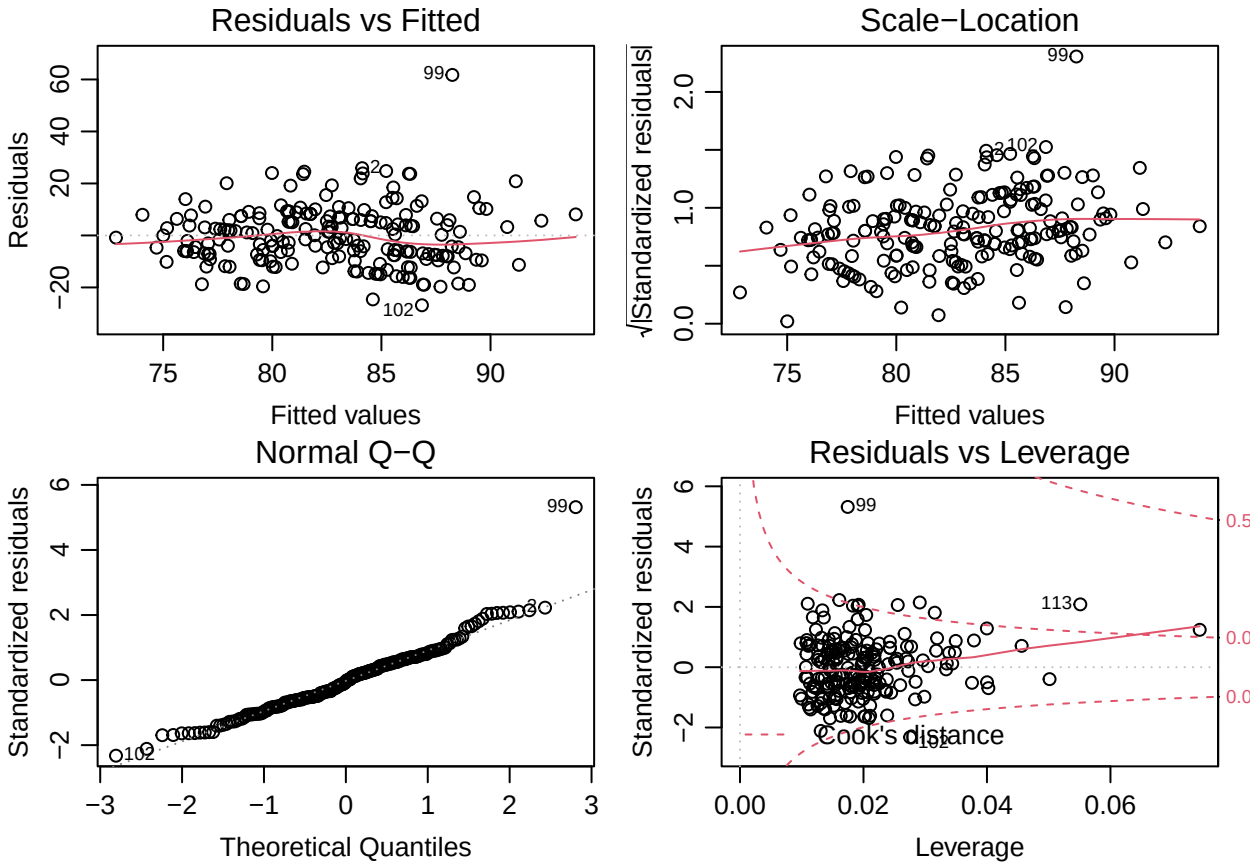
Tarea 1

Ejercicio 1

Inicialmente consideramos un modelo de regresión en donde tomábamos en consideración todas las variables (tcresult, age y sex) sin interacción. Sin embargo, a pesar de que este modelo inicial tenía sentido (pasó la prueba F de la tabla ANOVA), no paso ninguna prueba de los supuestos.

A continuación mostramos el p-value de la prueba F, gráficas mostrando el mal comportamiento del modelo respecto a los supuestos y una tabla donde mostramos el test realizado al modelo y el p-value obtenido.

```
##
## Call:
## lm(formula = bpdiast ~ tcresult + age + sex, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -26.857  -7.570  -0.532   7.041  61.750
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 67.26801    4.14280  16.237  < 2e-16 ***
## tcresult     0.05462    0.01935   2.822  0.00526 **
## age          0.12051    0.05345   2.255  0.02526 *
## sexMujer    -3.61209    1.66129  -2.174  0.03088 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.72 on 196 degrees of freedom
## Multiple R-squared:  0.1142, Adjusted R-squared:  0.1007
## F-statistic: 8.427 on 3 and 196 DF,  p-value: 2.699e-05
```




```
##          Test realizado: p-value obtenido:
## 1           car::ncvTest           0.00012
## 2           shapiro.test             1e-05
## 3 normtest::jb.norm.test                 0
## 4     car::residualPlots              <NA>
## 5                tcresult           0.02748
## 6                     age           0.17812
## 7                     sex              <NA>
## 8             Tukey test           0.57976
```

Dadas únicamente las gráficas, podríamos querer decir que no hay ningún problema con los supuestos ya que no hay grandes anomalías ni en la varianza, ni en la linealidad, ni en la normalidad. Salvo por los outliers, y si observamos las pruebas de hipótesis relacionadas a los supuesto vemos que el modelo tiene un muy mal comportamiento ya que no pasa ninguna prueba. Es decir, estos outliers nos están causando grandes problemas.

Mejorando el Modelo

Creamos un nuevo modelo en donde inicialmente sólo transformamos a la variable de interés, a bpdiast, ya que el modelo anterior no cumplía con el supuesto de homocedasticidad. Para esto usamos una transformación Box-Cox y el λ arrojado fue $\frac{-1}{2}$.

Este modelo con $bpdiast^{-1/2}$, además de tener sentido (pasó la prueba F de la tabla ANOVA) tuvo muchísimo mejores comportamientos como lo podemos ver a continuación en su respectiva tabla de pruebas de supuestos:

```
##          Test realizado: p-value obtenido:
## 1           car::ncvTest           0.06832
## 2           shapiro.test           0.32017
```

```
## 3 normtest::jb.norm.test             0.854
## 4     car::residualPlots              <NA>
## 5                tcresult           0.02917
## 6                     age           0.20716
## 7                     sex              <NA>
## 8             Tukey test           0.22832
```

Sin embargo aún hay un pequeño detalle con la linealidad, y es que el test `residualPlots` nos indica que podríamos agregar a la variable $tcresult^2$.

Este detalle nos llevó a también transformar tcresult. Para esto hicimos una transformación Box-Tidwell y el λ sugerido fue de 9.

Nuestro modelo definitivo fue este último indicado, con $bpdiast^{-1/2}$ y $tcresult^9$. A continuación podemos ver su tabla de pruebas de supuestos, donde apreciamos su buen comportamiento respecto a los supuestos:

```
##          Test realizado: p-value obtenido:
## 1           car::ncvTest           0.42704
## 2           shapiro.test            0.3422
## 3 normtest::jb.norm.test             0.555
## 4     car::residualPlots              <NA>
## 5          I(tcresult^9)           0.95108
## 6                    age           0.12593
## 7                    sex              <NA>
## 8            Tukey test            0.1861
```

Preguntas del Investigador

Para responder a la pregunta

"¿Se puede indicar que para una persona de cierta edad y sexo, tener un nivel de colesterol alto se asocia con una alta presión arterial diastólica?"

debemos traducirla a nuesto modelo. Y dado a que nuestro modelo se ve de la manera

$E[bpdiast^{-1/2}] = \beta_0 + \beta_1 tcresult^2 + \beta_2 age + \beta_3 sex$

y además la función $f(x) = x^{-1/2}$ es decreciente para valores positivos como lo es tcresult, contrastamos la hipótesis

$E[bpdiast^{-1/2}|\ tcresult + 1, age^*, sex^*] < E[bpdiast^{-1/2}|\ tcresult, age^*, sex^*]$
$\Leftrightarrow$
$\beta_0 + \beta_1(tcresult + 1)^9 + \beta_2 age^* + \beta_3 sex^* < \beta_0 + \beta_1 tcresult^9 + \beta_2 age^* + \beta_3 sex^*$
$\Leftrightarrow$
$\beta_1((tcresult + 1)^9 - tcresult^9) < 0$
$\Leftrightarrow$
$\beta_1 < 0$, pues $tcresult \geq 0$
$\therefore H_0 : \beta_1 \geq 0$ *vs.* $H_a : \beta_1 < 0$
ya que un incremento en $tcresult$ sería un decremento en $bpdiast^{-1/2}$ por la transformación Box-Cox usada.

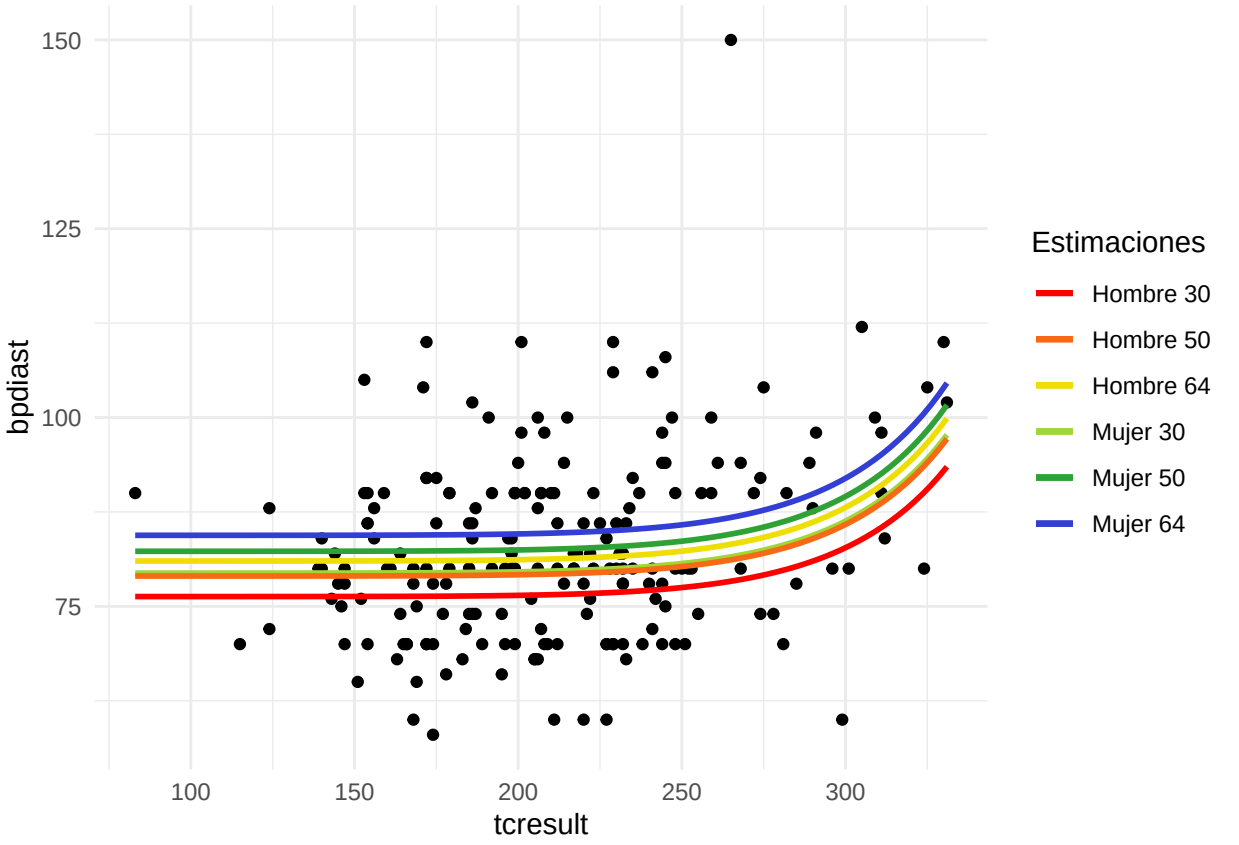
Contrastamos la hipótesis necesaria para responder al investigador y el resultado obtenido fue que encontramos fuerte evidencia de que se cumple lo preguntado, que a mayor nivel de colesterol mayor presión arterial diastólica para un paciente con edad y sexo fijo.

El p-value de la prueba fue de 0.0005 como se puede ver a continuación, evidencia muy fuerte a favor de la pregunta del investigador.

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
```

```
## Fit: lm(formula = I(bpdiast^(-1/2)) ~ I(tcresult^9) + age + sex, data = data)
##
## Linear Hypotheses:
##          Estimate Std. Error t value   Pr(<t)
## 1 >= 0 -2.318e-25  6.912e-26  -3.354 0.000479 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Graficando el Modelo



La grafica presentada nos muestra 6 curvas ajustadas según nuestro modelo, dejando diferentes edades y sexos fijos y muestra cómo cambia la presión arterial diastólica en los pacientes conforme los niveles de colesterol van aumentando para pacientes de dicha edad y dicho sexo indicado.

Podemos observar que aproximadamente con un puntaje del tcresult mayor a 275 el cambio en la presión arterial empieza a crecer exponencialmente. Antes de esta cota no hay prácticamente influencia en las variables, casi es sólo el parámetro del intercepto (haciendo las transformaciones correspondientes es aproximadamente $\beta_0^{-2} \approx 69$).

Una última observación es que este efecto de mayor presión arterial cuando se tienen altos niveles de colesterol afecta en general más a las mujeres, ya que las líneas azul y verde oscuro están por encima de todas las demás, siendo las correspondientes a mujeres con 64 y 50 años de edad respectivamente, mientras que las mujeres de 30 años sufren de este efecto prácticamente con la misma intensidad que los hombres de 50 años.

Ejercicio 2

El modelo que presentamos es uno gaussiano inverso usando la liga identidad y tomando a la variable tcresult pero elevada al cubo.

Dado a que la liga es la identidad, la expresión matemática para modelar la esperanza de los valores de presión arterial esuna muy sencilla de interpretar (sólo debemos asumir que las observaciones provienen de una distribución inversa gaussiana, como su nombre lo indica). Dicha expresión es la siguiente:

$$E[bpdiast|tcresult, age, sex] = \beta_0 + \beta_1 tcresult^3 + \beta_2 age + \beta_3 sex$$

Donde sex=1 cuando el paciente es mujer, 0 en el caso de los hombres.

El procedimiento para obtener este modelo fue tomar una malla de polinomios y potencias para tcresult, pues en el Ejercicio 1 habíamos visto que esta variable era la que causaba problemas en la linealidad de los supuestos (además de que nos interesa modelar el cambio de la presión arterial respecto al cambio en los niveles de colesterol). Después de tomar esta malla tomamos todas las posibles combinaciones entre las distribuciones Gaussiana, Gamma, Inversa Gaussiana, y las diferentes ligas útiles para estas distribuciones: identidad, logarítmica, inversa e inversa cuadrática.

Después ordenamos toda esta malla de modelos por su AIC y BIC. A continuación les mostramos algunos de los mejores modelos calificados por su AIC y BIC:

```
##      Index     AICs Index     BICs
## [1,]   157 1533.467   157 1549.959
## [2,]   158 1533.869   158 1550.361
## [3,]    17 1533.932   147 1550.534
## [4,]   147 1534.042   159 1550.765
## [5,]   159 1534.274   148 1550.920
## [6,]    18 1534.408   160 1551.171
```

```
## [1] "Modelo con Index=157:"
```

```
##
## Family: inverse.gaussian
## Link function: identity
```

```
## [1] "bpdiast ~ I(tcresult^(3)) + age + sex"
```

```
## [1] "Modelo con Index=158:"
```

```
##
## Family: inverse.gaussian
## Link function: log
```

```
## [1] "bpdiast ~ I(tcresult^(3)) + age + sex"
```

```
## [1] "Modelo con Index=147:"
```

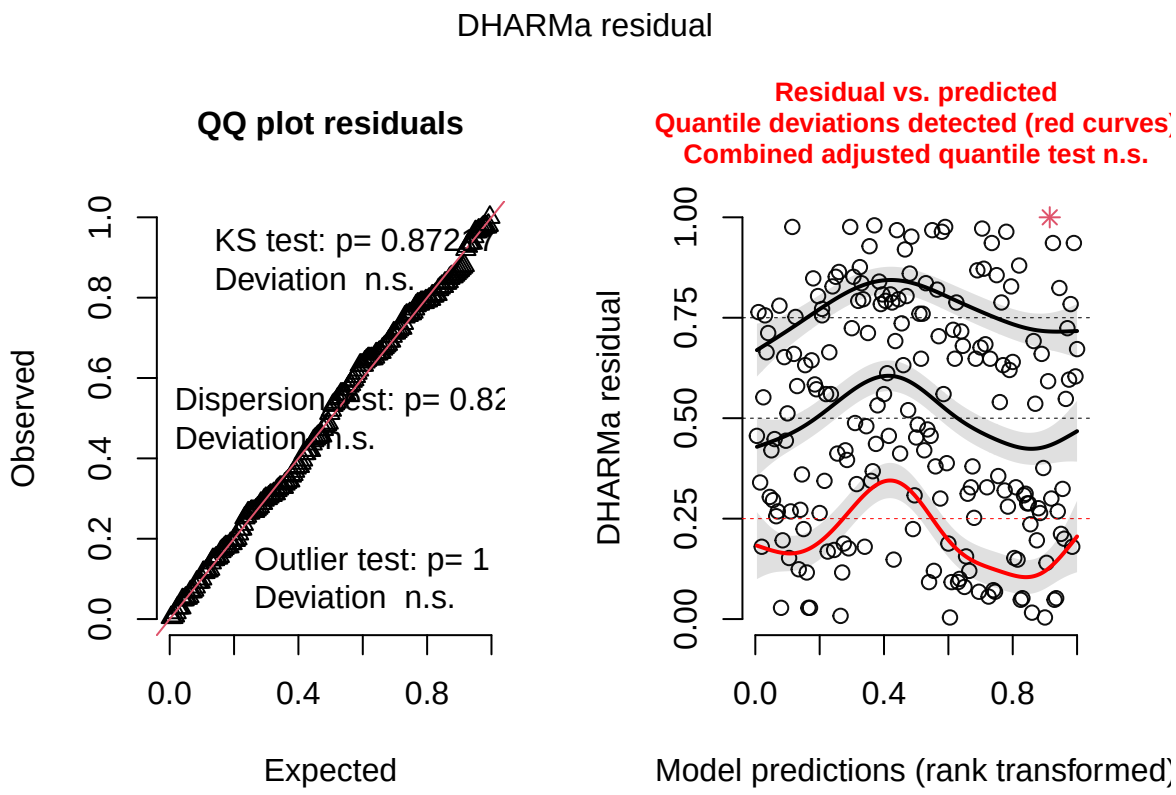
```
##
## Family: inverse.gaussian
## Link function: identity
```

```
## [1] "bpdiast ~ I(tcresult^(2.5)) + age + sex"
```

También notamos que al seguir aumentando la potencia a la que elevamos tcresult nos mejora un poco el AIC, pero para no perder interpretación optamos por dejar la potencia igual a 3.

A este modelo le hicimos un análisis de supuestos. No presentó prácticamente ningún problema por lo que lo tomamos como nuestra mejor opción.

Mostramos la gráfica de residuales simulados como resumen de que cumple con los supuestos:



*Toleramos

Preguntas del Investigador

Ahora queremos saber si a mayores niveles de colesterol mayores niveles de presión arterial diastólica. Como nuestro modelo es sencillo de interpretar, la prueba es más directa. En ecuaciones se ve así:

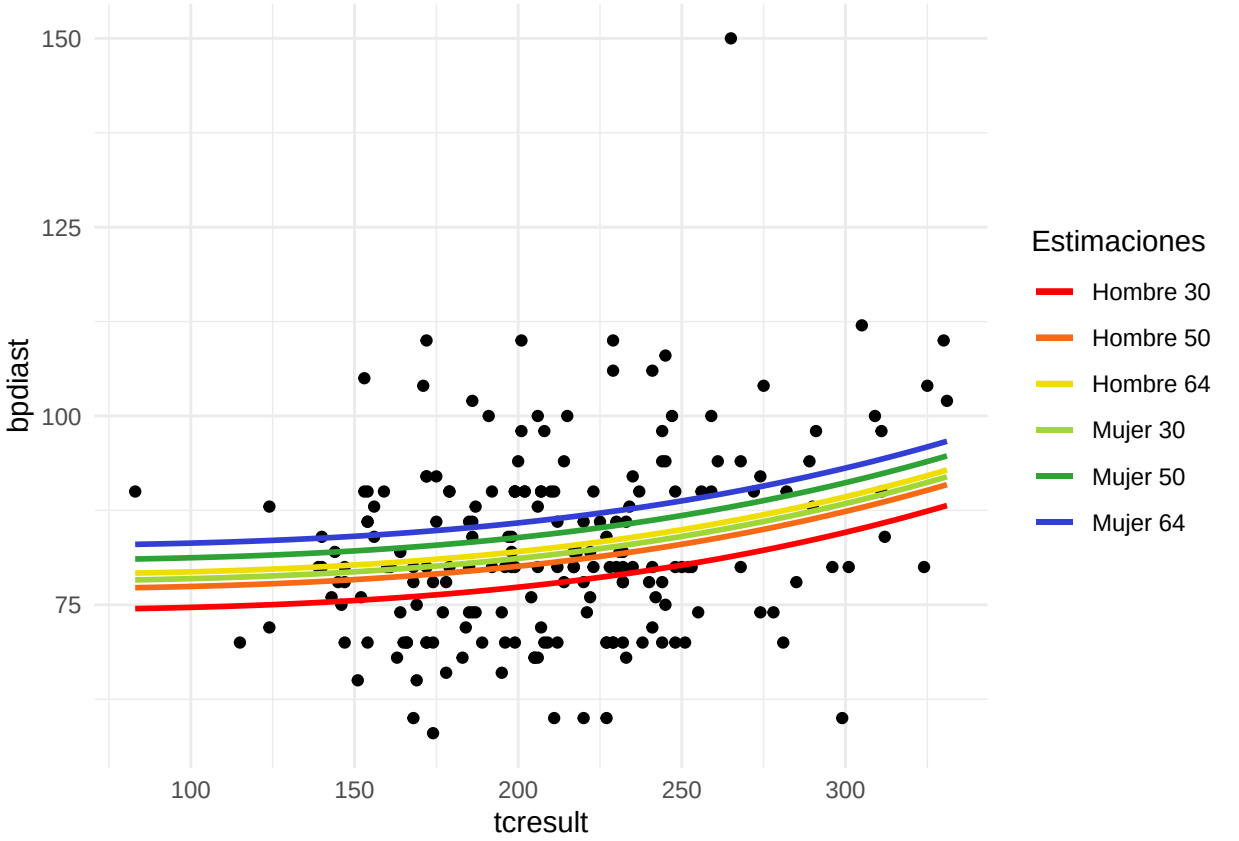
$E[bpdiast|tcresult+1, age^*, sex^*] > E[bpdiast|tcresult, age^*, sex^*]$
$\Leftrightarrow$
$\beta_0 + \beta_1(tcresult+1)^3 + \beta_2 age + \beta_3 sex > \beta_0 + \beta_1 tcresult^3 + \beta_2 age + \beta_3 sex$
$\Leftrightarrow$
$\beta_1((tcresult+1)^3 - tcresult^3) > 0$
$\Leftrightarrow$
$\beta_1 > 0$, pues $tcresult \geq 0$ Por lo tanto contrastamos $H_0 : \beta_1 \leq 0$

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: glm(formula = bpdiast ~ I(tcresult^3) + age + sex, family = inverse.gaussian(link = "identity"),
##     data = data)
##
## Linear Hypotheses:
##         Estimate Std. Error z value  Pr(>z)
## 1 <= 0 3.823e-07  1.274e-07       3 0.00135 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

Como el p-value = 0.001, rechazamos H_0 y concluimos lo que el investigador sospechaba, que a mayor colesterol mayor presión arterial.

Graficando el Modelo

Haremos las mismas gráficas que en el Ejercicio 1. Para tres edades específicas y sus combinaciones con los dos sexos.



Las interpretaciones de esta gráfica son prácticamente a las de el Ejercicio 1, pues la principal diferencia en estas es que la del modelo de regresión lineal múltiple tiene un crecimiento más exponencial cuando estamos en valores de colesterol mayores a 275, y ese modelo con la suposición de que las observaciones provienen de una inversa gaussiana, es más lineal el crecimiento. Se penaliza menos el tener altos niveles de colesterol.

Elección de un Modelo Definitivo

Optamos por elegir un modelo según lo que se necesite explicar Por ejemplo, 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 colesterol es más acertado que el de la suposición de que las observaciones provienen de una inversa gaussiana. El problema con éste, es que la interpretación es muy complicada por las transformaciones hechas a la variable bpdiast.

Mientras que para fines didácticos o explicativos con valores más exactos de cómo cambia la presión arterial diastólica con respecto a los niveles de colesterol en sangre, nos quedamos con el modelo lineal generalizado, pues al ser una interpretación directa sobre la variable bpdiast, esto nos facilita el entender cómo afecta el cambiar una variable respecto a la otra.

Además, por último, contrastamos los AIC de estos dos modelos, siendo nuevamente el de la regresión lineal múltiple el mejor calificado (se tuvo que hacer un ajuste a su AIC por la transformación que se hizo a bpdiast).

Mostramos la transformación de AIC y los AIC correspondientes a los dos modelos a continuación:

$$L(\theta|\underline{y}) = \prod_{i=1}^{200} f(y_i) =$$

$$\Pi_{i=1}^{200} f_Z(z_i) \prod_{i=1}^{200} \frac{d}{dy} z_i =$$

$$L(\theta|\underline{z}) \prod_{i=1}^{200} \frac{-1}{2y^{3/2}} =$$

$$L(\theta|\underline{z}) \prod_{i=1}^{200} \frac{1}{2y^{3/2}}$$

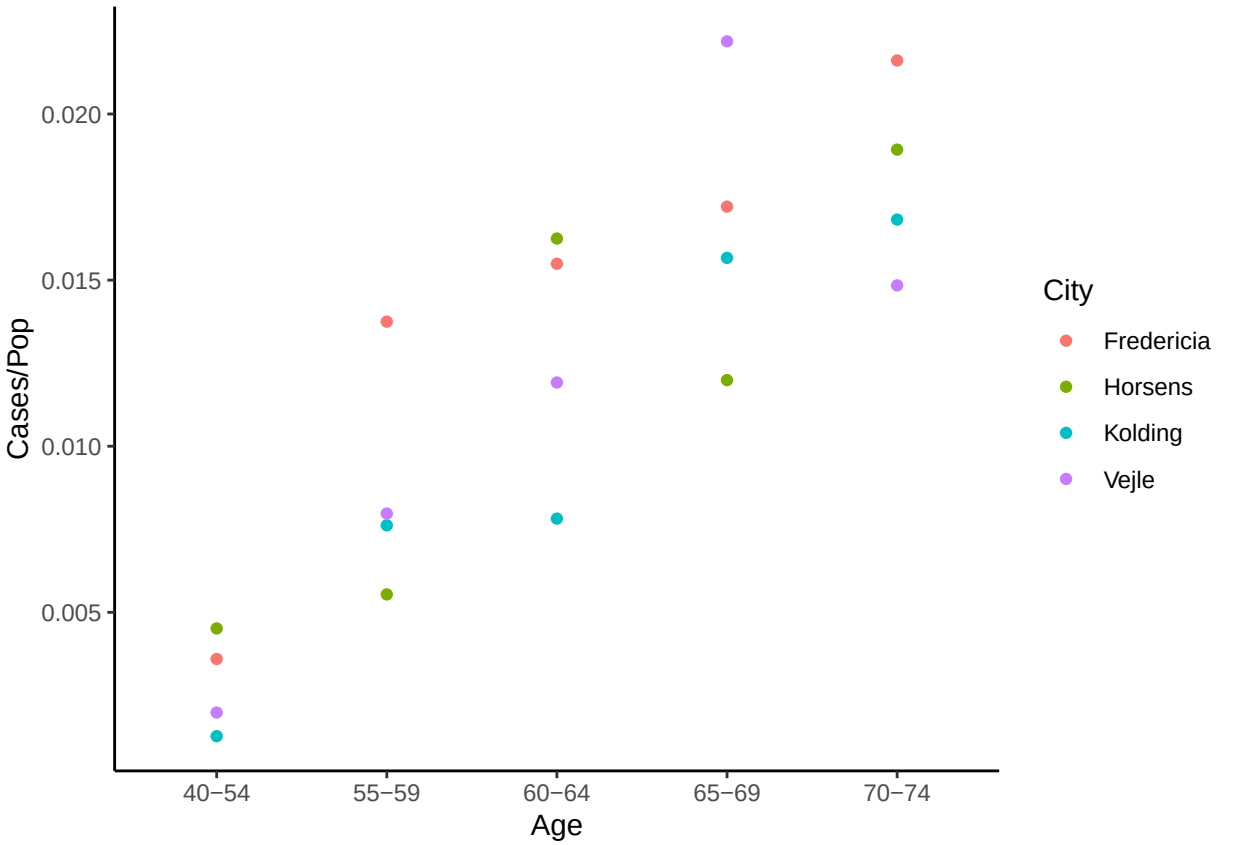
Por lo tanto, llevándolo a forma de AIC

$$AIC = AIC_{(y^{-1/2})} + 2\sum_{i=1}^{200} ln(2y^{3/2})$$

```
## [1] "AICes"
## [1] "RLM:"            "1530.70041163605"
## [1] "InversaG:"       "1533.46705379946"
```

Ejercicio 4

Analicemos los datos.



Podemos observar una fuerte tendencia en los datos respecto a la tasa de incidencia de los casos de cáncer de pulmón contra la edad, hay un claro incremento conforme avanzamos en los grupos de edad.

También podríamos argumentar que la distinción por ciudades no hace un gran cambio a simple vista. Probablemente la inclusión de esta variable cause ruido en nuestro modelo, pero no podemos descartar que afecte la ciudad para la tasa hasta hacer una prueba formal.

Modelos con Poisson

Para descartar este hecho, hemos hecho dos modelos y los contrastamos. En el primero consideraremos todas las interacciones entre la edad y la ciudad, y en el segundo únicamente a los grupos de edad.

En ambos casos los modelos fueron basados en la distribución Poisson con liga logarítmica.

Nuestro primer modelo con todas las interacciones entre ciudad y grupos de edad no cumple siquiera con los supuestos, mientras que el segundo únicamente considerando los grupos de edad sí los cumple.

Además de esto, el primer modelo generó una barbaridad de coeficientes. Deseamos poder considerar al segundo modelo en su lugar. En este caso usamos criterios como el AIC, BIC y una prueba de bondad de ajuste ya que estos modelos están anidados y queremos tomar al segundo modelo, que es una versión reducida del primero.

Primero los AIC y BIC:

```
c(AIC(fit.poi1), AIC(fit.poi2))
```

```
## [1] 121.4730 108.4512
```

```
c(BIC(fit.poi1), BIC(fit.poi2))
```

```
## [1] 141.3876 113.4299
```

Tenemos menores AIC y BIC para el segundo modelo.

Ahora la prueba de bondad de ajuste:

```
## Analysis of Deviance Table
##
## Model 1: Cases ~ Age * City + offset(logPop)
## Model 2: Cases ~ Age + offset(logPop)
##   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 de 0.32, no tenemos evidencia en contra de que debemos tomar el modelo completo ante el reducido. Es decir, podemos quedarnos con el reducido.

Y esto hace mucho sentido ya que desde un inicio, en la gráfica pudimos observar que independientemente de la ciudad, las tasas de incidencia parecían ser muy iguales entre todas.

Modelo Binomial Negativo

Ahora que hemos seleccionado un buen modelo Poisson con liga logarítmica, también probaremos un ajuste con la distribución Binomial Negativa (liga logarítmica de igual manera).

Este nuevo modelo con la distribución Binomial Negativa también cumple con los supuestos.

Por lo que podemos comparar los AIC y BIC de este modelo con el Poisson anterior y decidirnos por alguno:

```
c(AIC(fit.nb), AIC(fit.poi2))
```

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

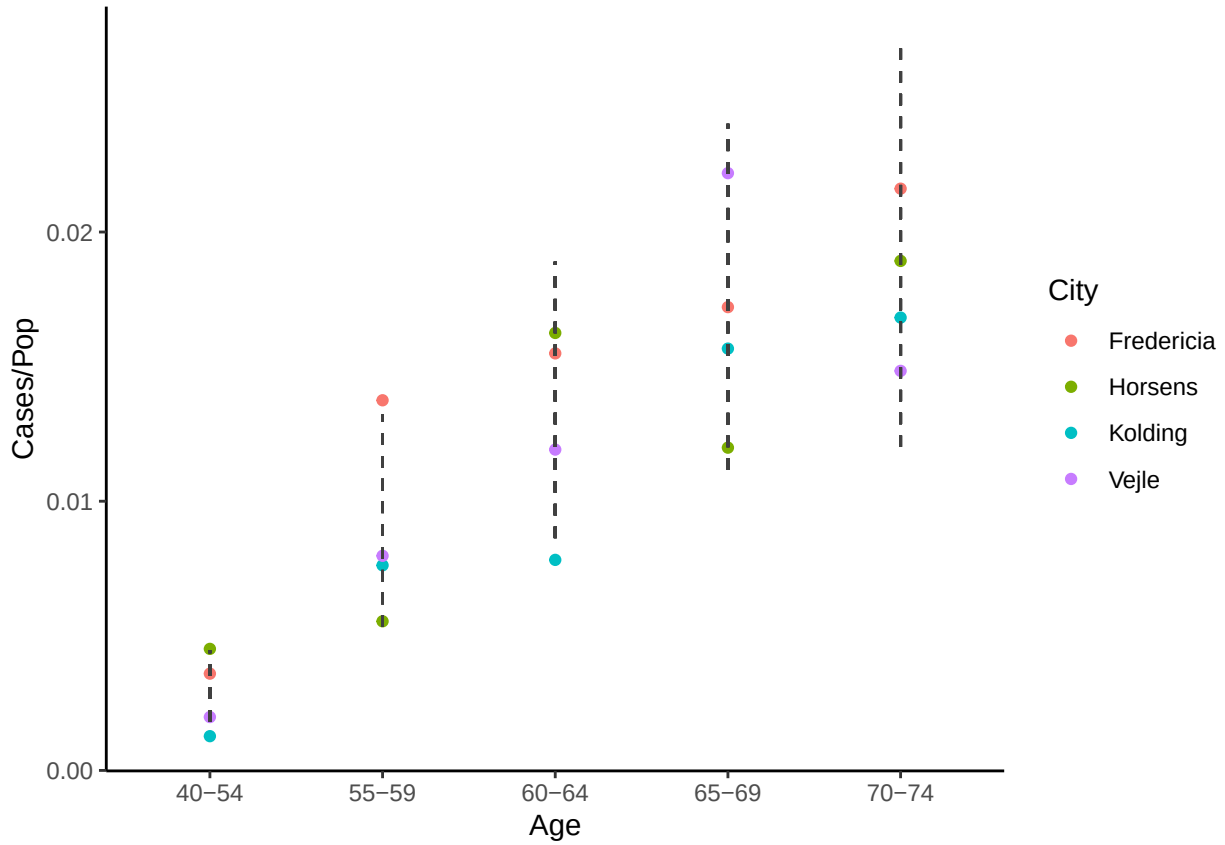
```
c(BIC(fit.nb), BIC(fit.poi2))
```

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

El modelo Poisson tiene menores índices, por lo que también lo seleccionamos por encima del binomial negativo.

Intervalos de Confianza: Grupos de Edad

Mostraremos intervalos de confianza al 95% para los diferentes grupos de edades:



A ojo gracias a estos intervalos de confianza podemos suponer que en efecto, a mayor edad mayores tasas de incidencias. Pero haremos una prueba para verificar esto.

Contrastamos los dos primeros grupos de edad, 40-54 y 55-59, contra los dos últimos, 65-69 y 70-74, ya que a ojo es evidente que a gran escala sí podemos concluir que hay un aumento en la tasa de incidencia de casos de cáncer de pulmón.

Serán pruebas simultáneas siguiendo la lógica

$H_0 : \mu(Cases/Pop \mid AgeGroup_{70-74}) \leq \mu(Cases/Pop \mid AgeGroup_{40-54})$

Y más estrictamente contrastamos la siguiente hipótesis:

$H_0 : \beta_4 - \beta_0 \leq 0 \ \cap \ \beta_3 - \beta_0 \leq 0 \ \cap \ \beta_4 - \beta_1 \leq 0 \ \cap \beta_3 - \beta_1 \leq 0$

Es decir, las pruebas del primer grupo contra los últimos dos grupos de edad, y del segundo grupo contra los dos últimos grupos de edad.

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Fit: glm(formula = Cases ~ Age + offset(logPop), family = poisson(link = "log"),
##     data = data)
##
## Linear Hypotheses:
##        Estimate Std. Error z value  Pr(>z)
## 1 <= 0   7.7095     0.3824  20.162 < 0.001 ***
## 2 <= 0   7.6125     0.3787  20.100 < 0.001 ***
## 3 <= 0   0.7649     0.2372   3.225 0.00208 **
## 4 <= 0   0.6679     0.2312   2.889 0.00684 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

En todas las pruebas obtuvimos un p-value menor al 0.05, por lo que en efecto tenemos evidencia en contra de que al aumentar la edad se reduce el riesgo.

Es decir, que lo que suponíamos se cumple. A mayor edad mayor riesgo de padecer cáncer de pulmón.

Modelo con Edad Continua

Por último consideramos un modelo suponiendo que la edad es una variable continua, no categórica. Para eso les dimos una edad puntual a cada observación de cada grupo. Ésta fue el punto medio de su grupo de edad (por ejemplo al Grupo 40-54 se le asignó la edad de 47 años)

Creamos 4 modelos, 2 con Poisson y liga logarítmica y 2 con Binomial Negativa y la misma función liga. Para cada distribución consideramos un modelo con sólo interacciones de la edad vista como variable continua, y para su segundo modelo respectivo uno donde se tome en cuenta la edad y la edad^2 como variables continuas.

Desafortunadamente sólo 2 de los 4 modelos cumplen con los supuestos, y coinciden en ser los dos modelos donde se toma en cuenta la influencia de la edad y la edad^2 (es decir tenemos un modelo Poisson y uno Binomial Negativo).

A continuación mostramos sus AIC y BIC:

```
c(AIC(Poi), AIC(BN))
```

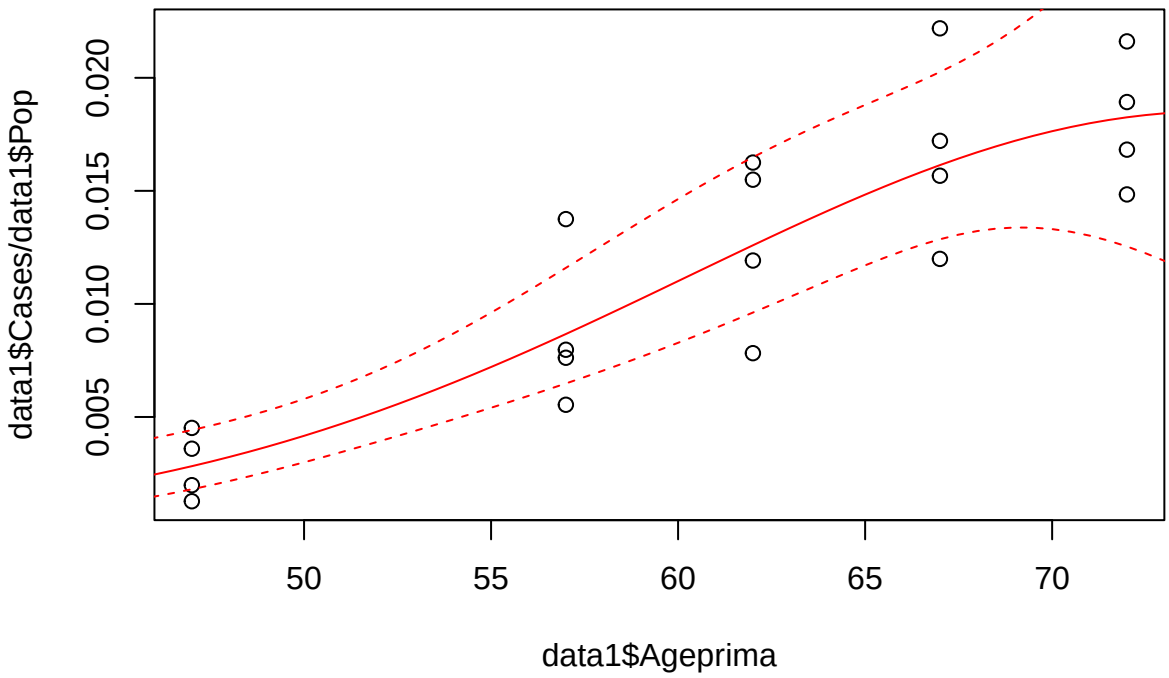
```
## [1] 104.5095 106.5098
```

```
c(BIC(Poi), BIC(BN))
```

```
## [1] 107.4967 110.4927
```

Por lo que consideraremos el Poisson.

Para finalizar este reporte, mostraremos intervalos de confianza continuos usando este último modelo tomando a la edad como variable continua para apoyar el hecho de que a mayor edad, mayor incidencia de cáncer de pulmón.



Para dar un poco más de rigurisidad veamos que esto tiene sentido, ya que nuestra variable edad sólo nos interesa en valores entre 40 y 74 años, es decir $40 \leq x \leq 74$, donde x representa la edad de algún paciente. Y si a mayor edad, mayor probabilidad de tener cáncer, eso nos implicaría que nuestra función de incidencia es creciente respecto a la edad. Lo que a su vez lo podemos traducir en que si tenemos una derivada positiva de esta función para el intervalo $[40,\ 74]$, entonces en efecto podemos concluir lo deseado.

Esto se cumple ya que, equivalentemente, la derivada de nuestra función liga se ve así:

$\frac{d}{dx}\mu(Cases/Pop \mid Age = x) = \frac{d}{dx}(\beta_0 + \beta_1 x + \beta_2 x^2) = \beta_1 + 2\beta_2 x$, y dado a que nuestros valores del modelo de β_1 y β_2 son respectivamente 0.3723 y -0.0025, la recta $\beta_1 + 2\beta_2 x$ es positiva en el intervalo $(-\infty, \frac{-\beta_1}{2\beta_2}) = (-\infty, 74.46)$.

En particular la derivada $\frac{d}{dx}\mu(Cases/Pop \mid Age = x) = \beta_1 + 2\beta_2 x$ es positiva en el intervalo $[40,\ 74]$.

Con lo que concluimos que en efecto a mayor edad, mayor indicencia en los casos de cáncer de pulmón, pues la función de tasa de incidencia es creciente en el intervalo deseado.

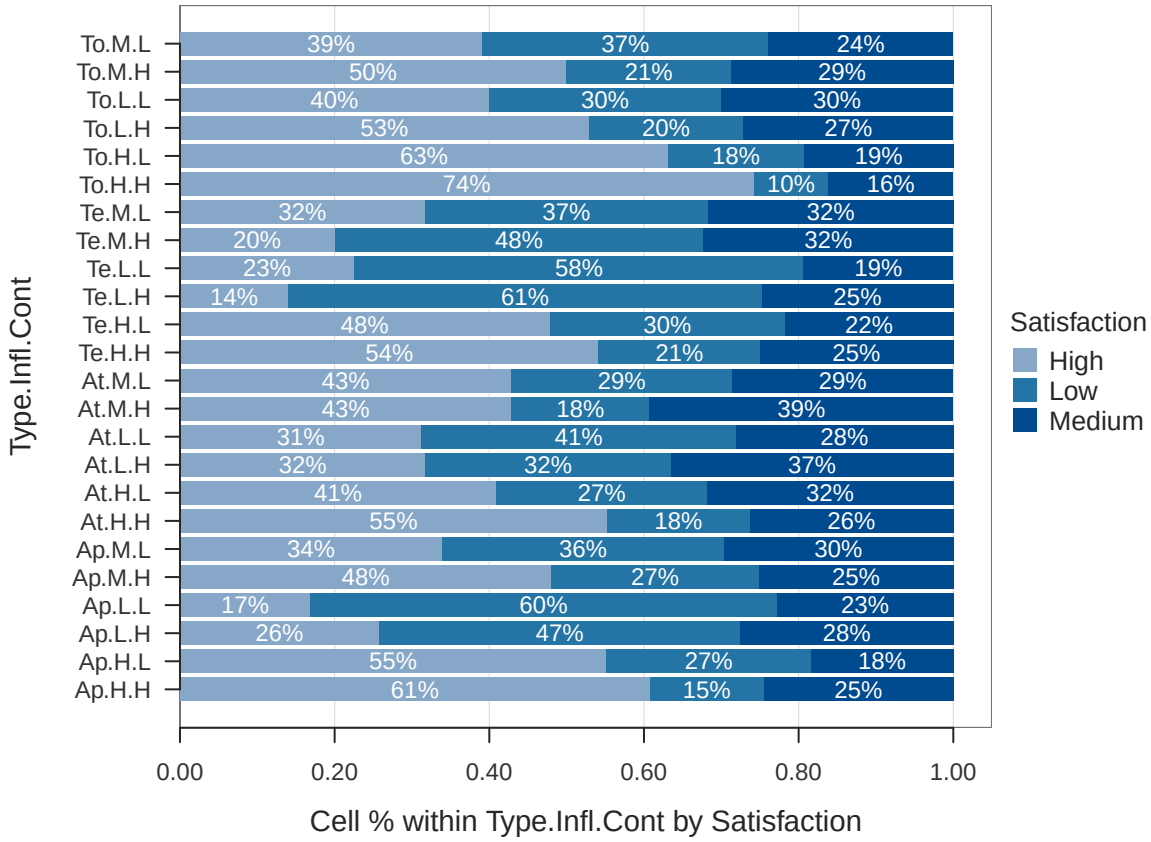
Ejercicio 5

Presentaremos la gráfica de los datos, donde la notación usada por simplicidad es la siguiente.

Para el tipo de viviendas: Tower=To, Atrium=At, Apartment=Ap, Terrace=Te

Para los niveles ordinales: Low=L, Medium=M, High=H

Además los datos son presentados en el formato "Vivienda.Influencia.Contacto". Por ejemplo la etiqueta To.M.L nos indica que hablamos de los inquilinos con vivienda tipo Tower, que perciben su influencia en el mantenimiento de la vivienda como Medium, y que su contacto con el resto de inquilinos es Low.



En general no podemos observar grandes patrones, salvo unos pocos.

Por ejemplo, aquellos inquilinos que se consideran con influencia en el mantenimiento como High y a su vez tienen buena comunicación con los demás, su nivel de satisfacción siempre es mayor al 50%, esto hace sentido ya que sentir que su opinión importa y tener buena comunicación con los demás puede influir en sentirse cómodos en su vivienda.

Por otro lado, cuando se consideran con influencia Low y comunicación Low, salvo por los inquilinos que viven en Tower, tienen un índice menor al 40% en sentirse poco satisfechos.

Otro dato a notar, es que en general los individuos que viven en Tower tienden a ser los que se sienten más satisfechos, seguidos por los que viven en Apartment.

Ajuste de Modelo Multinomial

Ajustamos dos modelos multinomiales, uno donde consideramos todas las interacciones entre todas las variables y uno sin interacciones.

De estos dos modelos nos quedamos con el que no tiene interacciones, ya que hicimos una prueba de ajuste y no encontramos evidencia para descartar al modelo reducido.

Además de que tiene menores AIC y BIC este último modelo sin interacciones. A continuación mostramos la prueba realizada así como los AIC y BIC de ambos modelos.

```
## Likelihood ratio test
##
## Model 1: Sat ~ Type * Infl * Cont
## Model 2: Sat ~ Type + Infl + Cont
##    #Df  LogLik Df  Chisq Pr(>Chisq)
```

```
## 1 3314 -1715.7
## 2 3348 -1735.0 34 38.662     0.2671
```

```
## [1] "Completo" "Reducido"
```

```
## [1] "AIC:"
```

```
## [1] 3527.422 3498.084
```

```
## [1] "BIC:"
```

```
## [1] 3787.925 3574.064
```

Por estos motivos, consideramos como un mejor modelo el reducido, ya que tenemos menos parámetros y mejores criterios AIC y BIC.

Modelos Logísticos Acumulativos

También consideramos dos modelos logísticos acumulativos. Uno con el supuesto de curvas paralelas de probabilidad acumulada y otro sin este supuesto.

De estos dos modelos optamos por el modelo con el supuesto de curvas paralelas de probabilidad acumuladad. Para tomar esta decisión hicimos una prueba de ajuste y obtuvimos un p-value mayor al 0.05, por lo que no encontramos evidencia para no descartar al modelo con más parámetros.

Encima de esto, este modelo con el supuesto de curvas paralelas tiene menores AIC y BIC. Mostramos la prueba de ajuste, así como los AIC y BIC de ambos modelos a continuación.

```
## Likelihood ratio test
##
## Model 1: Sat ~ Type + Infl + Cont
## Model 2: Sat ~ Type + Infl + Cont
##    #Df  LogLik Df  Chisq Pr(>Chisq)
## 1 3354 -1739.6
## 2 3348 -1735.3 -6 8.5706     0.1992
```

```
## [1] "No-Paralelas" "Paralelas"
```

```
## [1] "AIC:"
```

```
## [1] 3498.579 3495.149
```

```
## [1] "BIC:"
```

```
## [1] 3574.559 3538.566
```

Elección Final de Modelo

Después de tomar en consideración todos estos modelos, optamos por el modelo logístico acumulativo con el supuesto de curvas paralelas, ya que encima de ser mejor en interpretación al que no tenía este supuesto, también resultó mejor al multinomial bajo los criterios del AIC y BIC. Lo mostramos a continuación:

```
## [1] "Paralelos"      "Mult. Reducido"
```

```
## [1] "AIC:"
```

```
## [1] 3495.149 3498.084
```

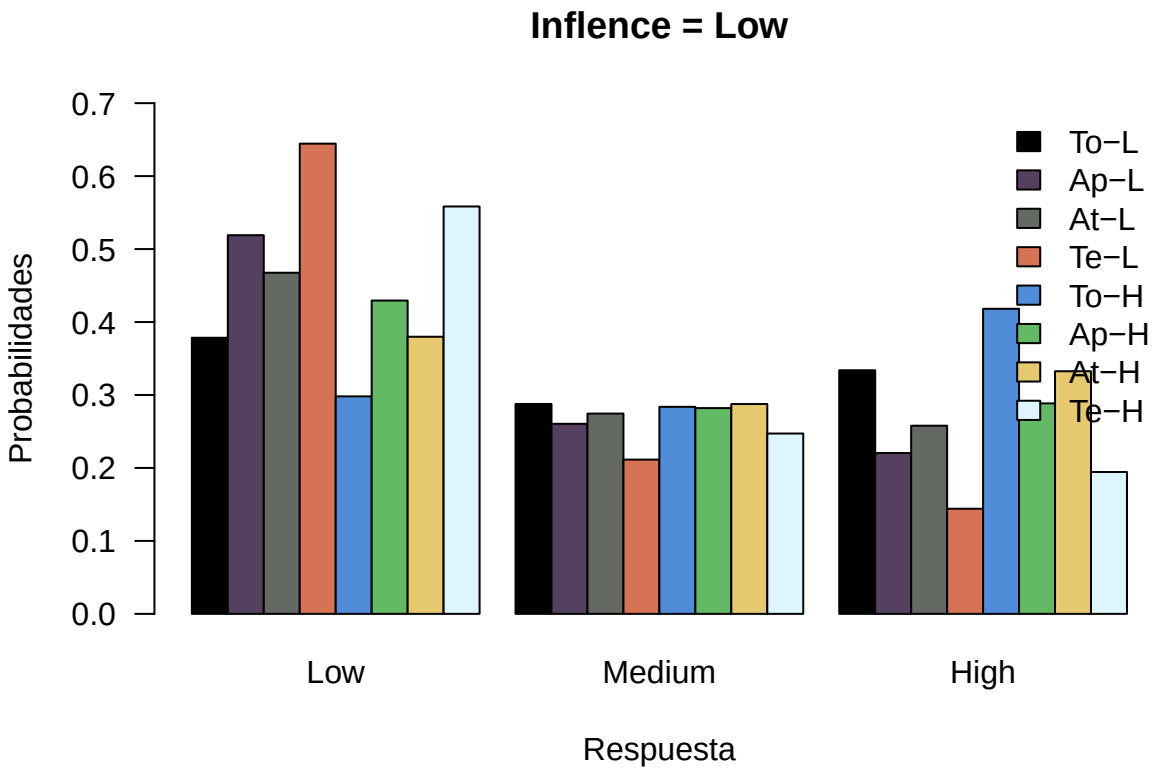
```
## [1] "BIC:"
```

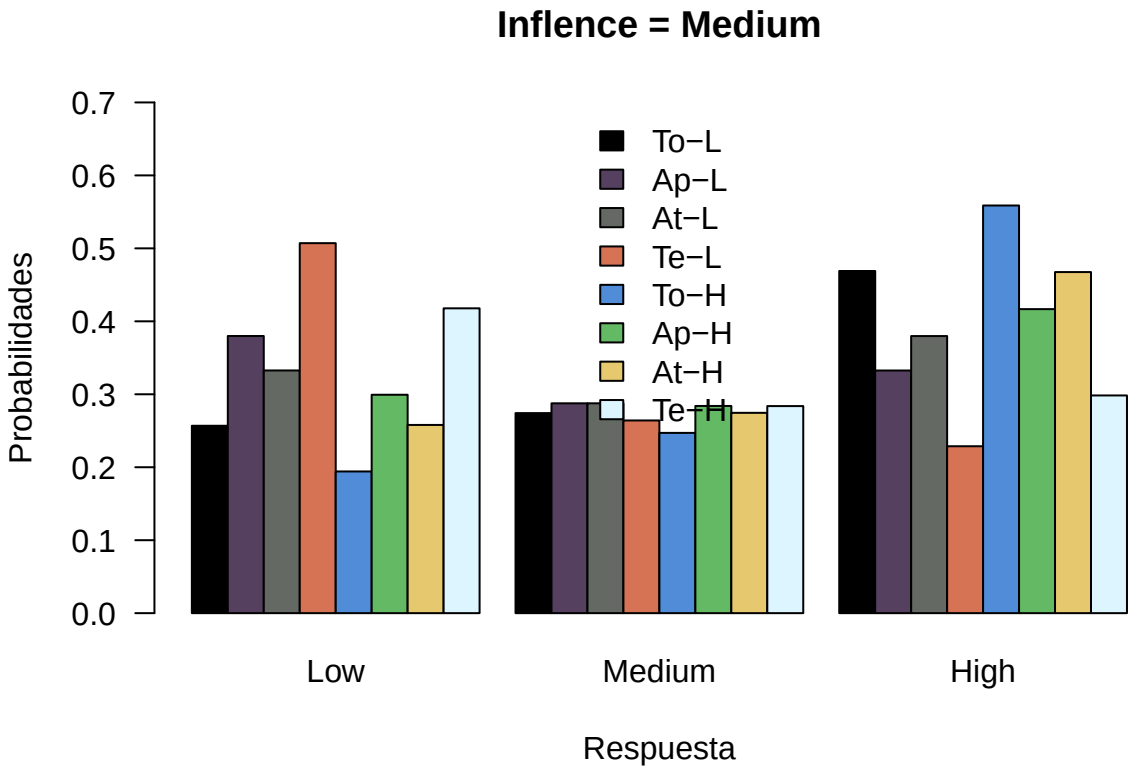
```
## [1] 3538.566 3574.064
```

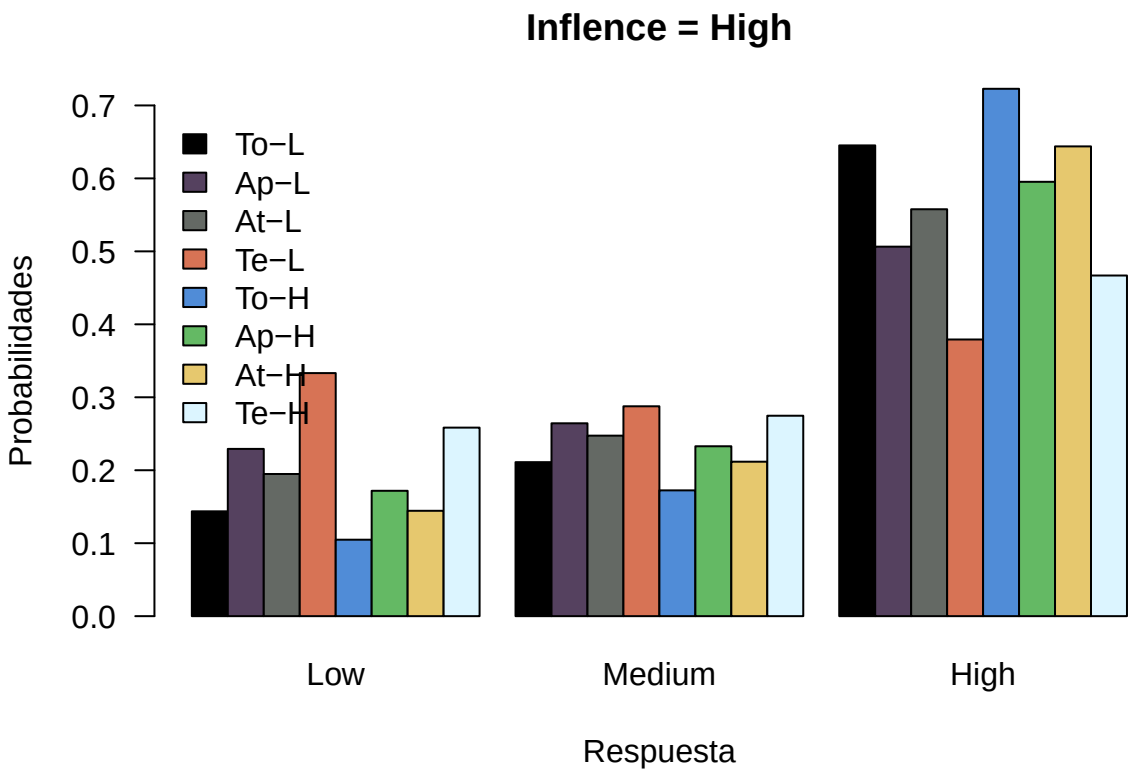
Interpretación Usando el Modelo Final

Usando este modelo final elegido, interpretamos las probabilidaes arrojadas por el modelo de caer en cada nivel de satisfacción dependiendo de cada uno de las 24 posibles combinaciones de categorías (4 del tipo de vivienda, 3 del nivel de influencia en las decisiones del mantenimiento, 2 en la comunicación con otros inquilinos. $4 \times 3 \times 2 = 24$)

Mostramos a continuación una serie de gráficas, cada una dependiendo el nivel de influencia autopercibida en las decisiones del mantenimiento de los inquilinos




Inflence = Medium
Probabilidades
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
To−L
Ap−L
At−L
Te−L
To−H
Ap−H
At−H
Te−H
Low
Medium
High
Respuesta



Por simplicidad interpretemos lo arrojado por el modelo para la variable Infl. Dejemos fijo que la vivienda del inquilino sea Tower y que el nivel de contacto con los demás inquilinos sea Low. Es decir, fijémonos en las barras de color negro de las tres gráficas.

Para el caso en el que autoperciban su influencia en las decisiones del mantenimiento como Low, las probablidades (aproximadas) arrojadas por el modelo son las siguientes:
Sentirse Low satisfecho, 38%
Sentirse Medium satisfecho, 29%
Sentirse High satisfecho, 33%

Para el caso cuando autoperciben su nivel de influencia como Medium, las probabilidades aproximadas son:
Sentirse Low satisfecho, 25%
Sentirse Medium satisfecho, 27%
Sentirse High satisfecho, 48%

Para el caso cuando autoperciben su nivel de influencia como High, las probabilidades aproximadas son:
Sentirse Low satisfecho, 15%
Sentirse Medium satisfecho, 20%
Sentirse High satisfecho, 65%

Podemos repetir esta interpretación de las gráficas para cada combinación del tipo “Type.Infl.Cont”.