



**Facultad de  
Ciencias**  
UNAM

SEMINARIO DE ESTADÍSTICA

---

## Tarea 1B

### INTRODUCCIÓN A LOS MODELOS LINEALES GENERALIZADOS

---

Enríquez Hernández Leobardo  
Tlahuiz Tenorio Giovanni Saúl

11 de abril de 2024

# Índice

1. Regresión lineal múltiple. . . . .	2
i) Modelo de RLM reducido para $E(\text{bpsystol}; \text{bmi}, \text{sex}, \text{age})$ con datos originales. . . . .	2
ii) Modelo adecuado con transformación de datos. . . . .	2
iii) Asociación entre masa corporal y presión arterial sistólica. . . . .	6
iv) Gráfica resumen con la estimación puntual de la relación $\text{bpsystol}$ y $\text{bmi}$ . . . . .	6
2. Modelos lineales generalizados para datos continuos . . . . .	8
i) Explorando modelos con variable dependiente continua. . . . .	8
ii) Asociación entre masa corporal y presión arterial sistólica, y estimación puntual. . . . .	13
iii) Comparativo modelo de regresión lineal múltiple contra modelo lineal generalizado. . . . .	14
3. Modelos lineales generalizados para datos binarios . . . . .	16
i) Gráfica de dispersión de dosis del insecticida y la proporción de insectos muertos. . . . .	16
ii) Ajuste modelos para datos binarios 1 . . . . .	16
iii) Ajuste modelos para datos binarios 2 . . . . .	19
iv) Modelo adecuado. Comparaciones, probabilidades y prueba de hipótesis. . . . .	22
4. Modelos lineales generalizados para datos de conteos . . . . .	25
i) Gráfica de dispersión de grupos de edad e incidencia . . . . .	25
ii) Distribución Poisson con liga logarítmica y un segundo modelo. . . . .	25
iii) Modelo binomial negativo, comparación e intervalo de confianza simultáneo. . . . .	29
5. Modelos lineales generalizados para datos categóricos . . . . .	33
i) Gráfica de frecuencias relativas . . . . .	33
ii) Modelo logístico multinomial nominal . . . . .	34
iii) Modelo logístico acumulativo (cumulative logit) ordinal . . . . .	35
iv) Selección del modelo e interpretación de resultados . . . . .	38

## 1. Regresión lineal múltiple.

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

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

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

```
datos<-read_csv("Preg1B.csv", show_col_types = FALSE)
datos$sex<-factor(datos$sex) #declaramos la variable como factor
#Realizamos un releve para poner como referencia "1" hombre
datos$sex<-relevel(datos$sex,"1")
```

```
modelo1<-lm(data=datos, bpsystol ~ bmi + sex + age)
#summary(modelo1)
```

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

```
table_tests<-nice_assumptions(modelo1) #usamos esta función de biblioteca rempsyc
table_tests_fin<-subset(table_tests, select = -c(Model,Dagnostic) )

kable(t(table_tests_fin)) %>%
kable_styling(bootstrap_options = "striped", full_width = F)
```

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

### ii) Modelo adecuado con transformación de datos.

```
#AIC Usando la función performance_aicc de biblioteca performance
modelo2 <- lm(log(bpsystol) ~ bmi + factor(sex) + age, datos)
#summary(modelo2)
AICY<-performance_aic(modelo2)
```

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

AIC, tenemos para el Modelo 1 es de 2507,213 y para el Modelo 2, considerando que la transformación a la variable dependiente fue logarítmica, es de 2485.7472039 lo cual es menor, esto favorece la elección del Modelo 2.

```
#Modelo 2 con logaritmos
modelo2=lm(data=datos, I(log(bpsystol)) ~ bmi + sex + age)
#summary(modelo2)

#stargazer(modelo1, modelo2)
```

## MODELOS

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

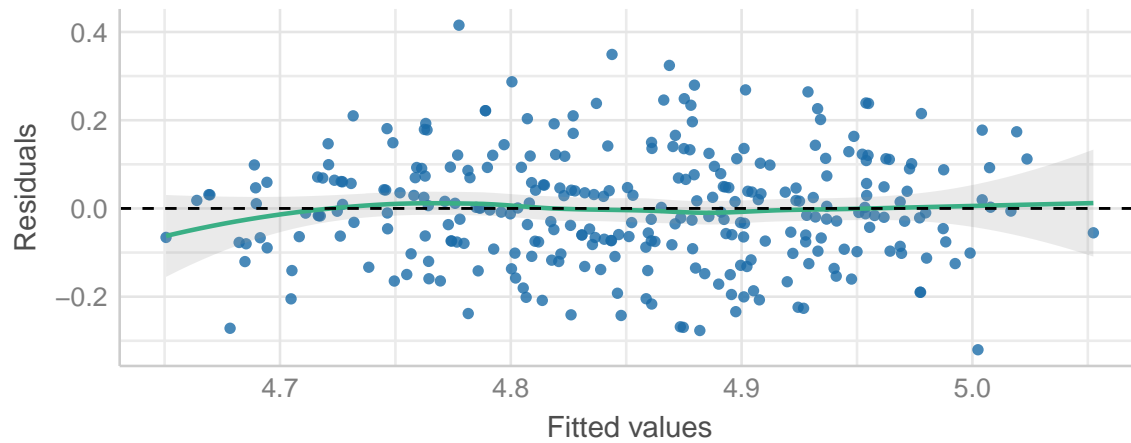
Para poder tener una interpretación válida de los coeficientes, veremos que el Modelo 2 cumple con los supuestos del modelo de regresión lineal. Primero se mostrarán algunas gráficas. La Gráfica **Residuals vs Fitted Values**, se utiliza para comprobar los supuestos de relación lineal, una línea horizontal,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

```
par(mfrow = c(2, 2))
##check_model() function of performance package

# return a list of single plots
diagnostic_plots <- plot(check_model(modelo2, panel = FALSE))
# linearity
diagnostic_plots[[2]]
```

## Linearity

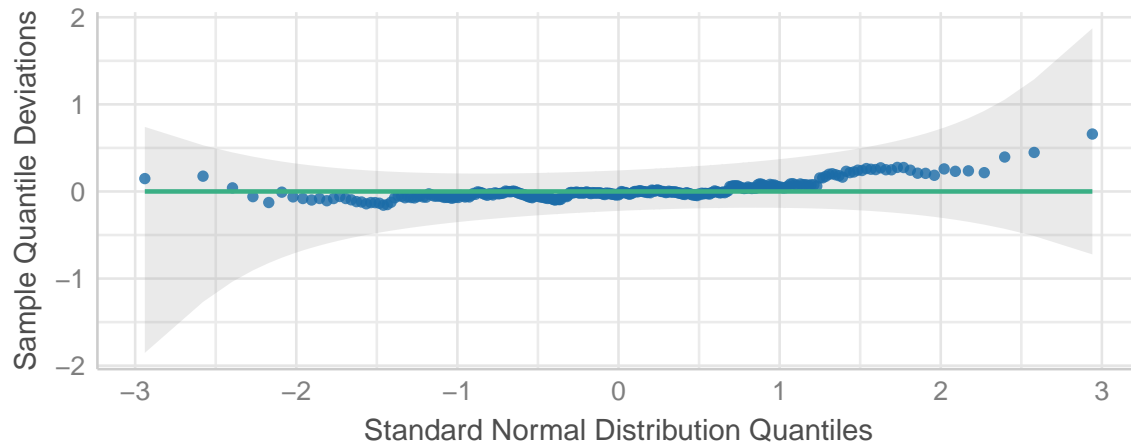
Reference line should be flat and horizontal



```
# normally distributed residuals  
diagnostic_plots[[6]]
```

## Normality of Residuals

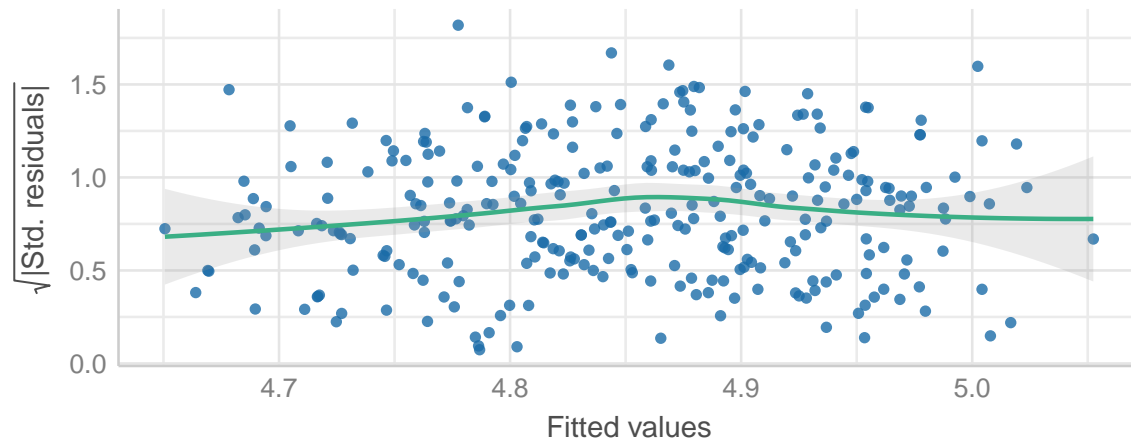
Dots should fall along the line



```
# homoscedasticity - homogeneity of variance  
diagnostic_plots[[3]]
```

## Homogeneity of Variance

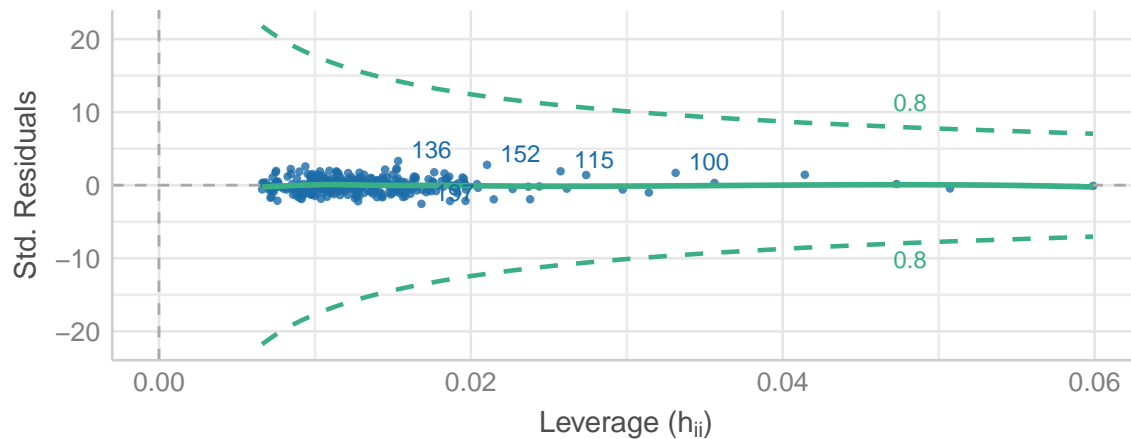
Reference line should be flat and horizontal



```
# influential observations - outliers
diagnostic_plots[[4]]
```

## Influential Observations

Points should be inside the contour lines



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

```
# nice_assumptions() function of rempsyc package
table_tests2<-nice_assumptions(modelo2)
table_tests_fin2<-subset(table_tests2, select = -c(Model,Diagnostic) )
# Table
kable(t(table_tests_fin2)) %>%
kable_styling(bootstrap_options = "striped", full_width = F)
```

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

```
#Las AIC no son comparables directamente por la transformación hecha
#AICmodelo1<-AIC(modelo1)
#AICmodelo2<-AIC(modelo2)
```

### iii) Asociación entre masa corporal y presión arterial sistólica.

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

```
#La primera prueba que se debe realizar es
#prueba asociada a la tabla ANOVA de regresión lineal múltiple
```

```
library(multcomp)
K=matrix(c(0,1,0,0), ncol=4, nrow=1)
m=0

prueba_dir<-glht(modelo2, linfct=K, rhs=m, alternative = "greater")
summary(prueba_dir)
```

```
##
##      Simultaneous Tests for General Linear Hypotheses
##
## Fit: lm(formula = I(log(bpsystol)) ~ bmi + sex + age, data = datos)
##
## Linear Hypotheses:
##           Estimate Std. Error t value      Pr(>t)
## 1 <= 0 0.009365    0.001520   6.161 0.0000000012 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```

```
# Se rechaza H0, lo que implica que se puede proceder al análisis del modelo
```

El resultado de la prueba con dirección Simultaneous Tests for General Linear Hypotheses con el ajuste  $lm(formula = I(\log(bpsystol)) \sim bmi + sex + age, data = datos)$  muestra un p-value de  $1.2e - 09$ , lo cual rechaza la hipótesis nula planteada. La información de los datos proporcionan evidencia en contra de la hipótesis nula por lo que con un nivel de confianza del 95 % es plausible decir que para una persona de cierta edad y sexo, tener un índice de masa corporal alto se asocia con una alta presión arterial sistólica.

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

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

```
curva_ajustada_mujer30 <- function(x) {exp(modelo2$coefficients[1] + modelo2$coefficients[2]*x +
                                           modelo2$coefficients[3] + modelo2$coefficients[4]*30)}
curva_ajustada_hombre30 <- function(x) {exp(modelo2$coefficients[1] + modelo2$coefficients[2]*x +
                                           modelo2$coefficients[4]*30)}

curva_ajustada_mujer50 <- function(x) {exp(modelo2$coefficients[1] + modelo2$coefficients[2]*x +
```

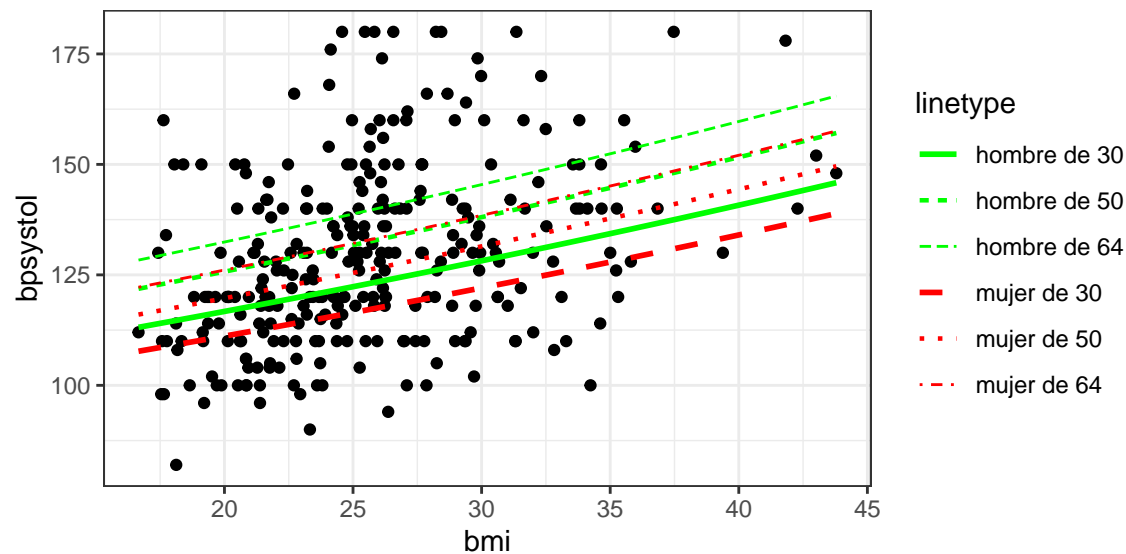
```

                                modelo2$coefficients[3] + modelo2$coefficients[4]*50))}
curva_ajustada_hombre50 <- function(x) {exp(modelo2$coefficients[1] + modelo2$coefficients[2]*x +
                                modelo2$coefficients[4]*50)}

curva_ajustada_mujer64 <- function(x) {exp(modelo2$coefficients[1] + modelo2$coefficients[2]*x +
                                modelo2$coefficients[3] + modelo2$coefficients[4]*64)}
curva_ajustada_hombre64 <- function(x) {exp(modelo2$coefficients[1] + modelo2$coefficients[2]*x +
                                modelo2$coefficients[4]*64)}

ggplot(datos, aes(bmi, bpsystol)) +
  geom_point() +
  geom_function(fun = curva_ajustada_mujer30, aes(linetype = "mujer de 30"), col="red", lwd = 1)+
  geom_function(fun = curva_ajustada_hombre30, aes(linetype = "hombre de 30"), col="green", lwd = 1)+
  geom_function(fun = curva_ajustada_mujer50, aes(linetype = "mujer de 50"), col="red", lwd = 0.8) +
  geom_function(fun = curva_ajustada_hombre50, aes(linetype = "hombre de 50"), col="green", lwd = 0.8)+
  geom_function(fun = curva_ajustada_mujer64, aes(linetype = "mujer de 64"), col="red", lwd = 0.5) +
  geom_function(fun = curva_ajustada_hombre64, aes(linetype = "hombre de 64"), col="green", lwd = 0.5)+
  theme_bw()

```





## 2. Modelos lineales generalizados para datos continuos

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

### i) Explorando modelos con variable dependiente continua.

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

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

```
datos<-read_csv("Preg1B.csv", show_col_types = FALSE)
datos$sex<-factor(datos$sex) #declaramos la variable como factor
#Realizamos un releve para poner como referencia "1" hombre
datos$sex<-relevel(datos$sex,"1")

#Seleccionar un modelo entre un conjunto de posibles glm, mediante mallas

#Comp lineal: a)transf BoxTidwell(potencias) a x y b) polinomio sobre x
#Mallas para el valor de potencia y grado de polinomio:
malla=seq(from = 1, to = 5, by = 1)
Poli <- cbind("poly", malla)
malla=seq(from = -3, to = 3, by = .5)
Pot <- cbind("pot", malla)
CompLin=rbind(Poli, Pot)

#Componente aleatorio: Y es continua y positiva, tenemos tres opciones:
#Malla con distribucion Normal, Gaussiana e Inversa Gaussiana
Distribuciones=c("gaussian", "Gamma", "inverse.gaussian")

#Funcion liga:inverse, identity, log, y 1/mu^2(solo para IG)
FunLigas=c("identity", "log", "inverse", "1/mu^2")

#Declaramos longitud o dimensión de los vectores
nFunLigas=length(FunLigas)
nDist=length(Distribuciones)
nCompLin=dim(CompLin)[1]

#Creación de variables para guardar resultados
ModelList=list(NA) #guardar resultados del ajuste, objeto glm
AICList=list(NA) #guardar el AIC del modelo
BICList=list(NA) #guardar el BIC del modelo
FormList=list(NA) #guardar la formula usada para el ajuste

#Modelos 18*2*3+18*1*4(tres funciones ligas para 2 distrib y 4 para una, IG)
```

```

#Generamos los ciclos y combinaciones de las mallas
index=0
for(k in 1:nCompLin){
  #definimos componente lineal y formula
  if(CompLin[k,1]=="poly"){
    formstring=paste0("bpsystol ~ age+sex+poly(bmi,", CompLin[k,2], ", raw=TRUE)")
  }else{
    if(CompLin[k,2]==0){
      formstring=paste0("bpsystol ~ age+sex+I(log(bmi))")}else
      {
        formstring=paste0("bpsystol ~ age+sex+I(bmi^(", CompLin[k,2], "))")
      }
    }
  form <- as.formula(formstring)
  for(j in 1:nDist){
    for(l in 1:nFunLigas){
      #definicion del argumento family
      if(FunLigas[l]=="1/mu^2"){
        if(Distribuciones[j]=="inverse.gaussian"){
          index=index+1
          Dist=get(Distribuciones[j]) #obtener la funcion a usar
          Mod.A.Prueba=glm(form, data=datos, family = Dist(link=FunLigas[l]))
          ModelList[[index]]=Mod.A.Prueba
          AICList[[index]]=AIC(Mod.A.Prueba)
          BICList[[index]]=BIC(Mod.A.Prueba)
          FormList[[index]]=formstring
        }
      }else{
        index=index+1
        Dist=get(Distribuciones[j])
        Mod.A.Prueba=glm(form, data=datos, family = Dist(link=FunLigas[l]))
        ModelList[[index]]=Mod.A.Prueba
        AICList[[index]]=AIC(Mod.A.Prueba)
        BICList[[index]]=BIC(Mod.A.Prueba)
        FormList[[index]]=formstring
      }
    }
  }
}

```

```

#El modelo con menor AIC
MinAIC=which.min(unlist(AICList))
ModMinAIC=ModelList[[MinAIC]]
summary(ModMinAIC)

```

```

##
## Call:
## glm(formula = form, family = Dist(link = FunLigas[l]), data = datos)
##
## Coefficients:
##              Estimate Std. Error t value      Pr(>|t|)
## (Intercept)  90.16891    3.90205   23.108 < 0.0000000000000002 ***
## age          0.48671    0.05732    8.491 0.000000000000000106 ***
## sex2        -7.05833    1.90756   -3.700    0.000258 ***
## I(bmi^(1.5)) 0.15131    0.02619    5.777 0.00000001948758568 ***

```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for inverse.gaussian family taken to be 0.0001280556)
##
##      Null deviance: 0.053716  on 294  degrees of freedom
## Residual deviance: 0.036270  on 291  degrees of freedom
## AIC: 2484
##
## Number of Fisher Scoring iterations: 5
```

```
#Mostrar los AIC, BIC, modelo y liga
ModMinAIC$family
```

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

```
AICList[[MinAIC]]
```

```
## [1] 2484.009
```

```
BICList[[MinAIC]]
```

```
## [1] 2502.443
```

```
FormList[[MinAIC]]
```

```
## [1] "bpsystol ~ age+sex+I(bmi^(1.5))"
```

Sin embargo, se elige el modelo más simple o parsimonioso sin el exponente de 1,5 para la variable bmi, pues al considerar bmi sin modificación se obtiene un AIC de 2484.1, el cual no parece ser muy diferente a 2484.009. En el siguiente Cuadro de MODELOS 1, columna (2) se muestra el modelo final elegido.

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

```
modelo_glm1<-glm(formula = bpsystol ~ age+sex+bmi, family = inverse.gaussian(link = identity), data = d
summary(modelo_glm1)
```

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

```
## Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	80.02163	5.25361	15.232	< 0.0000000000000002 ***
age	0.48269	0.05744	8.403	0.00000000000000195 ***
sex2	-6.88649	1.90588	-3.613	0.000356 ***
bmi	1.17620	0.20261	5.805	0.00000001680617507 ***

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for inverse.gaussian family taken to be 0.0001279792)
##
##      Null deviance: 0.053716  on 294  degrees of freedom
## Residual deviance: 0.036286  on 291  degrees of freedom
## AIC: 2484.1
##
```

```
## Number of Fisher Scoring iterations: 5
```

```
#stargazer(ModMinAIC, modelo_glm1)
```

# MODELOS 1

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

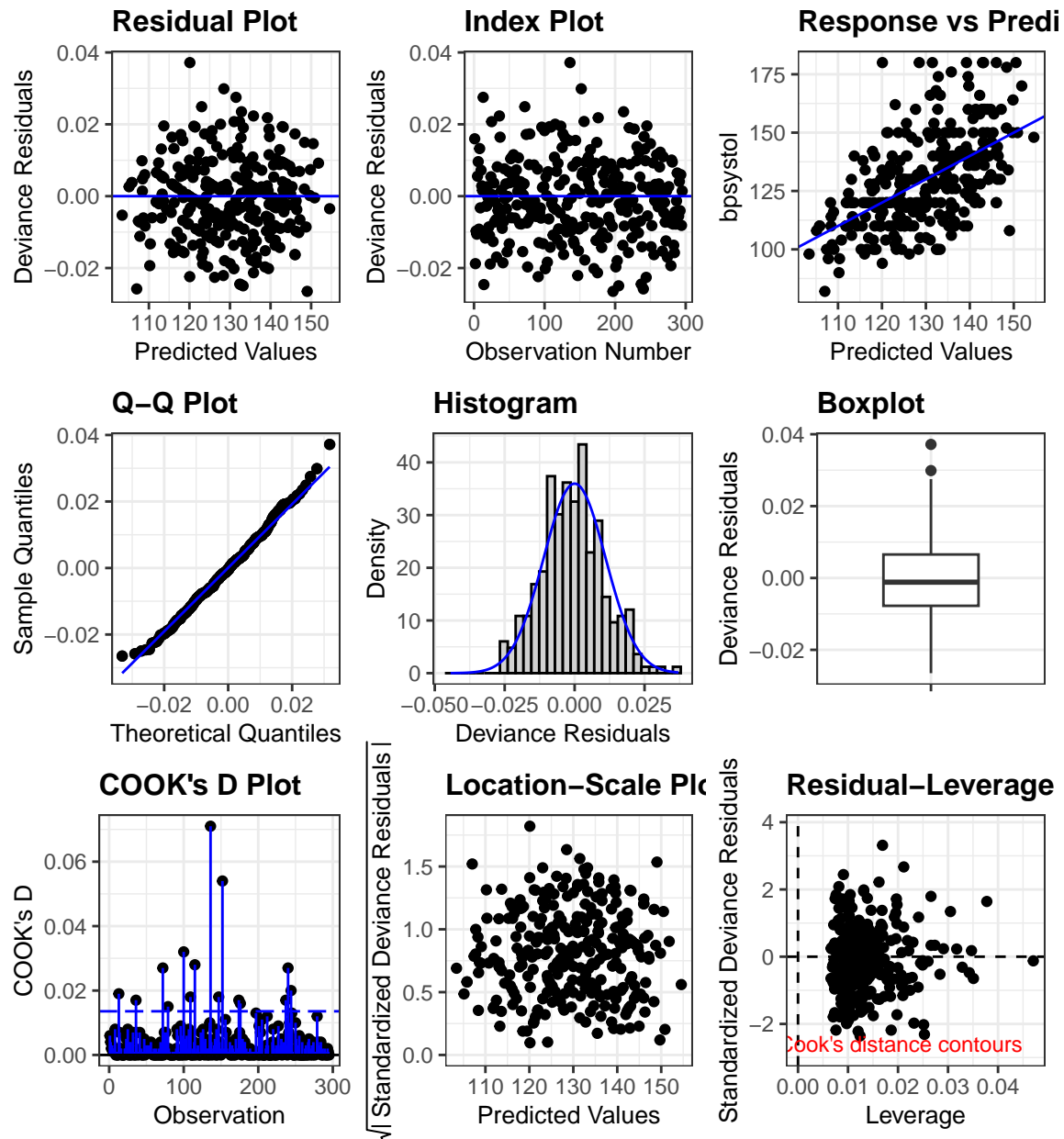
La prueba de hipótesis global con la chi-cuadrada del modelo lineal general **inversa gaussiana con liga identidad** muestra un valor Chisq de 142.2139 y un p-value muy pequeño ( $\Pr(>\text{Chisq})$ : 1.259176e-30), mucho menor a 0.05, es decir se rechaza la hipótesis nula, por lo que podemos proceder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

```
# Se incluye cierta aleatorización para datos binarios
library(statmod)
fitlogitqr <- qresid(modelo_glm1)
#qqnorm(fitlogitqr, las=1); qqline(fitlogitqr)
lilKS<-nortest::lillie.test(fitlogitqr)
Shapiro<-shapiro.test(fitlogitqr)
```

En la prueba de normalidad Lilliefors (Kolmogorov-Smirnov) **normality test** tenemos que el p-value es de 0.560576574569643, por lo que no se rechaza la hipótesis nula de normalidad. Por otra parte, pa la prueba de normalidad de Shapiro-Wilk **normality test** el p-value es de 0.44368350946798, lo que también no rechaza la hipótesis nula de normalidad.

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

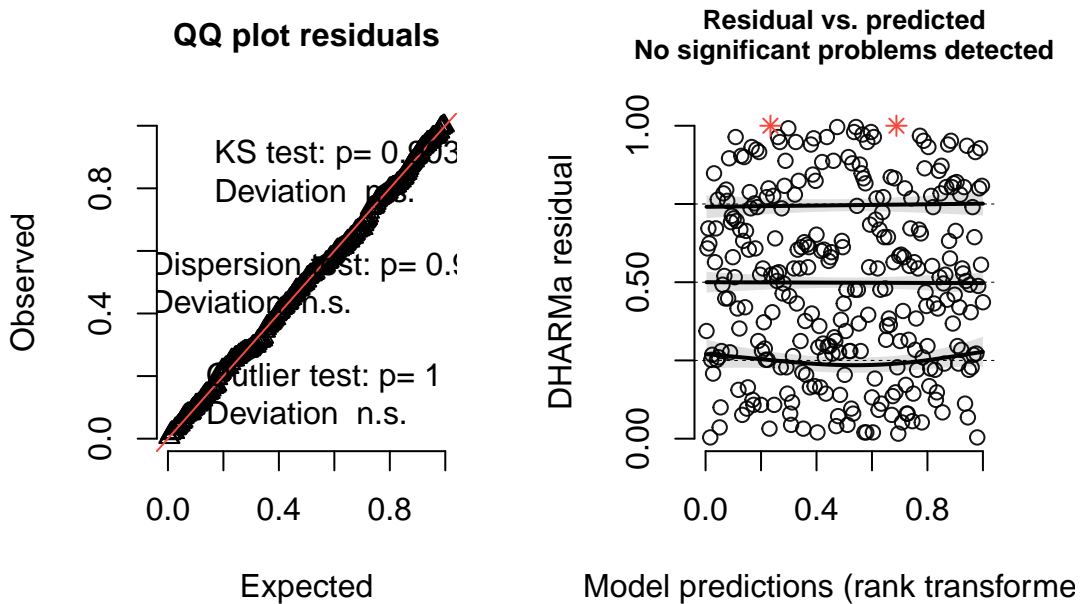
```
resid_panel(modelo_glm1, plots=c("all"),scale = 1)
```



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

```
set.seed(123)
fit2res <- simulateResiduals(fittedModel = modelo_glm1)
plot(fit2res)
```

## DHARMA residual



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

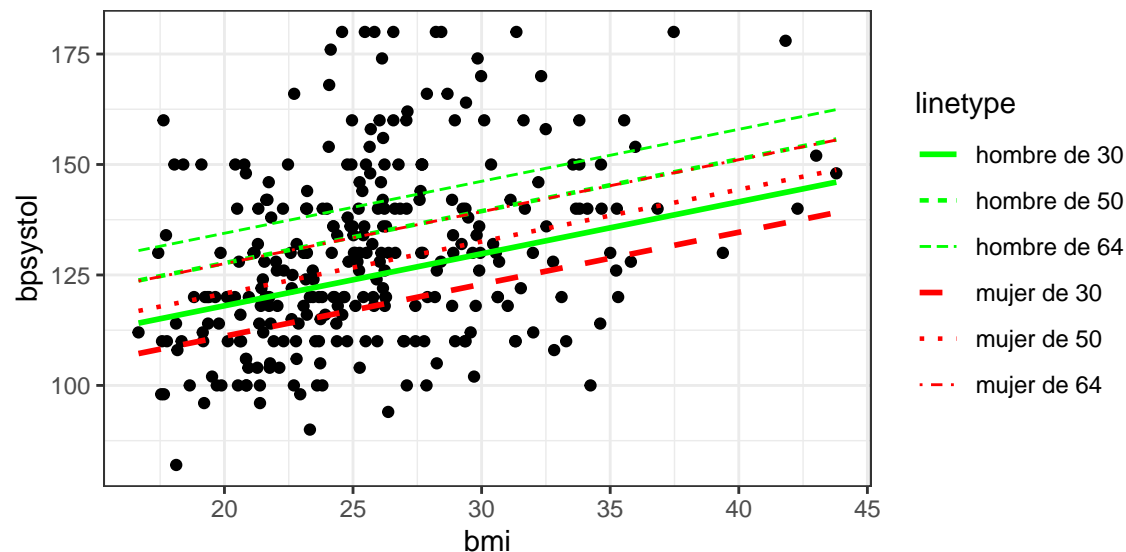
En esta sección describiremos la asociación entre masa corporal y presión arterial sistólica y la prueba de hipótesis de esta relación. Dado que lo que buscamos responder es si tener un índice de masa corporal alto se relaciona con tener una presión sistólica alta agregaremos una prueba de hipótesis con dirección, donde la hipótesis nula es  $H_0 : \beta_3 \leq 0$  contra la alternativa  $H_0 : \beta_3 > 0$ . El p-value asociado a la prueba es de  $\Pr(>z) = 3.22e-09$ , por lo tanto, rechazamos la hipótesis nula y por lo tanto hay relación asociación entre la masa corporal alta y la presión arterial sistólica alta para una persona de cierta edad y sexo.

```
K=matrix(c(0,0,0,1), ncol=4, nrow=1, byrow=TRUE)
m=0
prueba<-glht(modelo_glm1, linfct=K, rhs=m, alternative="greater")
#summary(prueba)
```

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

```
curva_ajustada_mujer30 <- function(x) {modelo_glm1$coefficients[1] + modelo_glm1$coefficients[4]*x + mo
curva_ajustada_hombre30 <- function(x) {modelo_glm1$coefficients[1] + modelo_glm1$coefficients[4]*x + m
curva_ajustada_mujer50 <- function(x) {modelo_glm1$coefficients[1] + modelo_glm1$coefficients[4]*x + mo
curva_ajustada_hombre50 <- function(x) {modelo_glm1$coefficients[1] + modelo_glm1$coefficients[4]*x + m
curva_ajustada_mujer64 <- function(x) {modelo_glm1$coefficients[1] + modelo_glm1$coefficients[4]*x + mo
curva_ajustada_hombre64 <- function(x) {modelo_glm1$coefficients[1] + modelo_glm1$coefficients[4]*x + m
```

```
ggplot(datos, aes(bmi, bpsystol)) +
  geom_point() +
  geom_function(fun = curva_ajustada_mujer30, aes(linetype = "mujer de 30"), col="red", lwd = 1) +
  geom_function(fun = curva_ajustada_hombre30, aes(linetype = "hombre de 30"), col="green", lwd = 1) +
  geom_function(fun = curva_ajustada_mujer50, aes(linetype = "mujer de 50"), col="red", lwd = 0.8) +
  geom_function(fun = curva_ajustada_hombre50, aes(linetype = "hombre de 50"), col="green", lwd = 0.8) +
  geom_function(fun = curva_ajustada_mujer64, aes(linetype = "mujer de 64"), col="red") +
  geom_function(fun = curva_ajustada_hombre64, aes(linetype = "hombre de 64"), col="green") + theme_bw()
```



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

```
#AIC Usando la función performance_aicc de biblioteca performance
modelo2 <- lm(log(bpsystol) ~ bmi + factor(sex) + age, datos)
#summary(modelo2)
AICY <- performance_aic(modelo2)
```

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

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

```
#Modelo Reg Lineal con logaritmos
modelo2=lm(data=datos, I(log(bpsystol)) ~ bmi + sex + age)
#summary(modelo2)
```

```
#summary(modelo_glm1)
#stargazer(modelo2, modelo_glm1)
```

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



### 3. Modelos lineales generalizados para datos binarios

La base de datos Preg3B.csv contiene información sobre 862 insectos que fueron expuestos a diferentes dosis en mg (Deposit) de tres insecticidas (Insecticide). La asignación a una dosis y al tipo de insecticida se realizó de forma aleatoria. Después de seis días se analizó si los insectos se habían muerto, de manera que la base de datos contiene también el número de insectos muertos (Killed) y el número total de insectos expuestos (Number) por cada dosis e insecticida. Dado que se asume que el costo de los insecticidas es el mismo, el objetivo del análisis es identificar para cada insecticida qué dosis es la mínima con la que se puede indicar que el 70 % de los insectos se muere, así como si considerando la menor de esas tres dosis se puede afirmar que un insecticida es el mejor comparado con el resto. El evento de interés es si el insecto muere o no (died).

#### i) Gráfica de dispersión de dosis del insecticida y la proporción de insectos muertos.

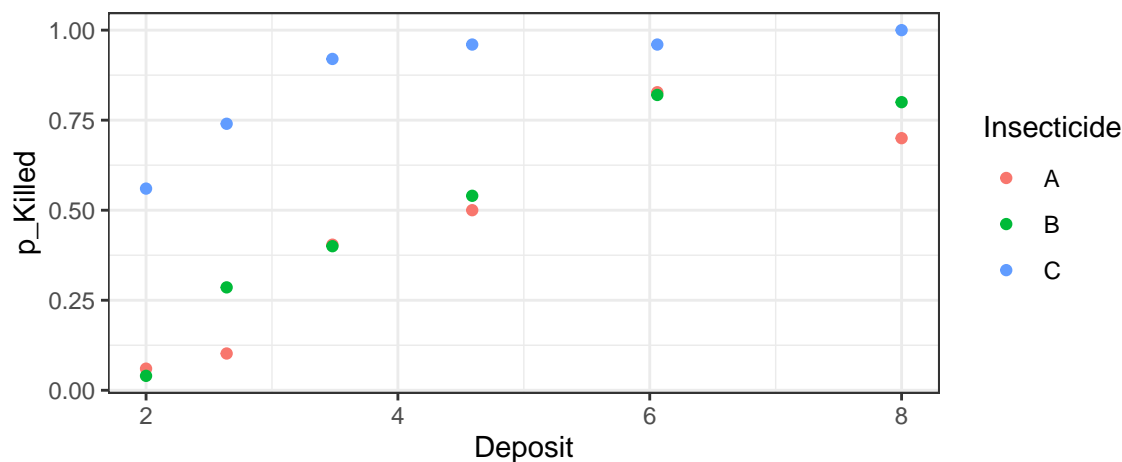
```
datos<-read_csv("https://raw.githubusercontent.com/LeobardoEnriquezH/Data/main/Preg3B.csv", show_col_types=FALSE)
datos$Insecticide<-factor(datos$Insecticide) #declaramos la variable como factor

#Realizamos un releve para poner como referencia "1"
datos$Insecticide<-relevel(datos$Insecticide,"A")

#Creamos dato de proporción de insectos que mueren
datos$p_Killed<-datos$Killed/datos$Number
```

Se presenta una gráfica de dispersión en donde en el eje  $x$  se incluye la dosis del insecticida (Deposit) y en el eje  $y$  la proporción de insectos muertos observados ( $p\_Killed$ ) para cada combinación dosis-insecticida (Deposit-Insecticide), distinguiendo con un color el insecticida asociado. Se puede observar que el insecticida C tiene una mayor tasa de mortalidad para todas las seis dosis consideradas (solamente la primera dosis es menor a 70 %). Para el caso de los insecticidas A y B, los resultados son muy parecidos, aunque marginalmente parece que el insecticida A tiene menor tasa de mortalidad, al menos de manera evidente en tres dosis distintas.

```
ggplot(data=datos, aes(x=Deposit,y=p_Killed, colour=Insecticide))+ geom_point()+ theme_bw()+
  theme(text = element_text(size = 11),element_line(linewidth =0.5))
```



#### ii) Ajuste modelos para datos binarios 1

Ajustaremos modelos para datos binarios (ligas: logit, probit, y cloglog) en donde se incluyen como covariables a Insecticide y  $\ln D$  ( $\ln D = \ln(\text{Deposit})$ ), así como su interacción. Se calcularon los tres modelos con interacciones y se muestran en el siguiente Cuadro. De acuerdo con el criterio AIC el modelo más adecuado es el de la liga probit, cuyo AIC fue de 789.28 (el del logit de 789.44 y cloglog de 800.46). Los términos de las interacciones no son significativas para los tres modelos (no se rechaza la hipótesis nula de que los coeficientes son cero),

mientras que para el intercepto, InsecticideC y lnD sí se rechaza la hipótesis nula. Esto sugiere que podría ser más adecuado el modelo reducido.

Adicionalmente, se calcularon los tres modelos (ligas logit, probit y cloglog) reducidos, sin las interacciones Insecticide-lnD. Todos tienen un menor AIC, en particular el modelo probit. Se puede observar que en estos casos incluso InsecticideB podría ser estadísticamente significativo si consideramos un nivel de significancia estadística del 10 %. Si consideramos el modelo reducido, el modelo probit tiene un mejor desempeño por su AIC y por ser más parsimonioso, con componente lineal o sistemático  $\beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 = -2,623 + 0,209InsecticideB + 1,672InsecticideC + 1,690lnD$ .

```
#Convertimos a datos desagrupados (más común en la práctica)
library(tidyverse)
datos$NoKilled<-datos$Number-datos$Killed
datna_1=datos %>% group_by(Killed,Insecticide,Deposit) %>%
  do( data.frame(died= rep(1, .$Killed)))
datna_2=datos %>% group_by(Killed,Insecticide,Deposit) %>%
  do( data.frame(died= rep(0, .$NoKilled)))
datna<-rbind(datna_1,datna_2)
```

```
#Generar variables con datos en logaritmos
datna$lnD=log(datna$Deposit)
#datos$lnD=factor(datos$lnD) #declaramos la variable como factor
```

```
#En general hay 4 ligas que se podrían usar para covariables factores: logit, probit, cloglog y log.
#Aquí emplearemos 3 ligas porque tenemos una variable numérica entre las covariables, omitimos liga log
fitlogit=glm(died~Insecticide*lnD, family = binomial(link="logit"), data=datna)
summary(fitlogit)
```

```
fitprob=glm(died~Insecticide*lnD, family = binomial(link="probit"), data=datna)
summary(fitprob)
```

```
fitcll=glm(died~Insecticide*lnD, family = binomial(link="cloglog"), data=datna)
summary(fitcll)
```

```
#En general hay 4 ligas que se podrían usar para covariables factores: logit, probit, cloglog y log.
#Aquí emplearemos 3 ligas porque tenemos una variable numérica entre las covariables, omitimos liga log
fitlogit_s=glm(died~Insecticide+lnD, family = binomial(link="logit"), data=datna)
summary(fitlogit)
```

```
fitprob_s=glm(died~Insecticide+lnD, family = binomial(link="probit"), data=datna)
summary(fitprob)
```

```
fitcll_s=glm(died~Insecticide+lnD, family = binomial(link="cloglog"), data=datna)
summary(fitcll)
```

```
stargazer(fitlogit, fitprob, fitcll, fitlogit_s, fitprob_s, fitcll_s)
```

```
#A continuación se muestran los valores AIC y BIC, respectivamente, para las ligas logit, probit y clog
#Se elige el modelo con el menor AIC, que es el de la liga probit.
#Además con el modelo completo también los criterios AIC y BIC coinciden
# ("AIC:logit probit cloglog")
# c(AIC(fitlogit), AIC(fitprob), AIC(fitcll))
# ("BIC:logit probit cloglog")
# c(BIC(fitlogit), BIC(fitprob), BIC(fitcll))
```

La prueba de hipótesis global con la chi-cuadrada del modelo **probit reducido** muestra un valor Chisq de 264.5619 y un p-value muy pequeño (Pr(>Chisq): 4.633875e-57), mucho menor a 0.05, es decir se rechaza la

	Dependent variable:					
	died					
	<i>logistic</i>	<i>probit</i>	<i>glm: binomial link = cloglog</i>	<i>logistic</i>	<i>probit</i>	<i>glm: binomial link = cloglog</i>
	(1)	(2)	(3)	(4)	(5)	(6)
InsecticideB	0.188 (0.722)	0.105 (0.400)	0.260 (0.530)	0.349* (0.206)	0.209* (0.120)	0.249* (0.135)
InsecticideC	2.110*** (0.790)	1.505*** (0.433)	2.350*** (0.485)	2.840*** (0.254)	1.672*** (0.141)	1.706*** (0.151)
lnD	2.727*** (0.349)	1.634*** (0.194)	1.861*** (0.234)	2.887*** (0.224)	1.690*** (0.122)	1.714*** (0.134)
InsecticideB:lnD	0.111 (0.487)	0.072 (0.270)	-0.004 (0.319)			
InsecticideC:lnD	0.661 (0.671)	0.137 (0.347)	-0.486 (0.327)			
Constant	-4.231*** (0.524)	-2.543*** (0.289)	-3.377*** (0.392)	-4.461*** (0.356)	-2.623*** (0.194)	-3.138*** (0.238)
Observations	862	862	862	862	862	862
Log Likelihood	-388.721	-388.640	-394.229	-389.246	-388.727	-395.786
Akaike Inf. Crit.	789.443	789.280	800.458	786.491	785.454	799.571

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

hipótesis nula, por lo que podríamos proceder con el análisis de los supuestos del modelo. Antes de continuar, revisaremos en el siguiente inciso algunos modelos que incluyan  $(\ln D)^2$ , y veremos si tienen menor AIC.

*#La primera prueba que se debe realizar es*

*#la similar a la prueba asociada a la tabla ANOVA en regresión lineal múltiple*

```
library(multcomp)
K=matrix(c(0,1,0,0,
           0,0,1,0,
           0,0,0,1), ncol=4, nrow=3, byrow=TRUE)
m=c(0,0,0)
summary(glht(fitprob_s, linfct=K, rhs=m), test=Chisqtest()) #Chisqtest() es apropiada para datos donde
```

```
##
## General Linear Hypotheses
##
## Linear Hypotheses:
## Estimate
## 1 == 0 0.2086
## 2 == 0 1.6721
## 3 == 0 1.6900
##
## Global Test:
## Chisq DF Pr(>Chisq)
## 1 264.6 3 4.634e-57
```

*# Se rechaza H0, lo que implica que se puede proceder  
# al análisis del modelo*

### iii) Ajuste modelos para datos binarios 2

A continuación incluiremos, adicional a los términos de las covariables anteriores, a la interacción de Insecticide con el término cuadrático  $(\ln D)^2$ . Para las ligas logit y probit, ninguna de las intersecciones con  $\ln D$  y  $(\ln D)^2$  rechazan la hipótesis nula, es decir ninguna aparece estadísticamente significativa porque el p-value asociado es mayor a 0.05. Para el caso del cloglog, la única intersección estadísticamente significativa al 5 % de significancia estadística es  $\text{InsecticideC}:\ln D$ . En los tres modelos se rechaza la hipótesis nula para el intercepto,  $\text{InsecticideC}$ ,  $\ln D$  y  $\ln D^2$ . Los AIC son 786.61, 786.92 y 786.06 para los modelos con liga logit, probit y cloglog, respectivamente, lo que indica que el mejor modelo por el criterio AIC es el de la liga cloglog.

Adicionalmente, se procedió a hacer un modelo reducido con sólo efectos principales, sin estas interacciones y el resultado es que hay menores AIC para los tres modelos considerando las variables explicativas  $\text{Insecticide}$ ,  $\ln D$  y  $\ln D^2$ , sin las interacciones. Por ejemplo, el menor AIC es de 780.01 para el caso de la liga probit, con componente lineal o sistemático  $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 = -3,920 + 0,195 \text{InsecticideB} + 1,701 \text{InsecticideC} + 3,775 \ln D - 0,750 (\ln D)^2$ .

```
datna$lnD2=datna$lnD*datna$lnD
```

```
#En general hay 4 ligas que se podrían usar para covariables factores: logit, probit, cloglog y log.  
#Aqui emplearemos 3 ligas porque tenemos una variable numérica entre las covariables, omitimos liga log  
fitlogit=glm(died~Insecticide*lnD+Insecticide*lnD2, family = binomial(link="logit"), data=datna)  
#summary(fitlogit)
```

```
fitprob=glm(died~Insecticide*lnD+Insecticide*lnD2, family = binomial(link="probit"), data=datna)  
#summary(fitprob)
```

```
fitcll=glm(died~Insecticide*lnD+Insecticide*lnD2, family = binomial(link="cloglog"), data=datna)  
#summary(fitcll)
```

```
#En general hay 4 ligas que se podrían usar para covariables factores: logit, probit, cloglog y log.  
#Aqui emplearemos 3 ligas porque tenemos una variable numérica entre las covariables, omitimos liga log  
fitlogit_s=glm(died~Insecticide+lnD +lnD2, family = binomial(link="logit"), data=datna)  
#summary(fitlogit_s)
```

```
fitprob_s=glm(died~Insecticide+lnD +lnD2, family = binomial(link="probit"), data=datna)  
#summary(fitprob_s)
```

```
fitcll_s=glm(died~Insecticide+lnD +lnD2, family = binomial(link="cloglog"), data=datna)  
#summary(fitcll_s)
```

```
#stargazer(fitlogit, fitprob, fitcll, fitlogit_s, fitprob_s, fitcll_s)
```

La prueba de hipótesis global con la chi-cuadrada del modelo **probit reducido** muestra un valor Chisq de 254.2325 y un p-value muy pequeño ( $\text{Pr}( > \text{Chisq})$ : 7.974736e-54), mucho menor a 0.05, es decir se rechaza la hipótesis nula, por lo que podemos proceder con el análisis de los supuestos de este modelo reducido más sencillo con el menor AIC.

```
#La primera prueba que se debe realizar es  
#la similar a la prueba asociada a la tabla ANOVA en regresión lineal múltiple
```

```
library(multcomp)  
K=matrix(c(0,1,0,0,0,  
           0,0,1,0,0,  
           0,0,0,1,0,  
           0,0,0,0,1), ncol=5, nrow=4, byrow=TRUE)  
m=c(0,0,0,0,0)  
summary(glht(fitprob_s, linfct=K, rhs=m), test=Chisqtest()) #Chisqtest() es apropiada para datos donde
```

	<i>Dependent variable:</i>					
	died					
	<i>logistic</i>	<i>probit</i>	<i>glm: binomial link = cloglog</i>	<i>logistic</i>	<i>probit</i>	<i>glm: binomial link = cloglog</i>
	(1)	(2)	(3)	(4)	(5)	(6)
InsecticideB	2.013 (2.589)	0.679 (1.361)	1.973 (2.085)	0.325 (0.204)	0.195 (0.120)	0.221* (0.133)
InsecticideC	6.150** (2.684)	2.934** (1.388)	6.139*** (1.825)	2.976*** (0.271)	1.701*** (0.145)	1.663*** (0.152)
lnD	9.085*** (2.778)	4.717*** (1.474)	8.599*** (2.173)	6.813*** (1.408)	3.775*** (0.782)	4.117*** (0.857)
lnD2	-2.167** (0.918)	-1.066** (0.499)	-2.198*** (0.691)	-1.407*** (0.491)	-0.750*** (0.276)	-0.844*** (0.295)
InsecticideB:lnD	-2.479 (3.663)	-0.773 (1.982)	-2.376 (2.790)			
InsecticideC:lnD	-5.238 (4.300)	-1.872 (2.198)	-5.572** (2.530)			
InsecticideB:lnD2	0.839 (1.231)	0.277 (0.678)	0.748 (0.895)			
InsecticideC:lnD2	1.971 (1.656)	0.628 (0.817)	1.558* (0.843)			
Constant	-8.512*** (1.993)	-4.560*** (1.022)	-8.123*** (1.637)	-6.946*** (0.967)	-3.920*** (0.524)	-4.647*** (0.594)
Observations	862	862	862	862	862	862
Log Likelihood	-384.307	-384.460	-384.028	-385.098	-385.006	-391.604
Akaike Inf. Crit.	786.613	786.919	786.055	780.196	780.011	793.208

Note:

\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

```
##
##   General Linear Hypotheses
##
## Linear Hypotheses:
##       Estimate
## 1 == 0    0.1947
## 2 == 0    1.7008
## 3 == 0    3.7753
## 4 == 0   -0.7501
##
## Global Test:
##   Chisq DF Pr(>Chisq)
## 1 254.2  4 7.975e-54
```

*# Se rechaza H0, lo que implica que se puede proceder  
# al análisis del modelo*

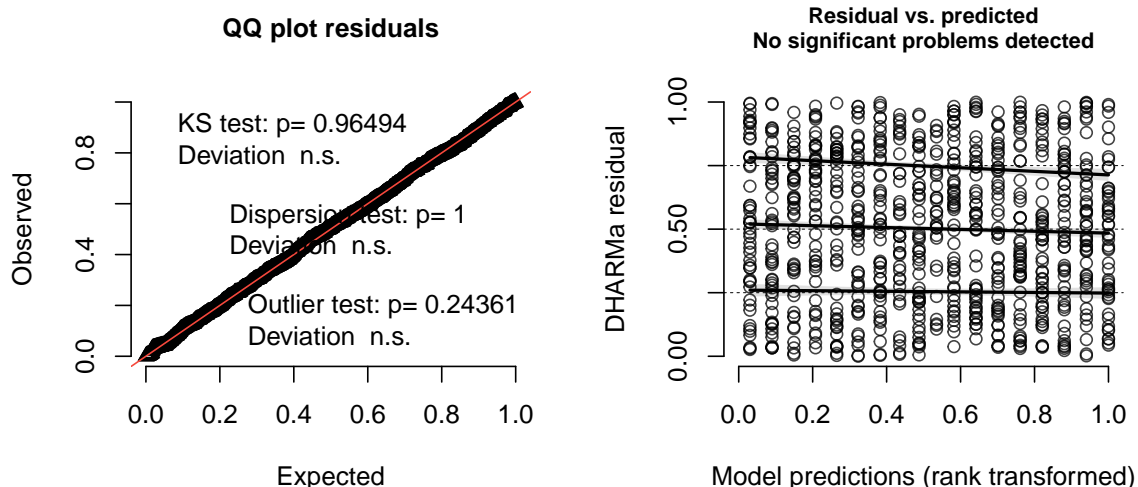
Se puede notar que una ventaja de introducir el componente  $(\ln D)^2$  es que los AIC disminuyeron, por lo que nos quedamos con este modelo probit reducido, para los análisis subsecuentes.

```
# Se incluye cierta aleatorización para datos binarios
library(statmod)
fitlogitqr <- qresid(fitprob_s)
#qqnorm( fitlogitqr, las=1 ); qqline( fitlogitqr)
lilKS<-nortest::lillie.test(fitlogitqr)
Shapiro<-shapiro.test(fitlogitqr)
```

En la prueba de normalidad Lilliefors (Kolmogorov-Smirnov) normality test tenemos que el p-value es de 0.616389122230475, por lo que no se rechaza la hipótesis nula de normalidad. Por otra parte, pa la prueba de normalidad de Shapiro-Wilk normality test el p-value es de 0.592709734522069, lo que también no rechaza la hipótesis nula de normalidad. Esto se observa en la siguiente Gráfica.

```
library(DHARMA) #Los residuales simulados también son útiles en este caso
set.seed(123)
fitprobitres <- simulateResiduals(fittedModel = fitprob_s)
plot(fitprobitres)
```

DHARMA residual



```
deviance_df<-deviance(fitprob_s)/df.residual(fitprob_s)
```

La regla de dedo para verificar el **parámetro de dispersión** de 1, con la devianza de residuales entre grados de libertad, muestra un valor de 0.8984962, lo cual se acerca a 1.

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

```
newdata <- data.frame(Insecticide = c("A", "A","A","A", "A","A","B","B", "B", "B","B", "B","C","C", "C")
newdata$prob_s <- predict(fitprob_s, newdata[,1:3], type = c("response"), se.fit=TRUE)$fit

newdata$Deposit<-exp(newdata$lnD)

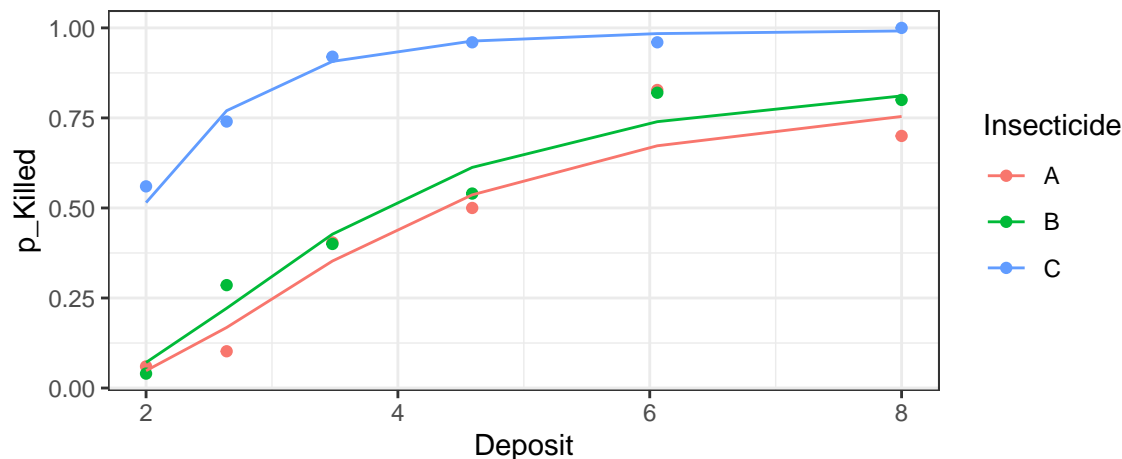
kable(newdata)
```

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

La siguiente gráfica de dispersión muestra en el eje  $x$  la dosis del insecticida (Deposit) y en el eje  $y$  la proporción de insectos muertos observados (se generó la variable  $p\_Killed$ ) para cada combinación dosis-insecticida (Deposit-Insecticide), distinguiendo el insecticida asociado por colores. Adicionalmente se agregaron las curvas con las probabilidades obtenidas con el modelo probit para cada dosis e insecticida. Con el modelo se obtuvieron probabilidades muy cercanas a las proporciones o tasas de mortalidad observadas, especialmente para el insecticida C.

```
newdata1<- newdata[, -c(1,5)]
datosfin<-cbind(datos,newdata1)

ggplot(data=datos, aes(x=Deposit,y=p_Killed, colour=Insecticide))+ geom_point() +theme_bw()+
  geom_line(data=datosfin, aes(x=Deposit,y=prob_s, colour=Insecticide))+theme_bw()
```



A continuación se muestra un cuadro de la dosis mínima para cada insecticida con la que se puede indicar que el 70 % de los insectos se muere. Para ello recordemos que  $\Phi^{-1}(0,7) = \beta_0 + \beta_3 \ln D + \beta_4 (\ln D)^2$ ,  $\Phi^{-1}(0,7) = \beta_0 + \beta_1 \text{InsecticidaB} + \beta_3 \ln D + \beta_4 (\ln D)^2$  y  $\Phi^{-1}(0,7) = \beta_0 + \beta_2 \text{InsecticidaC} + \beta_3 \ln D + \beta_4 (\ln D)^2$ , por lo que resolviendo para cada insecticida, se obtienen los respectivos valores de  $\ln D$  y por lo tanto de  $D$  que es la dosis en mg (Deposit). Es decir, para A, resolveremos  $\beta_4 (\ln D)^2 + \beta_3 \ln D + (\beta_0 - \Phi^{-1}(0,7))$ , para B  $\beta_4 (\ln D)^2 + \beta_3 \ln D + (\beta_0 + \beta_1 - \Phi^{-1}(0,7))$  y para C  $\beta_4 (\ln D)^2 + \beta_3 \ln D + (\beta_0 + \beta_2 - \Phi^{-1}(0,7))$ .

*#Resolver función cuadrática*

```
quad <- function(a, b, c)
{
  a <- as.complex(a)
  answer <- c((-b + sqrt(b^2 - 4 * a * c)) / (2 * a),
              (-b - sqrt(b^2 - 4 * a * c)) / (2 * a))
  if(all(Im(answer) == 0)) answer <- Re(answer)
  if(answer[1] == answer[2]) return(answer[1])
  answer
}
```

*#Para insecticida A*

```
a=fitprob_s$coefficients[5]
b=fitprob_s$coefficients[4]
c=fitprob_s$coefficients[1]-qnorm(.7)
```

```
D_A<-exp(quad(a, b, c))
```

*#Para insecticida B*

```
a=fitprob_s$coefficients[5]
b=fitprob_s$coefficients[4]
c=fitprob_s$coefficients[1]+fitprob_s$coefficients[2]-qnorm(.7)
```

```
D_B<-exp(quad(a, b, c))
```

*#Para insecticida C*

```
a=fitprob_s$coefficients[5]
b=fitprob_s$coefficients[4]
c=fitprob_s$coefficients[1]+fitprob_s$coefficients[3]-qnorm(.7)
```

```
D_C<-exp(quad(a, b, c))
```

Como se observa en la Gráfica anterior, el insecticida C es mejor, pues con menores dosis se tienen mayor





## 4. Modelos lineales generalizados para datos de conteos

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

### i) Gráfica de dispersión de grupos de edad e incidencia

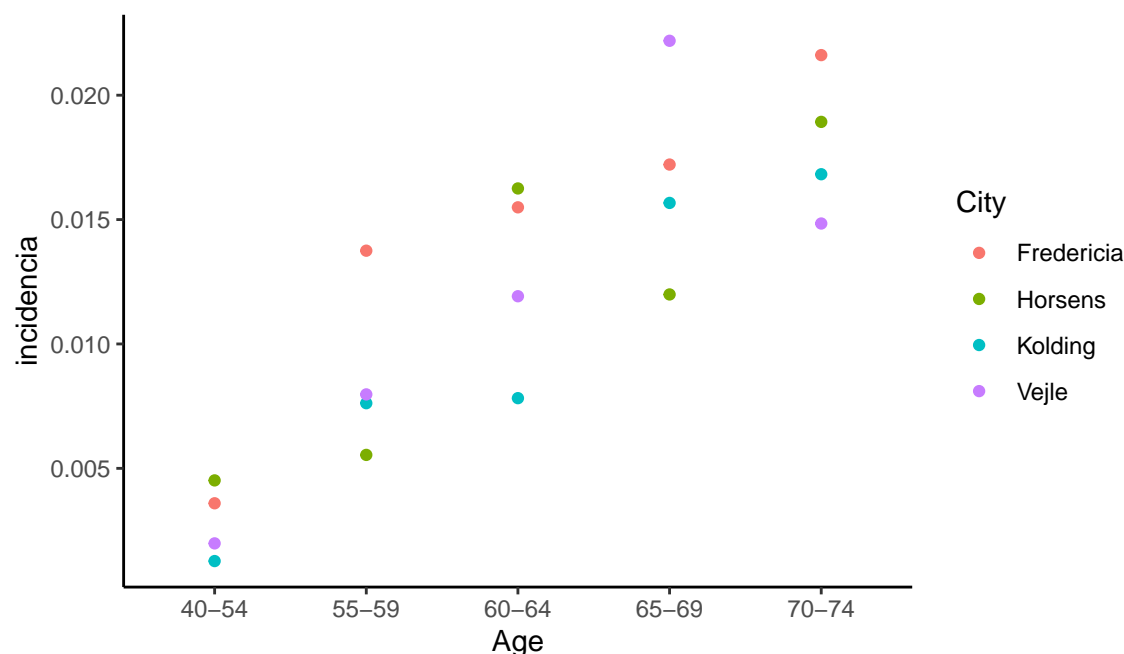
Podemos apreciar de la siguiente Gráfica presentada que por cada grupo de edad la incidencia en cada ciudad va en aumento, por ejemplo en el grupo de edad de 40-54 la incidencia de cáncer esta por debajo de 0.005 pero conforme avanzan los grupos de edad los niveles aumentan para todas las ciudades.

```
data4 <- read_csv("Preg4.csv")
```

```
#Primero haremos factor las variables de ciudad y edad  
data4$City <- as.factor(data4$City)
```

```
#Tambien crearemos la variable que indicara la tasa de incidencia  
data4$incidencia <- data4$Cases/data4$Pop
```

```
ggplot(data=data4, aes(x=Age,y=incidencia, colour=City))+ geom_point()+ theme_classic()+  
  theme(text = element_text(size = 11),element_line(linewidth =0.5))
```



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

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

```
#Lo que nos interesa es ver si es estudiar si se puede indicar que a mayor edad existe mayor incidencia
```

```

# Notar que para dejar en terminos de la variable con conteos,
# al considerar la liga log se debe incluir log(data4$Pop) [offset]

#### log(mu_y/t)=b0+b1x
#### log(mu_y)=log(t)+b0+b1x

data4$logPop=log(data4$Pop)

ajuste1 <- glm(Cases ~ offset(logPop)+ Age * City , family=poisson(link="log"), data=data4)
summary(ajuste1)

##
## Call:
## glm(formula = Cases ~ offset(logPop) + Age * City, family = poisson(link = "log"),
## data = data4)
##
## Coefficients:
## Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.62795 0.30151 -18.666 < 0.0000000000000002 ***
## Age55-59 1.34123 0.42640 3.145 0.001658 **
## Age60-64 1.46058 0.42640 3.425 0.000614 ***
## Age65-69 1.56578 0.43693 3.584 0.000339 ***
## Age70-74 1.79340 0.42640 4.206 0.000026 ***
## CityHorsens 0.22770 0.40967 0.556 0.578343
## CityKolding -1.03837 0.58387 -1.778 0.075335 .
## CityVejle -0.59463 0.53936 -1.102 0.270257
## Age55-59:CityHorsens -1.13671 0.65223 -1.743 0.081368 .
## Age60-64:CityHorsens -0.17991 0.57045 -0.315 0.752471
## Age65-69:CityHorsens -0.58918 0.60649 -0.971 0.331320
## Age70-74:CityHorsens -0.36029 0.58487 -0.616 0.537886
## Age55-59:CityKolding 0.44798 0.74620 0.600 0.548271
## Age60-64:CityKolding 0.35483 0.75807 0.468 0.639737
## Age65-69:CityKolding 0.94450 0.72926 1.295 0.195268
## Age70-74:CityKolding 0.78788 0.73684 1.069 0.284945
## Age55-59:CityVejle 0.04961 0.72434 0.068 0.945398
## Age60-64:CityVejle 0.33237 0.69413 0.479 0.632058
## Age65-69:CityVejle 0.84855 0.67995 1.248 0.212051
## Age70-74:CityVejle 0.21891 0.71191 0.307 0.758469
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
## Null deviance: 115.43443313608482014843 on 19 degrees of freedom
## Residual deviance: -0.00000000000000088818 on 0 degrees of freedom
## AIC: 121.47
##
## Number of Fisher Scoring iterations: 3
# Regla de dedo para analizar si hay un problema por
# considerar el parametro de dispersion igual a 1
dev_res1<-deviance(ajuste1)/df.residual(ajuste1)

```

El AIC obtenido con este Modelo 1 es de 121.47 y de la regla del dedo para analizar si hay un problema por considerar el parametro de dispersion igual a 1 obtuvimos un valor de  $-\infty$ , lo cual nos dice que no es

un buen modelo, de todas formas se hizo la verificación de los supuestos que por cuestión de espacio no se muestra, pero salió muy mal en estos por lo que se decidió no usarse.

```
library(DHARMA)
set.seed(1234)
fitres <- simulateResiduals(fittedModel = ajuste1)
plot(fitres)
```

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

```
##
## Call:
## glm(formula = Cases ~ offset(logPop) + Age, family = poisson(link = "log"),
##      data = data4)
##
## Coefficients:
##              Estimate Std. Error z value      Pr(>|z|)
## (Intercept)  -5.8623      0.1741 -33.676 < 0.0000000000000002 ***
## Age55-59       1.0823      0.2481   4.363  0.0000128549514374 ***
## Age60-64       1.5017      0.2314   6.489  0.000000000000865407 ***
## Age65-69       1.7503      0.2292   7.637  0.00000000000000222 ***
## Age70-74       1.8472      0.2352   7.855  0.00000000000000040 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##      Null deviance: 115.434  on 19  degrees of freedom
## Residual deviance:  16.978  on 15  degrees of freedom
## AIC: 108.45
##
## Number of Fisher Scoring iterations: 4
#Regla de dedo para analizar si hay un problema por
# considerar el parametro de dispersion igual a 1
dev_res<-deviance(ajuste2)/df.residual(ajuste2)
```

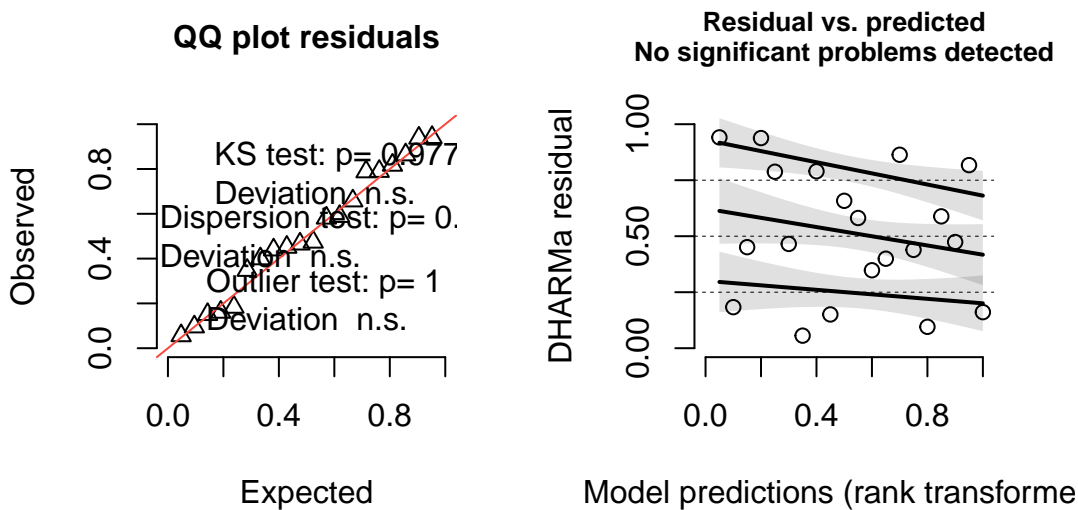
Vemos que este Modelo 2 tuvimos un valor de 1.131886, muy cercano a 1 con la regla de dedo para analizar si hay un problema por considerar el parametro de dispersion igual a 1, lo cual nos dice que es buen modelo, por lo que procedimos con la verificación de los supuestos del modelo de manera gráfica.

```
library(DHARMA)
set.seed(1234)
fitres_2 <- simulateResiduals(fittedModel = ajuste2)
plot(fitres_2)
```

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

*Note:* <sup>28</sup>  
\*p<0.1; \*\*p<0.05; \*\*\*p<0.01

## DHARMA residual



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

```
anova(ajuste1,ajuste2, test = "Chisq")
```

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

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

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

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

```
library(MASS)
fit_negbin3 <- glm.nb(Cases ~ offset(logPop)+ Age , data = data4, link = "log")
summary(fit_negbin3)
```

```
##
## Call:
## glm.nb(formula = Cases ~ offset(logPop) + Age, data = data4,
##   link = "log", init.theta = 152366.6524)
##
## Coefficients:
##           Estimate Std. Error z value      Pr(>|z|)
```

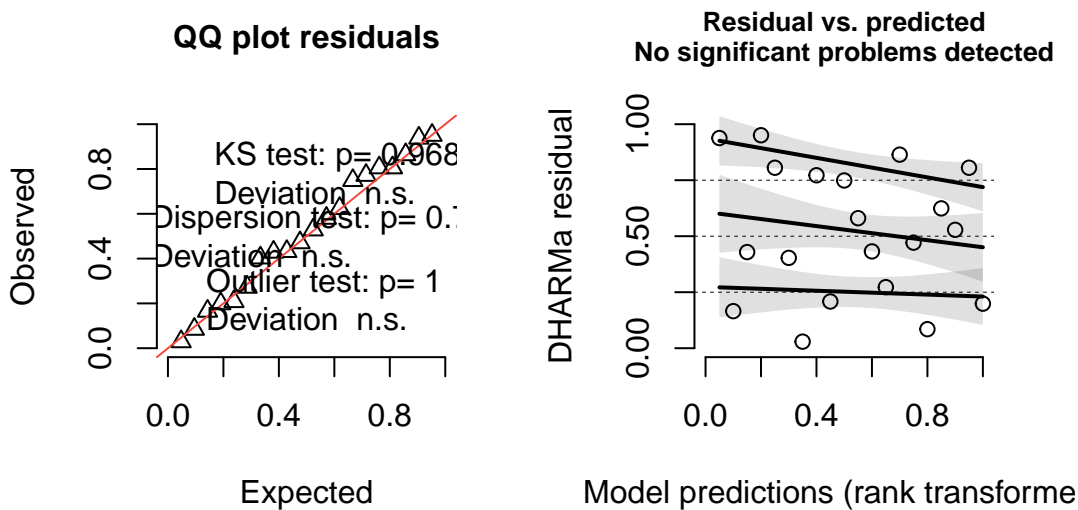
```
## (Intercept) -5.8623      0.1741 -33.675 < 0.0000000000000002 ***
## Age55-59      1.0823      0.2481   4.362  0.00001286240595221 ***
## Age60-64      1.5017      0.2314   6.489  0.000000000008667032 ***
## Age65-69      1.7503      0.2292   7.637  0.000000000000002227 ***
## Age70-74      1.8472      0.2352   7.855  0.000000000000000401 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(152366.7) family taken to be 1)
##
## Null deviance: 115.425  on 19  degrees of freedom
## Residual deviance:  16.977  on 15  degrees of freedom
## AIC: 110.45
##
## Number of Fisher Scoring iterations: 1
##
##
##          Theta: 152367
##      Std. Err.: 5232704
## Warning while fitting theta: iteration limit reached
##
## 2 x log-likelihood: -98.451
```

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

*#Verifiquemos los supuestos de este modelo con DHARMA*  
*#En este caso nuevamente omitiremos los resultados de las pruebas de los supuestos para este modelo ya*

```
set.seed(123)
fit_neg <- simulateResiduals(fittedModel = fit_negbin3)
plot(fit_neg)
```

## DHARMA residual



*#Podemos decir que este tambien parece ser buen modelo, al menos es mucho mejor #que el primero, sin em*

*#Comparemos los AIC y BIC de los ultimos 2 modelos para ver cual es mejor en este #aspecto.*

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

```
("AIC:(2), AIC:(3)")
```

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

```
c(AIC(ajuste2), AIC(fit_negbin3))
```

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

```
("BIC:(2), BIC:(3)")
```

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

```
c(BIC(ajuste2), BIC(fit_negbin3))
```

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

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

*#Para los intervalos de confianza lo que haremos sera definir una malla de valores para la edad.*

```
library(MASS)
```

```
Edad <- seq(from = 40, to =74, by = .5)
```

*#el calculo de los intervalos los haremos a una confianza del 95%*

*#E(y;x)= b0 + b1 incidencia^1.8 + b2 Age*

```
library(multcomp)
```

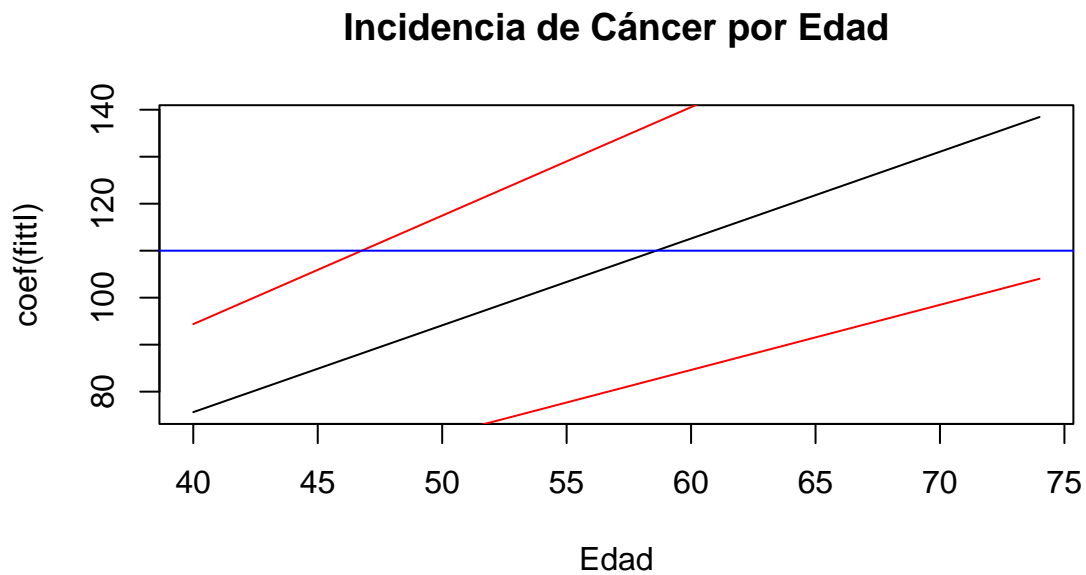
```
K_1 <- cbind(0,0,0,1,Edad)
```

*#banda del primer grupo de edad*



```
fittI <- glht(ajuste2, linfct = K_1)
fitci3 <- confint(fittI, level = 0.95)

plot(Edad, coef(fittI), col="black", type="l", main="Incidencia de Cáncer por Edad")+
lines(Edad, fitci3$confint[, "upr"], col="red")+
lines(Edad, fitci3$confint[, "lwr"], col="red")+
abline(h=110, col="blue")
```



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

```
#Aplicando la funcion inversa para obtener los valores sobre la esperanza de y, omitiremos este resulta  
#exp(fitci3$confint)
```

OBSERVACIÓN: REVISAR Y MODIFICAR GRÁFICA, ASÍ COMO LAS CONCLUSIONES.

## 5. Modelos lineales generalizados para datos categóricos

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

### i) Gráfica de frecuencias relativas

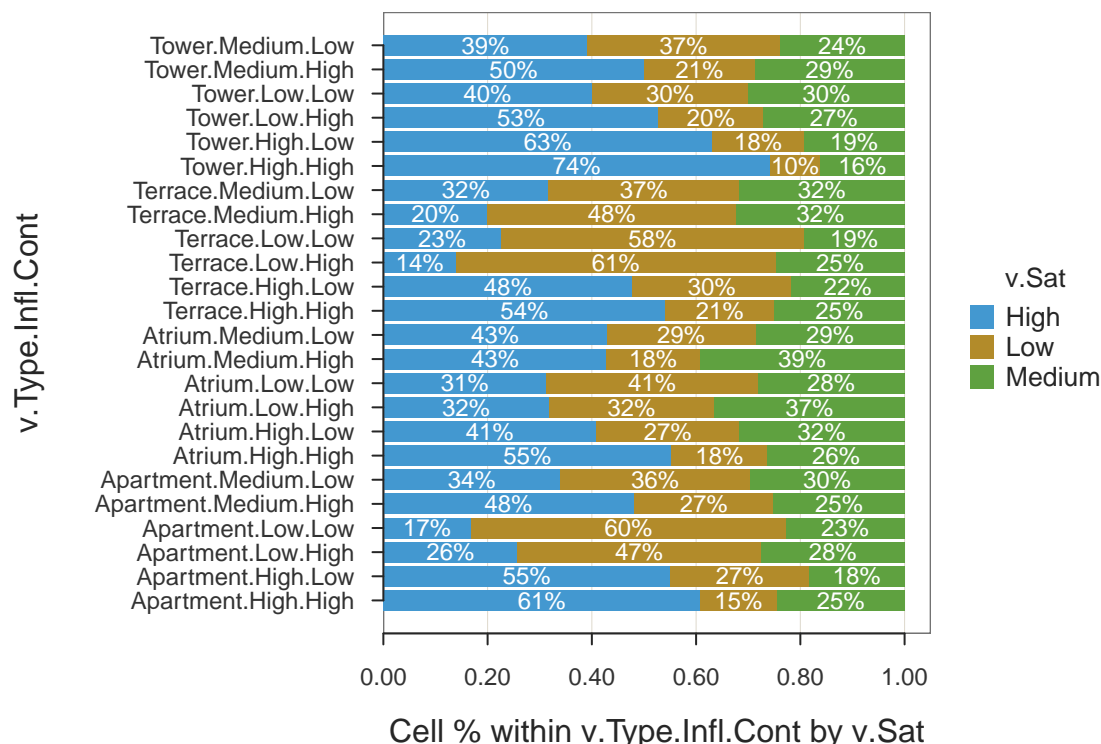
```
Datos[sapply(Datos, is.character)] <- lapply(Datos[sapply(Datos, is.character)], as.factor)
```

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

```
Datos$TypeInflCont=factor(paste(Datos$Type, Datos$Infl, Datos$Cont, sep="."))
v.Type.Infl.Cont=(Datos$TypeInflCont)
v.Sat=Datos$Sat
BarChart(x=v.Type.Infl.Cont , by=v.Sat, stack100=TRUE, srt=0, horiz=TRUE)
```

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

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



```
## >>> Suggestions
```

```
## Plot(v.Type.Infl.Cont, v.Sat) # bubble plot
```

```
## BarChart(v.Type.Infl.Cont, by=v.Sat, horiz=TRUE) # horizontal bar chart
```

```
## BarChart(v.Type.Infl.Cont, fill="steelblue") # steelblue bars
```

```
##
## Cramer's V: 0.252
##
## Chi-square Test of Independence:
##      Chisq = 213.070, df = 46, p-value = 0.000
```

## ii) Modelo logístico multinomial nominal

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

*#Primero ajustamos un modelo de Satisfacción (Sat) con todas las interacciones posibles  
#Entre influencia sobre mantenimiento, tipo de vivienda y contacto con otros vecinos (Infl, Type y Cont.*

```
fit_comp <- vglm(Sat ~ Infl*Type*Cont, #Fórmula
                family = multinomial(refLevel = "Low"), #Familia
                data = Datos) #Datos
#Funciona de forma equivalente a los glm
```

```
#Posteriormente un modelo que no contenga ninguna interacción
fit_no_int <- vglm(Sat ~ Infl+Type+Cont,
                  family = multinomial(refLevel = "Low"),
                  data = Datos)
```

```
#summary(fit_comp)
#coef(fit_comp, matrix = TRUE) # Le indicamos que nos los muestre como matriz
```

```
#summary(fit_no_int)
#coef(fit_no_int, matrix = TRUE) # Le indicamos que nos los muestre como matriz
```

```
#Escribimos los dos modelos y la salida es la prueba
anova(fit_no_int, fit_comp, test="LRT", type = "I")
```

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

```
AIC_fit_comp<-AIC(fit_comp)
AIC_fit_no_int<-AIC(fit_no_int)
```

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

```
summary(fit_no_int)
```

```
##
## Call:
```

```
## vglm(formula = Sat ~ Infl + Type + Cont, family = multinomial(refLevel = "Low"),
##      data = Datos)
##
## Coefficients:
##              Estimate Std. Error z value      Pr(>|z|)
## (Intercept):1    1.2201    0.1585   7.699 0.0000000000000137 ***
## (Intercept):2    0.1709    0.1825   0.936    0.349200
## InflLow:1       -1.6126    0.1671  -9.649 < 0.0000000000000002 ***
## InflLow:2       -0.6649    0.1863  -3.568    0.000359 ***
## InflMedium:1    -0.8778    0.1641  -5.348 0.00000000890670893 ***
## InflMedium:2    -0.2185    0.1872  -1.167    0.243151
## TypeAtrium:1     0.3277    0.1886   1.737    0.082391 .
## TypeAtrium:2     0.5671    0.1947   2.913    0.003577 **
## TypeTerrace:1   -0.6767    0.1756  -3.854    0.000116 ***
## TypeTerrace:2   -0.2309    0.1748  -1.321    0.186653
## TypeTower:1      0.7356    0.1553   4.738 0.0000021614222276 ***
## TypeTower:2      0.4357    0.1725   2.525    0.011562 *
## ContLow:1       -0.4818    0.1241  -3.881    0.000104 ***
## ContLow:2       -0.3609    0.1324  -2.726    0.006420 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: log(mu[,1]/mu[,2]), log(mu[,3]/mu[,2])
##
## Residual deviance: 3470.084 on 3348 degrees of freedom
##
## Log-likelihood: -1735.042 on 3348 degrees of freedom
##
## Number of Fisher scoring iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
## Reference group is level 2 of the response
```

### iii) Modelo logístico acumulativo (cumulative logit) ordinal

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

```
Datos$Sat=factor(Datos$Sat, ordered=TRUE, levels = c("Low", "Medium", "High"))

#Ajustamos utilizando de nuevo la función vglm: modelo de probabilidades NO proporcionales
fit_ord_nopar <- vglm(Sat ~ Infl+Type+Cont,
                      family = cumulative(parallel = FALSE), #La diferencia es esta
                      data = Datos)

#El orden viene de que respuesta es tipo factor y a esta le pusimos orden

#Ajustamos utilizando de nuevo la función vglm: modelo de probabilidades proporcionales
```

```

fit_ord_par <- vglm(Sat ~ Infl+Type+Cont,
                    family = cumulative(parallel = TRUE), #La diferencia es esta
                    data = Datos)

coef(fit_ord_nopar, matrix=TRUE)

##               logitlink(P[Y<=1]) logitlink(P[Y<=2])
## (Intercept)      -1.4946708      -0.41887754
## InflLow          1.2191686       1.30775003
## InflMedium       0.6267059       0.75794760
## TypeAtrium       -0.4099844      -0.05488411
## TypeTerrace      0.4778385       0.58106044
## TypeTower        -0.6011588      -0.53793695
## ContLow          0.4304870       0.29566055

summary(fit_ord_par)

##
## Call:
## vglm(formula = Sat ~ Infl + Type + Cont, family = cumulative(parallel = TRUE),
##       data = Datos)
##
## Coefficients:
##               Estimate Std. Error z value      Pr(>|z|)
## (Intercept):1 -1.57289    0.12519 -12.564 < 0.0000000000000002 ***
## (Intercept):2 -0.38604    0.11938  -3.234    0.001222 **
## InflLow        1.28882    0.12670  10.172 < 0.0000000000000002 ***
## InflMedium     0.72242    0.12372   5.839    0.00000000524 ***
## TypeAtrium     -0.20616    0.13993  -1.473    0.140675
## TypeTerrace    0.51866    0.13358   3.883    0.000103 ***
## TypeTower      -0.57235    0.11875  -4.820    0.00000143628 ***
## ContLow        0.36028    0.09536   3.778    0.000158 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])
##
## Residual deviance: 3479.149 on 3354 degrees of freedom
##
## Log-likelihood: -1739.575 on 3354 degrees of freedom
##
## Number of Fisher scoring iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
## Exponentiated coefficients:
##      InflLow InflMedium TypeAtrium TypeTerrace TypeTower ContLow
## 3.6284964 2.0594206 0.8137002 1.6797833 0.5641981 1.4337373

coef(fit_ord_par, matrix=TRUE)

##               logitlink(P[Y<=1]) logitlink(P[Y<=2])
## (Intercept)      -1.5728877      -0.3860443
## InflLow          1.2888184       1.2888184

```

```
## InflMedium          0.7224247          0.7224247
## TypeAtrium          -0.2061633         -0.2061633
## TypeTerrace         0.5186648          0.5186648
## TypeTower           -0.5723499         -0.5723499
## ContLow             0.3602845          0.3602845
```

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

```
AIC_fit_ord_par<-AIC(fit_ord_par)
AIC_fit_ord_no_par<-AIC(fit_ord_nopar)
```

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

```
summary(fit_ord_par)
```

```
##
## Call:
## vglm(formula = Sat ~ Infl + Type + Cont, family = cumulative(parallel = TRUE),
##       data = Datos)
##
## Coefficients:
##              Estimate Std. Error z value      Pr(>|z|)
## (Intercept):1 -1.57289    0.12519 -12.564 < 0.0000000000000002 ***
## (Intercept):2 -0.38604    0.11938  -3.234     0.001222 **
## InflLow        1.28882    0.12670  10.172 < 0.0000000000000002 ***
## InflMedium     0.72242    0.12372   5.839     0.00000000524 ***
## TypeAtrium     -0.20616    0.13993  -1.473     0.140675
## TypeTerrace    0.51866    0.13358   3.883     0.000103 ***
## TypeTower     -0.57235    0.11875  -4.820     0.00000143628 ***
## ContLow        0.36028    0.09536   3.778     0.000158 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: logitlink(P[Y<=1]), logitlink(P[Y<=2])
##
## Residual deviance: 3479.149 on 3354 degrees of freedom
##
## Log-likelihood: -1739.575 on 3354 degrees of freedom
##
## Number of Fisher scoring iterations: 4
##
## No Hauck-Donner effect found in any of the estimates
##
## Exponentiated coefficients:
##      InflLow InflMedium TypeAtrium TypeTerrace TypeTower ContLow
## 3.6284964 2.0594206 0.8137002 1.6797833 0.5641981 1.4337373
```

#### iv) Selección del modelo e interpretación de resultados

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

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

```
#Agruparemos los datos desagrupados
Datos$SatTypeInflCont=factor(paste(Datos$Sat, Datos$TypeInflCont, sep="."))
by_SatTypeInflCont <- Datos %>% group_by(SatTypeInflCont)
by_SatTypeInflCont<-by_SatTypeInflCont %>% tally()
Dat <- data.frame(do.call('rbind', strsplit(as.character(by_SatTypeInflCont$SatTypeInflCont), '.', fixed=
names(Dat) <- c("Sat", "Type", "Infl", "Cont")
Dat$SatTypeInflCont<-factor(paste(Dat$Sat, Dat$Type, Dat$Infl, Dat$Cont, sep="."))
DatosAg<-merge(Dat, by_SatTypeInflCont, by = "SatTypeInflCont", all = TRUE)

#Obtenemos las combinaciones, aprovechando la tabla agregada
combinaciones <- unique(DatosAg[,3:5]) %>%
  arrange(Type, Infl, Cont)

#También podríamos obtenerla con los datos desagregados como:
#combinaciones <- unique(datos[,1:2]) %>% arrange(Sexo, Edad)

#Con esto aplicamos la función predict, con tipo "response" y tenemos:
probas <- predict(fit_ord_par, combinaciones, type = "response")

#Finalmente, unimos las combinaciones con sus probabilidades
datos_modelo <- data.frame(cbind(combinaciones, probas))

#Obtenemos las combinaciones, con la tabla agregada

DatosAg_Apartment<-DatosAg%>%filter(Type %in% "Apartment") #Filtramos apartment
combinaciones <- unique(DatosAg_Apartment[,3:5]) %>%
  arrange(Type, Infl, Cont)

#También podríamos obtenerla con los datos desagregados como:
#combinaciones <- unique(datos[,1:2]) %>% arrange(Sexo, Edad)

#Con esto aplicamos la función predict, con tipo "response" y tenemos:
probas <- predict(fit_ord_par, combinaciones, type = "response")

#Finalmente, unimos las combinaciones con sus probabilidades
datos_modelo <- data.frame(cbind(combinaciones, probas))
```

```

#Le cambiamos el nombre, para más adelante (en caso de que no aparezcan los nombres)
colnames(datos_modelo)<-c("Type", "Infl", "Cont", "Low", "Medium", "High")

#En este caso, necesitamos los datos en tipo long
data_long <- datos_modelo %>%
  pivot_longer(cols = c(`Low`, Medium, `High`),
    names_to = "Sat",
    values_to = "Probabilidad") %>%
  #Aplicamos orden al tipo factor
  mutate("Respuesta"=factor(Sat, levels = c("Low", "Medium", "High")), "Infl"=factor(Infl, levels = c("Low", "Medium", "High")))

#Con esto podemos crear la gráfica
ggplot(data_long, aes(x = Infl, y = Probabilidad, fill = Sat)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.8),
    width = 0.7) +
  geom_text(aes(label = round(Probabilidad, 2)),
    position = position_dodge(width = 0.8),
    vjust = -0.7, size = 3) + # Añadir etiquetas con probabilidades
  facet_grid(. ~ Cont) + #Este nos permite separar por Cont
  labs(
    title = "Probabilidades de Satisfacción (Sat) por Contacto (Cont) e Influencia (Infl)",
    x = "Influencia (Infl) en mantenimiento del Apartment",
    y = "Probabilidad"
  ) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "top"
  ) + theme_bw()

```



Probabilidades de Satisfacción (Sat) por Contacto (Cont) e Influencia

