 6 | 0.2755962 |

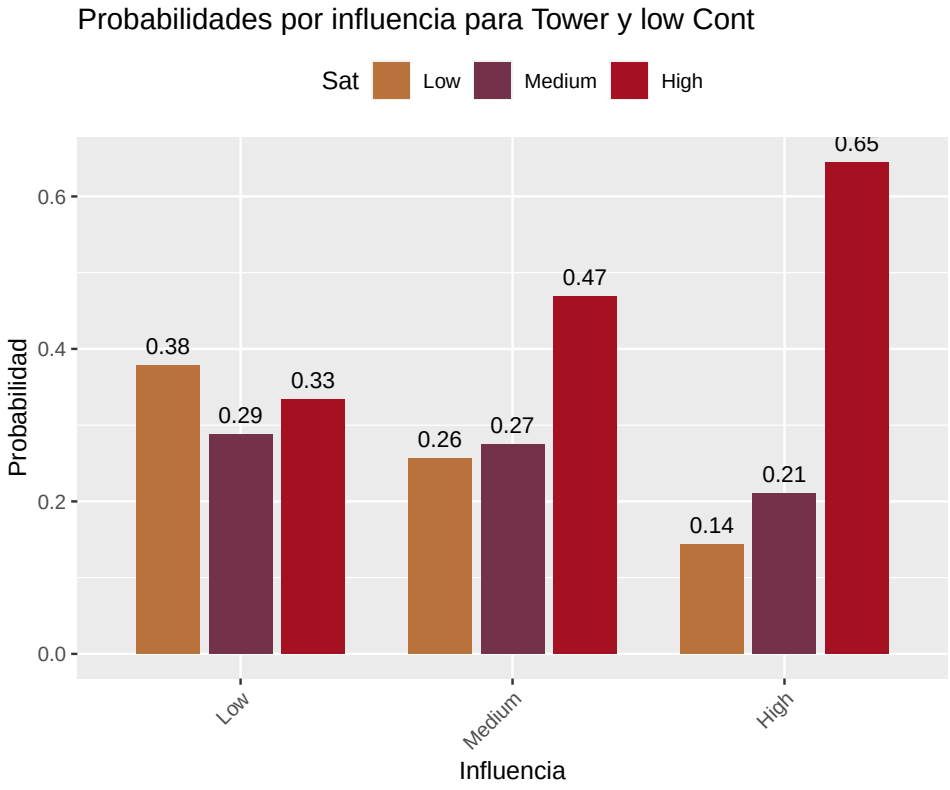



Figura 2: Probabilidades para la vivienda Tower

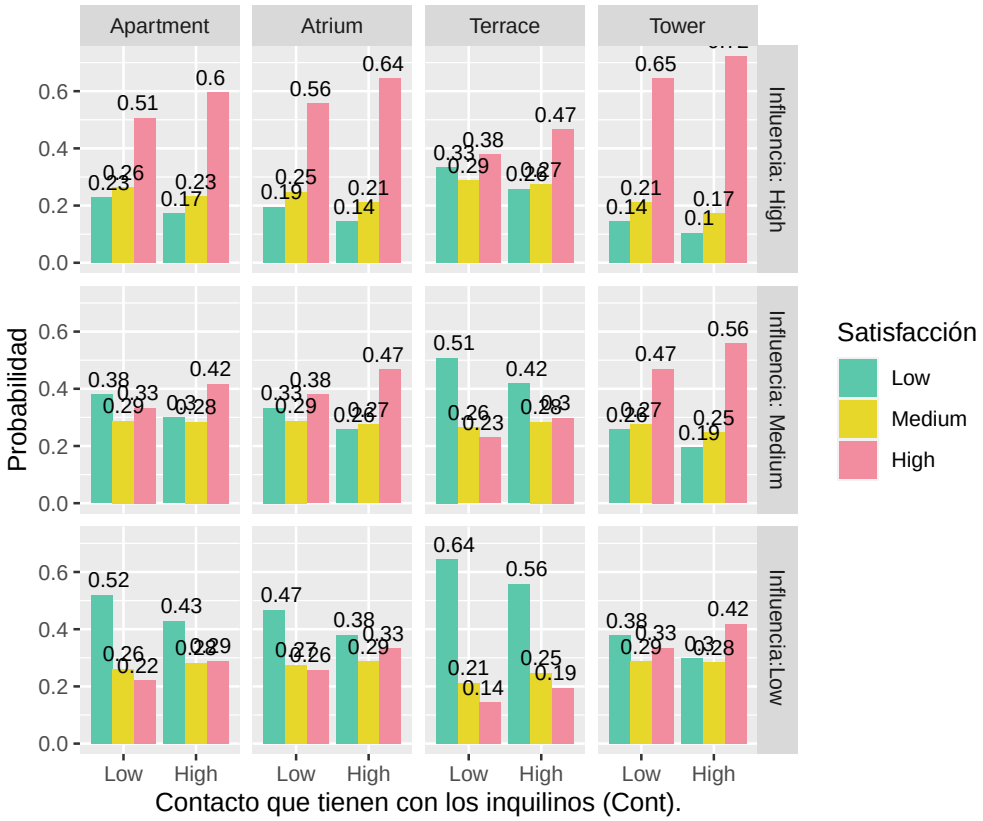



Figura 3: Probabilidades por nivel de satistacción