

# Práctica 2 - Modelos Lineales 2

Hernandez Bello Diana Patricia '20183808', Manrique Urbina Justo Andres '20091107',  
Moreano Roldan Juan Pablo '20184093', Parillo Apaza Jorge Hernan '19947810', Urbano  
Burgos Alejandrina Margarita '20047278'

10/26/2019

## Pregunta 1

Sea  $Y$  una variable aleatoria discreta con distribución binomial negativa  $\mu$  y parámetro de dispersión  $\phi$ , cuya función de distribución es dada por

$$f(y) = \frac{\Gamma(y + \phi)}{\Gamma(y + 1)\Gamma(\phi)} \left(\frac{\mu}{\mu + \phi}\right)^y \left(\frac{\phi}{\mu + \phi}\right)^\phi, y = 0, 1, 2 \dots$$

Demuestre que pertenece a la familia exponencial, para  $\phi$  conocido.

**Solución:** La función de probabilidad de  $Y$  se puede reexpresar como:

$$f(y) = \frac{\Gamma(y + \phi)}{\Gamma(y + 1)\Gamma(\phi)} (1 - p)^y + p^\phi.$$

Posteriormente, la función se puede expresar de la siguiente forma:

$$f(y) = \frac{\Gamma(y + \phi)}{\Gamma(y + 1)\Gamma(\phi)} \exp(y \log(1 - p)\phi \log(p)).$$

En donde  $p = \frac{\phi}{\mu + \phi}$  y  $1 - p = \frac{\mu}{\mu + \phi}$

Se observa que pertenecería a la familia exponencial en tanto  $\phi$  es conocido, pues:

- $\theta = \log(1 - p)$
- $b(\theta) = \phi \log(p)$  o, asimismo,  $b(\theta) = \phi \log(1 - e^\theta)$
- $c(y, \phi)$  es igual a la función gamma.
- $\phi = 1$

Demuestre que para  $\phi$  conocido, la distribución de  $Y$  pertenece a la familia exponencial.

Encuentre la función de varianza y la función de enlace canónica.

**Solución:** La función de varianza se define como:

$$V(Y) = \frac{-1}{\phi} b''(\theta)$$

Resolviendo, se tiene que:

$$\frac{\partial^2}{\partial \theta^2} b(\theta) = \frac{\partial}{\partial \theta} \left( \frac{\phi e^\theta}{1 - e^\theta} \right) = \phi \frac{\partial}{\partial \theta} \left( \frac{e^\theta}{1 - e^\theta} \right) = \phi \frac{e^\theta}{(1 - e^\theta)^2} = \frac{\phi(1 - p)}{p^2}$$

Reemplazando  $p$  y  $1 - p$ , se tiene que:

$$\frac{\frac{\phi \mu}{\mu + \phi}}{\frac{\phi^2}{(\mu + \phi)^2}}$$

Por lo tanto, la varianza es:

$$\mu + \frac{\mu^2}{\phi}.$$

La función de enlace canónica se definiría de la siguiente forma:

$$\theta = \frac{\mu}{\phi + \mu}$$

## Pregunta 2

a) Demuestre la matriz de información de Fisher para

$$\beta = (\beta_0, \beta_1)^T$$

En general, la matriz de información de Fisher está dada por:

$$I(\theta) = \phi \sum_{i=1}^n w_i x_i x_j$$

Esto, de forma matricial, se escribe de:

$$\phi X^T w X$$

Considerando que, para Poisson, el siguiente enlace:

$$\eta_i = \log(\mu_i)$$

y varianza  $V(\mu) = \mu$ .

Asimismo, se tiene que:

$$w_i = \frac{(\frac{\partial}{\partial \eta_i})^2}{V_i}$$

$$w_i = \frac{(\frac{\partial \eta_i}{\partial})^{-2}}{V_i}$$

En dónde  $\frac{\partial \eta_i}{\partial \mu_i} = \frac{1}{\mu}$

Por lo tanto, reemplazando en la ecuación anterior, se tiene que:

$$w_i = \frac{(\frac{1}{\mu})^{-2}}{V_i}$$

$$w_i = \frac{\mu^2}{\mu} = \mu$$

Entonces se tiene que  $w_i = \mu_i$ . Por lo tanto, la matriz de información de Fisher es:

$$\phi X^T w X$$

en dónde se tiene lo siguiente:

$$\phi = 1$$

$$w = diag(\mu_1, \mu_2, \dots, \mu_n)$$

$$X = \begin{pmatrix} 1 & X_1 - \bar{X} \\ 1 & X_2 - \bar{X} \\ \vdots & \vdots \\ 1 & X_n - \bar{X} \end{pmatrix}$$

b) Encuentre una expresión de varianza de  $\hat{\beta}_0 - \hat{\beta}_1$

Se sabe que  $Var(\hat{\beta}) = (X^T w X)^{-1}$  para Poisson.

Para hallar  $(X^T w X)$  (matriz de información de Fisher) se aplica:

$$H(\beta) = -\frac{\partial^2 l(\beta_0, \beta_1)}{\partial \beta \partial \beta'} = -\sum_{i=1}^n \frac{\partial^2 l_i(\beta_0, \beta_1)}{\partial \beta \partial \beta'}$$

Para el cálculo de  $-\frac{\partial^2 l(\beta_0, \beta_1)}{\partial \beta \partial \beta'}$  se consideran las siguientes notaciones :

- $L_i(\beta_0, \beta_1) = u_i^{y_i} (\exp(-\mu_i))$   
como contribución de la observación i
- $L_i(\beta_0, \beta_1) = \prod_{i=1}^n L_i(\beta_0, \beta_1)$   
como función de verosimilitud
- $l_i(\beta_0, \beta_1) = y_i \log(\mu_i) - u_i = y_i(\beta_0 + \beta_1 X_i) - \exp(\beta_0 + \beta_1 X_i)$   
como logaritmo de la primera expresión

Luego

$$\frac{\partial^2 l_i(\beta_0, \beta_1)}{\partial \beta_0^2} = -\exp(\beta_0 + \beta_1 X_1) = -\mu_i$$

$$\frac{\partial^2 l_i(\beta_0, \beta_1)}{\partial \beta_1^2} = -\exp(\beta_0 + \beta_1 X_i) X_i^2 = -\mu_i X_i^2$$

$$\frac{\partial^2 l_i(\beta_0, \beta_1)}{\partial \beta_0 \partial \beta_1} = -\exp(\beta_0 + \beta_1 X_i) X_i = -\mu_i X_i$$

De lo anterior se tiene que  $H(\beta)$

$$H(\beta) = \begin{pmatrix} \sum_{i=1}^n \mu_i & \sum_{i=1}^n \mu_i x_i \\ \sum_{i=1}^n \mu_i x_i & \sum_{i=1}^n \mu_i x_i^2 \end{pmatrix}$$

Dado que  $H(\beta) = X^T w X$  y  $Var(\hat{\beta}) = (X^T w X)^{-1}$  se tiene que :

$$Var(\hat{\beta}) = (H(\beta))^{-1} = \frac{1}{(\sum_i^n \mu_i)(\sum_i^n \mu_i x_i^2) - (\sum_i^n \mu_i X_i)^2} \begin{pmatrix} \sum_i^n \mu_i x_i^2 & -\sum_i^n \mu_i x_i \\ -\sum_i^n \mu_i x_i & \sum_i^n \mu_i \end{pmatrix}$$

Donde :

$$(\sum_{i=1}^n \mu_i)(\sum_{i=1}^n \mu_i x_i^2) - (\sum_{i=1}^n \mu_i X_i)^2 = Z$$

Se solicita :

$$Var(\hat{\beta}_0 - \hat{\beta}_1) = Var(\hat{\beta}_0) + Var(\hat{\beta}_1) - 2Cov(\hat{\beta}_0, \hat{\beta}_1)$$

$$Z\left(\sum_{i=1}^n \mu_i X_i^2\right) + Z\left(\sum_{i=1}^n \mu_i\right) + 2 \sum_{i=1}^n \mu_i x_i$$

### Pregunta 3

```
datos <- read.csv("~/Documents/maestria-pucp/2019-2/modelos-lineales-2/clase-7/Preg3.csv", sep=",", file=
```

```
library(glm2)
library(faraway)
library(car)
library(spm)
library(MASS)
library(hnp)
library(ggplot2)
```

En el archivo Preg3.csv se presentan los siguientes variables medidas durante un año en una región :

- reclamos : números de reclamos en un seguro de autos para responsabilidad civil frente a terceros.
- accidentes : números de accidentes en la región
- poblacion : población en la región

a) Estime un modelo de regresión de Poisson para explicar tasa de reclamos por habitante de la región considerando como covariables el logaritmo del número de accidentes. Presente formalmente el modelo e interprete los coeficientes estimados.

$$Y_i \sim Poisson(\mu_i)$$

$$n_i = \beta_0 + \beta_1 x_i$$

$$\log(u_i) = n_i$$

$$u_i = t_i * \lambda_i$$

- $t_i$  = habitante de la región
- $\lambda_i$  = tasa de reclamos por habitante de la región

Donde :

- $Y_i$  = número de reclamos por habitante de la región
- $x_i$  = logaritmo del número de accidentes en la región

Análisis previo:

```
attach(datos)
head(datos)

##   reclamos accidentes poblacion
## 1      1103       2304    124850
## 2      1939       2660    143500
## 3      4339       7381    470700
## 4      1491       3217    311300
## 5      3801       6655    584900
## 6       387       2013    106350
```

```

dim(datos)

## [1] 176   3

Para obtener reclamos por habitante de la región

y<-reclamos/poblacion
x<-log(accidentes)

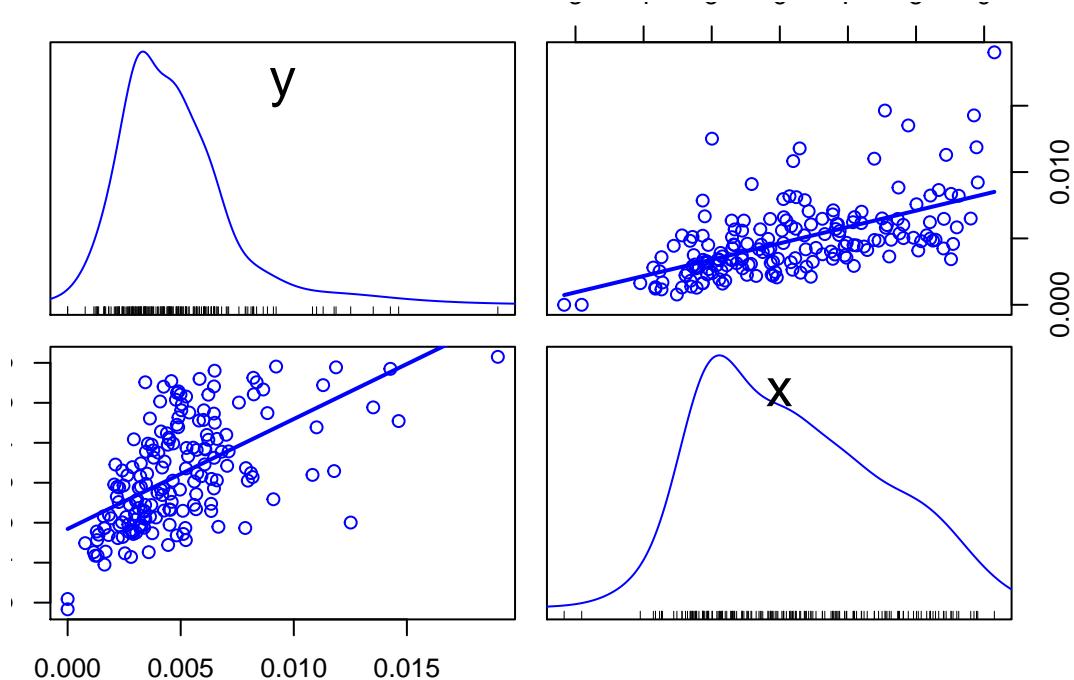
nueva_data<-as.data.frame(cbind(y,x))
head(nueva_data)

##          y      x
## 1 0.008834602 7.742402
## 2 0.013512195 7.886081
## 3 0.009218186 8.906664
## 4 0.004789592 8.076205
## 5 0.006498547 8.803124
## 6 0.003638928 7.607381

```

Analizando la data mediante el gráfico de dispersión

```
scatterplotMatrix(nueva_data,smooth = FALSE)
```



Se observa que la varianza no es constante, a medida que aumenta  $X$  se observa que aumenta la varianza

A continuación presentamos el primer modelo propuesto :

```

Modelo1 <- glm(y ~ x, data=nueva_data, family=poisson(link = "log"))

summary(Modelo1)

##
## Call:
## glm(formula = y ~ x, family = poisson(link = "log"), data = nueva_data)
##
```

```

## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.065468 -0.024434 -0.005905  0.012462  0.119165
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.9083    5.3907 -1.282   0.200
## x           0.2468    0.7967  0.310   0.757
##
## (Dispersion parameter for poisson family taken to be 1)
##
## Null deviance: 0.26056 on 175 degrees of freedom
## Residual deviance: 0.16506 on 174 degrees of freedom
## AIC: Inf
##
## Number of Fisher Scoring iterations: 8
exp(coef(Modelo1))

```

```

## (Intercept)      x
## 0.0009994421 1.2798795770

```

### Interpretación

$$\beta_0 \text{ estimado} = -6.9083$$

$$\beta_1 \text{ estimado} = 0.2468$$

$$\beta_0 \text{ estimado} = \exp(-6.9083) \cong 0$$

Se espera que el número de reclamos por habitante de la región sea aproximadamente cero cuando el logaritmo del número de accidentes de la región es cero, es decir cuando el número de accidentes de la región es uno.

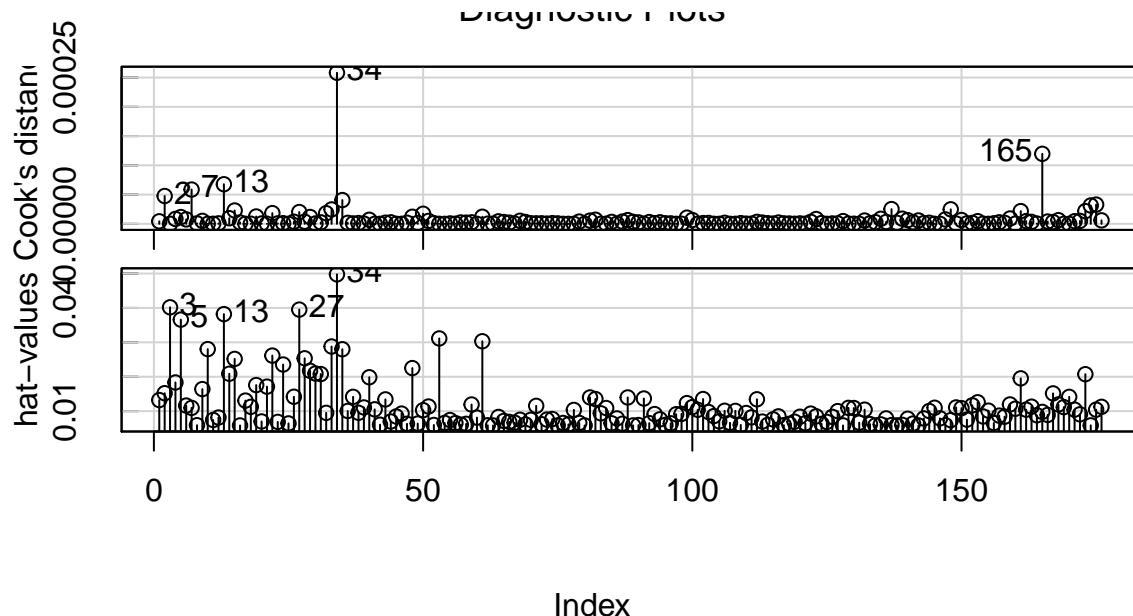
$$\beta_1 \text{ estimado} = \exp(0.2468) \cong 1.28$$

Cuando ocurre un incremento de uno en el logaritmo del número de accidentes de la región, se espera que esto genere un incremento de 29% en el número de reclamos por habitante de la región.

**b)** Realice gráficos de leverage, distancia de Cook, residuos versus valores ajustados, residuos con bandas de confianza. Comente sus resultados.

### Gráfico de Leverage y distancia de Cook

```
influenceIndexPlot(Modelo1, vars=c ("Cook", "hat"), id=list(n=5))
```



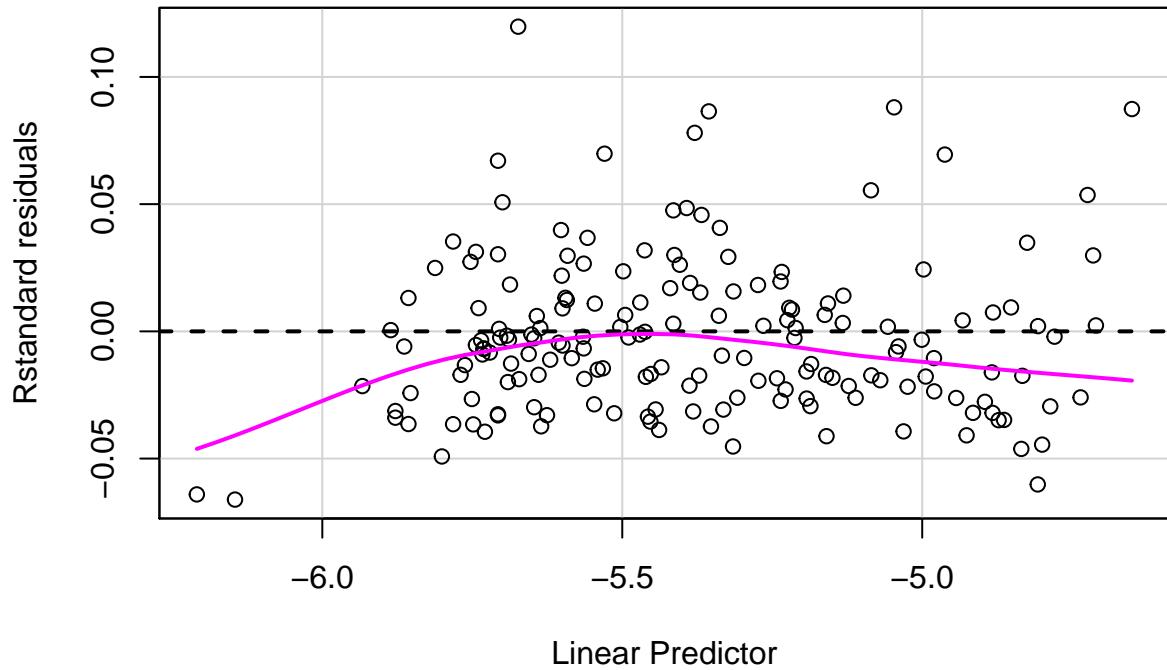
**Comentario** En el gráfico anterior de distancias de “Cook D” nos ayuda a identificar puntos que son potencialmente influyentes debido a su ubicación en el rango de los datos.

También hay que notar el segundo gráfico donde se muestran los atípicos en las covariables x, h(hatvalues). A esto se le conoce como leverage o apalancamiento.

Se recomienda retirar los puntos con alta distancia de cook y alto leverage effect(atípico en la covariable) como por ejemplo el punto 34.

#### Gráfico de residuos versus valores ajustados

```
residualPlot(Modelo1, type="rstandard")
```



El modelo de Poisson es heterocedástico

Función de varianza:

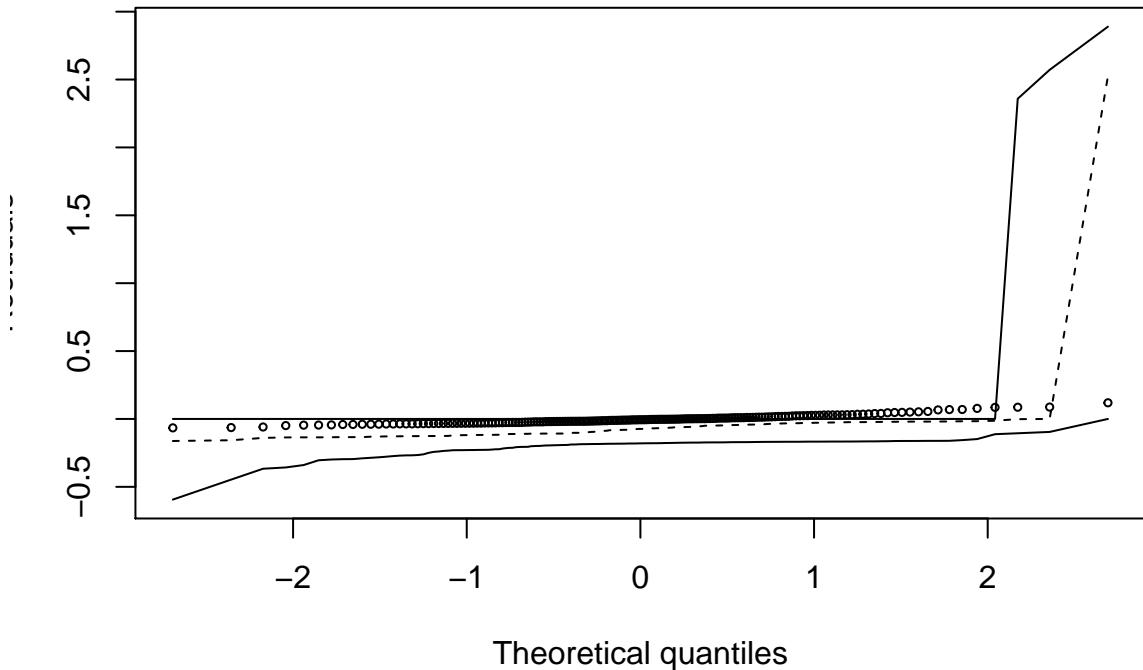
$$V(\text{??}) = \text{??}$$

No se muestra patrones, es casi casi un ruido blanco, pero parece que hubiera mayor dispersión que en un modelo Poisson.

#### Gráfico de residuos con bandas de confianza

```
library(hnp)
hnp(Modelo1, halfnormal = FALSE)

## Poisson model
```



### Comentario

Vemos que hay varios puntos que se quedan fuera de la banda, lo que indica que este modelo no ajusta bien a los datos.

Este gráfico de bandas y el anterior de residuos, nos hace pensar que necesitamos un modelo que contemple mayor varianza para que se ajuste mejor a los datos, como el modelo Binomial Negativa.

c) En base a sus resultados en b) de ser necesario proponga un nuevo modelo y realice un análisis de diagnóstico que incluya el estudio del efecto de posibles observaciones infuyentes. Indique cuál sería el modelo adecuado para este problema.

### Modelo 2 : Binomial Negativa

$$Y_i \sim BN(ui, \phi)$$

$$n_i = \beta_0 + \beta_1 x_i$$

$$\log(u_i) = n_i$$

$$E(Y_i) = u_i$$

$$Var(Y_i) = u_i + \frac{u_i^2}{\phi}$$

, tiene por propiedad más varianza que el Modelo de Poisson.

Donde :

- $Y_i$  : número de reclamos por habitante de la región
- $X_i$  : logaritmo del número de accidentes en la región

Además, se retira del modelo el caso 34, por lo visto en los gráficos de leverage y distancias de Cook.

```
Modelo2=glm.nb(y ~ x, data=nueva_data, subset=-34)
```

Coeficientes estimados

```

summary(Modelo2)

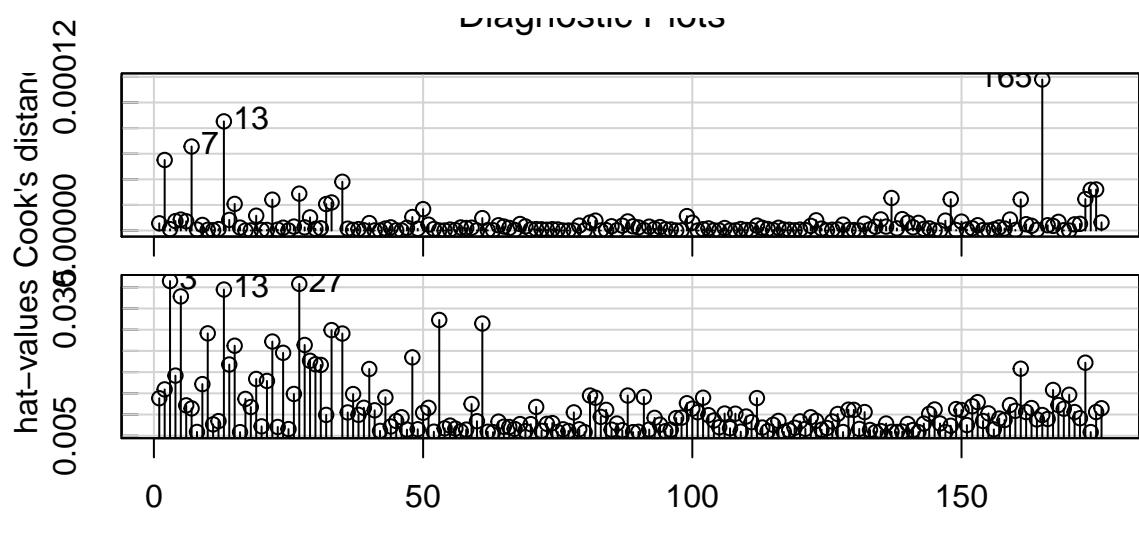
##
## Call:
## glm.nb(formula = y ~ x, data = nueva_data, subset = -34, init.theta = 70769.2186,
##         link = log)
##
## Deviance Residuals:
##      Min      1Q  Median      3Q     Max
## -0.2567 -0.1992 -0.1643  0.2433  0.3630
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.8135    5.4785 -1.244   0.214
## x            0.2307    0.8169  0.282   0.778
##
## (Dispersion parameter for Negative Binomial(70769.22) family taken to be 1)
##
## Null deviance: 8.9247 on 174 degrees of freedom
## Residual deviance: 8.8453 on 173 degrees of freedom
## AIC: 15.56
##
## Number of Fisher Scoring iterations: 1
##
##
## Theta: 70769
## Std. Err.: 11219953
## Warning while fitting theta: iteration limit reached
##
## 2 x log-likelihood: -9.56
exp(coef(Modelo2))

## (Intercept)      x
## 0.001098867 1.259423018

```

Gráfico de leverage y distancia de Cook

```
influenceIndexPlot(Modelo2, vars=c ("Cook", "hat"), id=list(n=3))
```



Index

Del gráfico anterior , se recomienda retirar los puntos con alta distancia de cook y alto leverage effect como por ejemplo el punto 13.

Gráfico de residuos versus valores ajustados

```
residualPlot(Modelo2,type="rstandard")
```

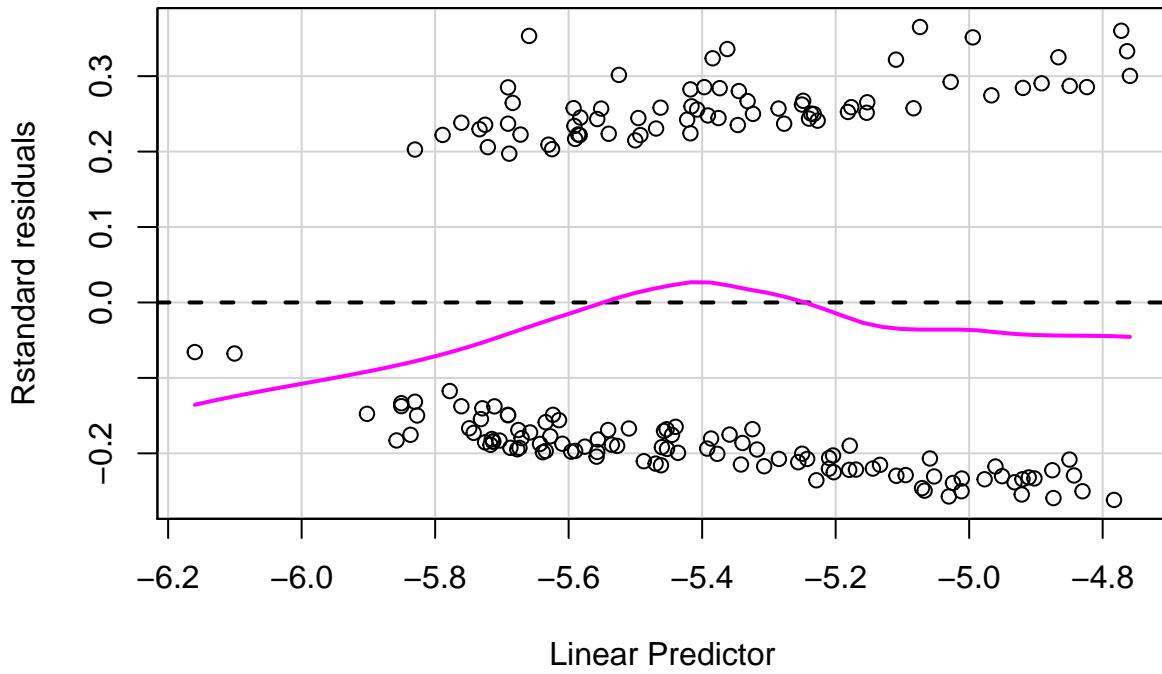
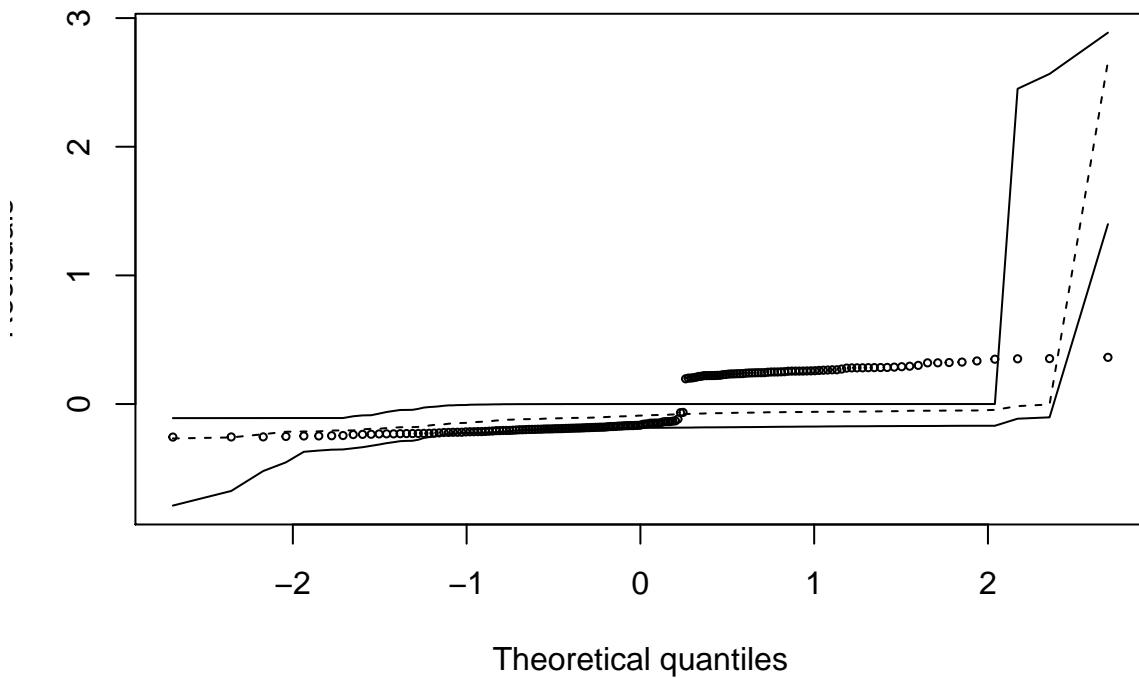


Gráfico de residuos con bandas de confianza

```
hnp(Modelo2,halfnormal = FALSE)
```

```
## Negative binomial model (using MASS package)
```



**Comentario** Vemos que los puntos están dentro de las bandas hasta un punto donde se salen de las bandas y forman una región fuera de ellas, lo que indica que este modelo tampoco ajusta bien a todos los datos.

El gráfico de bandas de este modelo y del anterior nos hace pensar que faltan considerar más covariables para encontrar un modelo que ajuste mejor.

d) Si en una región hubiera un aumento del 10% en el número de accidentes, calcule en forma puntual y por intervalo el efecto en la tasa de reclamos por habitante.

### Modelo 3 : Binomial Negativa

```
Modelo3=glm.nb(y ~ x, data=nueva_data, subset=-c(34,13))
```

Coefficientes estimados

```
summary(Modelo3)
```

```
##
## Call:
## glm.nb(formula = y ~ x, data = nueva_data, subset = -c(34, 13),
##         init.theta = 71125.50582, link = log)
##
## Deviance Residuals:
##    Min      1Q   Median      3Q     Max
## -0.2565 -0.1984 -0.1651  0.2434  0.3634
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.7600    5.5458 -1.219   0.223
## x           0.2214    0.8321  0.266   0.790
##
## (Dispersion parameter for Negative Binomial(71125.51) family taken to be 1)
##
## Null deviance: 8.7910  on 173  degrees of freedom
## Residual deviance: 8.7206  on 172  degrees of freedom
```

```

## AIC: 15.423
##
## Number of Fisher Scoring iterations: 1
##
##
##           Theta:  71126
##           Std. Err.: 11404247
## Warning while fitting theta: iteration limit reached
##
## 2 x log-likelihood: -9.423
exp(coef(Modelo3))

## (Intercept)          x
## 0.001159269 1.247813474

```

$$\beta_0 \text{ estimado} = -6.7600$$

$$\beta_1 \text{ estimado} = 0.2214$$

$$\beta_0 \text{ estimado} = \exp(-6.7600) \cong 0$$

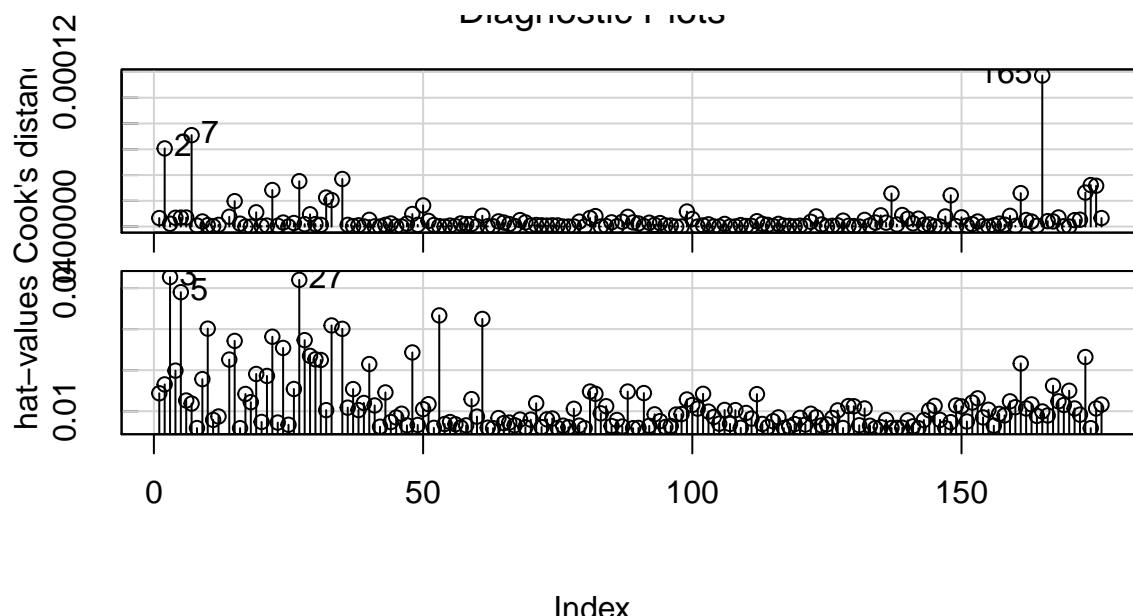
Se espera que el número de reclamos por habitante de la región sea aproximadamente cero cuando el logaritmo del número de accidentes de la región es cero, es decir cuando el número de accidentes de la región es uno.

$$\beta_1 \text{ estimado} = \exp(0.2214) \cong 1.25$$

Cuando ocurre un incremento de uno en el logaritmo del número de accidentes de la región, se espera que esto genere un incremento de 25% en el número de reclamos por habitante de la región.

#### Gráfico de Leverage y distancia de Cook

```
influenceIndexPlot(Modelo3, vars=c ("Cook", "hat"), id=list(n=3))
```



**Comentario** Se recomienda retirar los puntos con alta distancia de cook y alto leverage effect(atípico en la covariante) , en este modelo no tenemos puntos que cumplan ambas condiciones.

Gráfico de residuos versus valores ajustados

```
residualPlot(Modelo3,type="rstandard")
```

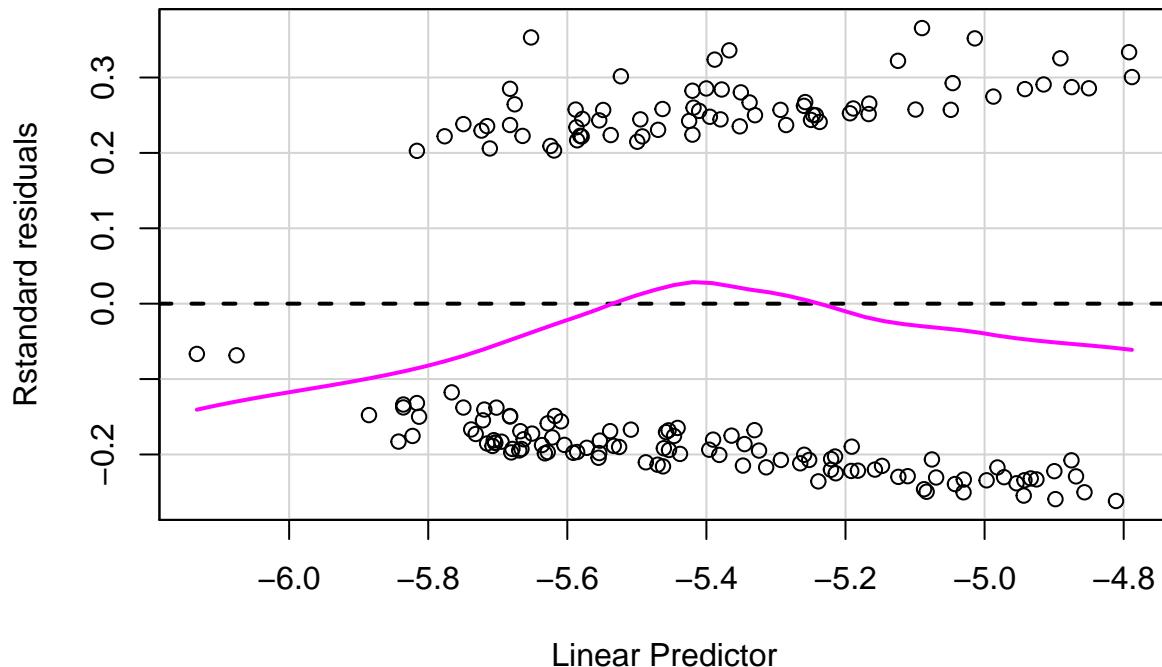
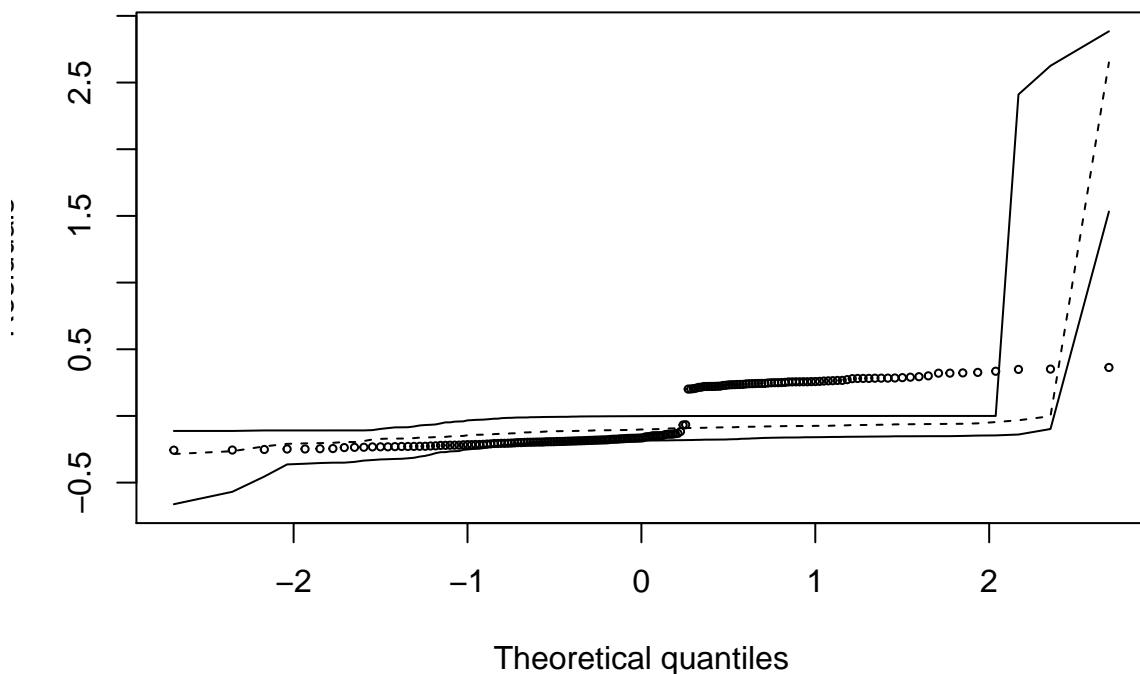


Gráfico de residuos con bandas de confianza

```
hnp(Modelo3,halfnormal = FALSE)
```

```
## Negative binomial model (using MASS package)
```



```
AIC(Modelo1)
```

```

## [1] Inf
AIC(Modelo2)

## [1] 15.55971
AIC(Modelo3)

## [1] 15.42255

```

Por el criterio AIC el mejor modelo es el **Modelo 3**

Pero el gráfico de bandas de residuos con bandas de confianza nos dice que ninguno de estos 3 modelos ajusta bien a los datos, sugerimos adicionar otras covariables que ayuden a explicar mejor.

## Pregunta 4

La base de datos utilizada para el presente informe contiene las siguientes variables:

- **Variable respuesta**
  - nsiniestros: Cantidad de siniestros ocurridos.
- **Covariables**
  - Asegurados\_Total: Cantidad de asegurados en cada una de las pólizas.
  - Planilla\_total: Monto mensual de salario pagado a los empleados en unidades monetarias dentro de cada póliza.
  - nivel\_riesgo: Esta variable fue construida clasificando las actividades de riesgo del 1 al 5, donde 5 significa que la actividad económica que desarrolla la empresa tiene mayor exposición al riesgo de accidente o enfermedad profesional y 1 que la exposición a estos riesgos es menor.

La presente base de datos contiene 14,064 observaciones. Estas observaciones fueron recabadas por un período de 3 años.

### Objetivo

El objetivo de este análisis es modelar el número de reclamos de una póliza de seguro de vida contratada para un grupo asegurado (que consiste en todos los empleados de una empresa), mediante las variables explicativas: número de asegurados, planilla mensual (salario mensual del grupo) y nivel de riesgo de la actividad económica de la empresa.

Los datos considerados en el análisis contemplan las Pólizas vigentes a corte dic/2017 de un producto de Seguro de Vida de una Compañía de Seguros de Perú. En cada una de estas pólizas al menos uno de los miembros del grupo asegurado ha presentado un reclamo de invalidez o los familiares de los miembros del grupo han presentado por lo menos un reclamo de fallecimiento en un periodo de tres años.

El estudio se compone de un análisis exploratorio de los datos, selección del mejor modelo, análisis de diagnóstico del mismo e interpretación de resultados.

### Análisis Exploratorio

Para realizar el análisis exploratorio inicial, realizaremos la carga de los datos en el siguiente código:

```

library(dplyr)
library(car)
library(ggplot2)
library(car)
library(GGally)
library(stargazer)
library(hnp)
setwd("/home/justomanrique/Documents/maestria-pucp/2019-2/modelos-lineales-2/clase-7/")
datos_preg4 <- readxl::read_excel("pregunta4_diana_v2.xlsx")

```

```

datos_preg4 = datos_preg4 %>%
  mutate ( ACTIVIDAD=as.factor(ACTIVIDAD),
          nivel_riesgo=as.factor(nivel_riesgo))

```

Realizamos un gráfico de dispersión mediante la función ggpairs, para identificar posibles relaciones entre los datos así como la distribución de las mismas:

```

nombres = c("ACTIVIDAD", "Asegurados_total", "Planilla_total", "nsiniestros", "nivel_riesgo")
datos_preg4 = subset ( datos_preg4 , select = nombres )
datos_preg4 = na.omit ( datos_preg4 )

```

```

datos_preg4 = datos_preg4 %>%
  mutate ( ACTIVIDAD=as.factor(ACTIVIDAD),
          nivel_riesgo=as.factor(nivel_riesgo))

datos_preg4 = datos_preg4[,2:5]

summary(datos_preg4)

```

	Asegurados_total	Planilla_total	nsiniestros	nivel_riesgo
## Min.	1.00	Min. : 2	Min. : 0.00000	1:1063
## 1st Qu.	6.00	1st Qu.: 8075	1st Qu.: 0.00000	2:1736
## Median	11.00	Median : 14300	Median : 0.00000	3:1183
## Mean	37.26	Mean : 76161	Mean : 0.01465	4:5705
## 3rd Qu.	25.00	3rd Qu.: 36496	3rd Qu.: 0.00000	5:4377
## Max.	:13401.00	Max. :50044818	Max. :24.00000	

En base a ello se observa:

- La cartera de pólizas en el presente informe es riesgosa, pues existe mayor proporción de pólizas con riesgo 4 y 5 que de pólizas con riesgo 1 a 3.
- En las variables Asegurados\_total, Planilla\_total, nsiniestros existen valores extremos pues la gráfica se encuentra distorsionada, con alta concentración de valores en un lado y una cola larga hacia la derecha.
- Se observa correlación fuerte entre la cantidad total de asegurados y planilla (correlación del 0.736).
- Se observa correlación media entre la cantidad de siniestros y la planilla total (correlación del 0.513).

## Selección de modelos

```

modelo1 <- glm(nsiniestros ~ log(Asegurados_total) + log(Planilla_total) + nivel_riesgo,
data=datos_preg4, family=poisson(link = "log"))

summary(modelo1)

##
## Call:
## glm(formula = nsiniestros ~ log(Asegurados_total) + log(Planilla_total) +
##       nivel_riesgo, family = poisson(link = "log"), data = datos_preg4)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.9422   -0.1022   -0.0617   -0.0429   13.1714
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -17.78446   0.81749 -21.755 < 2e-16 ***

```

```

## log(Asegurados_total)    0.07980   0.11724   0.681   0.49608
## log(Planilla_total)     1.04650   0.09911   10.559   < 2e-16 ***
## nivel riesgo2            0.65363   0.40554   1.612   0.10702
## nivel riesgo3            1.42285   0.43321   3.284   0.00102 **
## nivel riesgo4            1.30232   0.31807   4.094   4.23e-05 ***
## nivel riesgo5            1.71826   0.32990   5.208   1.90e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
## Null deviance: 2267.5  on 14063  degrees of freedom
## Residual deviance: 1163.7  on 14057  degrees of freedom
## AIC: 1413.9
##
## Number of Fisher Scoring iterations: 8
round(exp(coef(modelo1)),3)

##           (Intercept) log(Asegurados_total)  log(Planilla_total)
##               0.000          1.083          2.848
##       nivel riesgo2      nivel riesgo3      nivel riesgo4
##               1.923          4.149          3.678
##       nivel riesgo5
##               5.575

```

Del modelo inicial, se puede observar que:

- En la medida que incremente el nivel de riesgo (tomando como referencia el nivel 1), se espera un incremento en la cantidad de siniestros. El efecto se hace mayor, en la medida que el nivel de riesgo incrementa (ver tabla de coeficientes).
- En la medida que incremente la cantidad de asegurados, se espera que incremente el 8% la cantidad de siniestros.
- En la medida que aumente la planilla total, se espera que se incremente en 180% la cantidad de siniestros.

## Análisis de diagnóstico

La interpretación anteriormente indicada se sostiene en la medida que los supuestos del modelo se cumplan. Para ello, se realizó el siguiente análisis de diagnóstico:

### Residuales

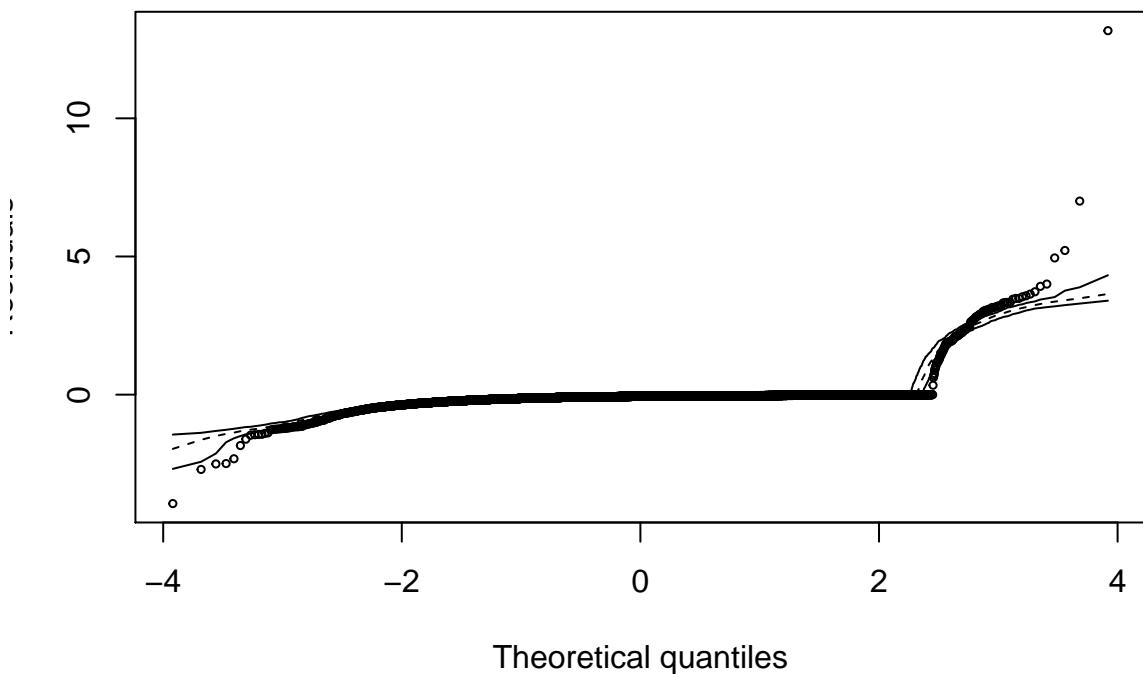
Ver a continuación el diagnóstico de residuales:

```

#### Gráfico de residuos con bandas de confianza
hnp(modelo1,halfnormal = FALSE)

```

```
## Poisson model
```

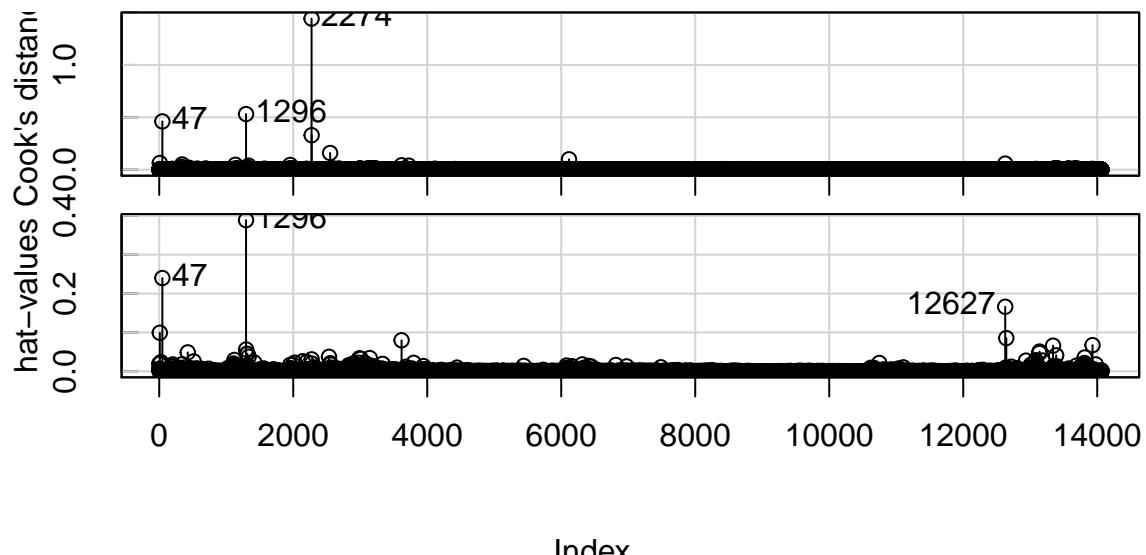


Se observa, en el diagnóstico de residuales, que los residuos se ajustan adecuadamente a las bandas de confianza. Sin embargo, en las colas existe mayor dispersión. Esto podría deberse a los valores atípicos que existen dentro de la muestra (esto se observó en el análisis exploratorio).

#### Visualización de puntos influyentes

```
### Gráfico de leverage y distancia de Cook
influenceIndexPlot(modelo1, vars=c ("Cook", "hat"), id=list(n=3))
```

Diagnóstico 1000



Se observa que las observaciones 1296, 457 y 2274 son valores influyentes en el modelo de acuerdo a la distancia de Cook. Asimismo, en relación a los hat-values, las observaciones 47, 1296 y 12627 son valores influyentes.

## Modelo final

En base al trabajo anterior, se eliminaron los valores influyentes en común para evaluar el modelo final. Ver a continuación

```
# Modelo 2: Poisson Regression y ~ x -c(1296,47)
modelo2 <- glm(nsiniestros ~ log(Asegurados_total) + log(Planilla_total) + nivel_riesgo, data=datos_preg4)

summary(modelo2)

##
## Call:
## glm(formula = nsiniestros ~ log(Asegurados_total) + log(Planilla_total) +
##      nivel_riesgo, family = poisson(link = "log"), data = datos_preg4,
##      subset = -c(1296, 47))
##
## Deviance Residuals:
##       Min      1Q   Median      3Q     Max
## -2.9718 -0.1027 -0.0620 -0.0425 13.0147
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -16.5470    0.9316 -17.761 < 2e-16 ***
## log(Asegurados_total) 0.2977    0.1307   2.279  0.02268 *
## log(Planilla_total)  0.8625    0.1132   7.620 2.53e-14 ***
## nivel_riesgo2  0.6561    0.4062   1.615  0.10628
## nivel_riesgo3  1.4274    0.4354   3.278  0.00105 **
## nivel_riesgo4  1.3058    0.3234   4.038  5.39e-05 ***
## nivel_riesgo5  1.7427    0.3315   5.257 1.47e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
## Null deviance: 1972.2 on 14061 degrees of freedom
## Residual deviance: 1142.3 on 14055 degrees of freedom
## AIC: 1387.4
##
## Number of Fisher Scoring iterations: 8
#ha cambiado la estimacion de los parametros beta y el log(asegurados_total se ha vuelto significativo)

round(exp(coef(modelo2)),3)

##             (Intercept) log(Asegurados_total) log(Planilla_total)
##                 0.000                  1.347                  2.369
## nivel_riesgo2          1.927                  4.168                  3.691
## nivel_riesgo5          5.713
```

Se observa que la estimación de los parámetros ha variado considerablemente para la cantidad total de asegurados (1.347 vs. 1.083), así como para la planilla (2.369 vs. 2.848). Asimismo, se observa que la cantidad total de asegurados se ha vuelto una variable significativa.

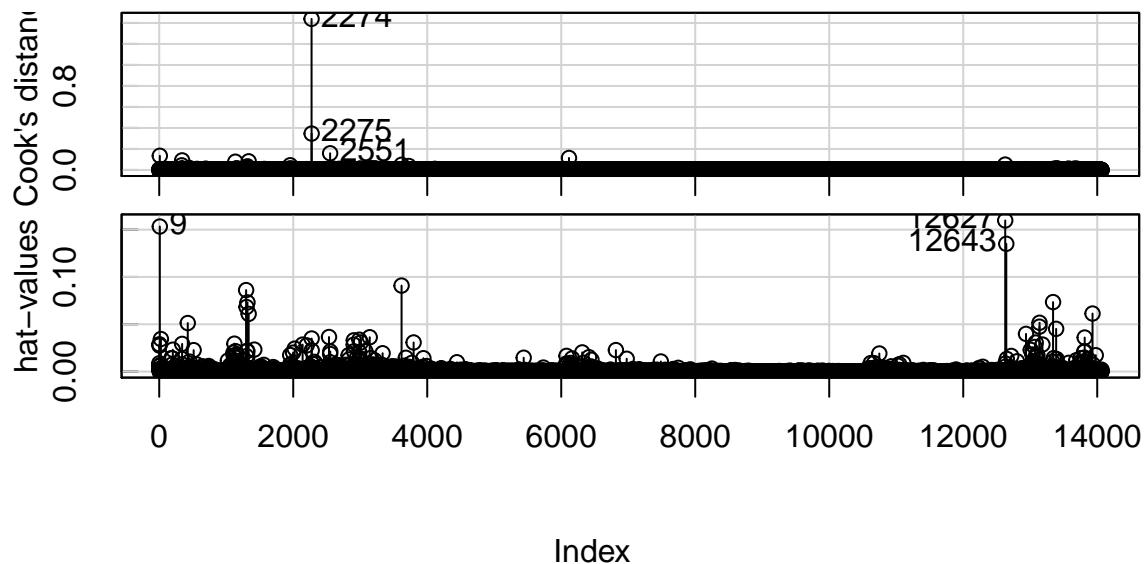
Asimismo, se han hecho los siguientes diagnósticos.

```
#####
```

```
## Grafico de leverage y distancia de Cook
```

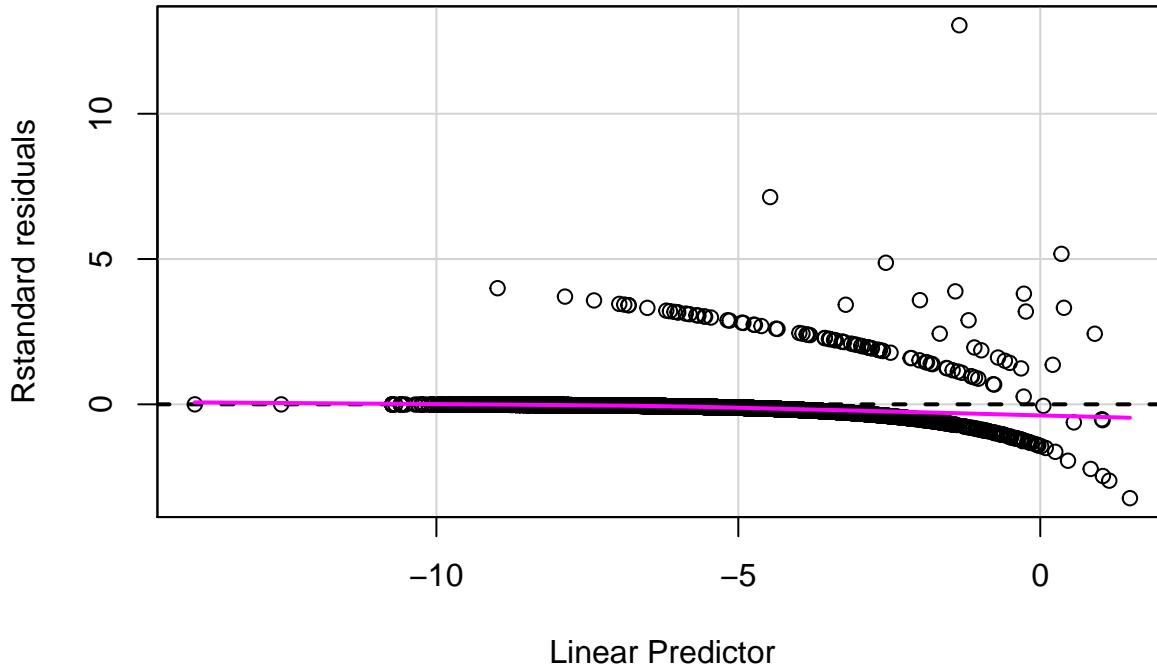
```
influenceIndexPlot(modelo2, vars=c ("Cook", "hat"), id=list(n=3))
```

Diagnostic plots



```
## Grafico de residuos versus valores ajustados
```

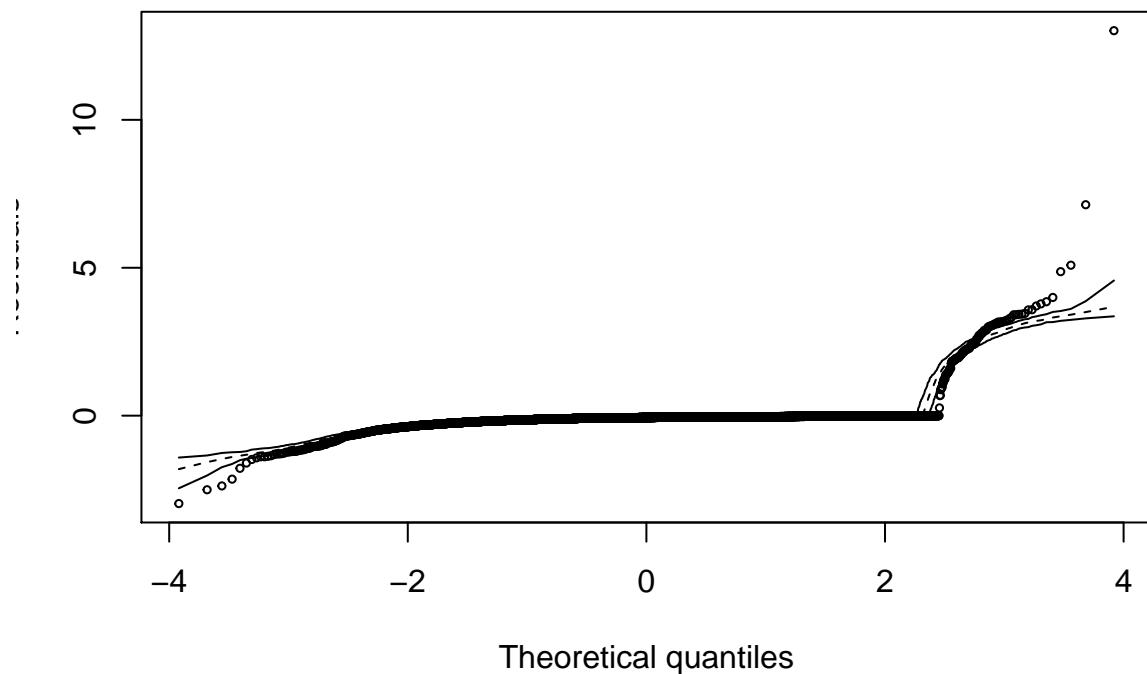
```
residualPlot(modelo2, type="rstandard")
```



```
## Grafico de residuos con bandas de confianza
```

```
hnp(modelo2, halfnormal = FALSE)
```

```
## Poisson model
```



En relación a los gráficos presentados, se visualiza lo siguiente:

- Se observa que los residuales se encuentran más cerca a las bandas de confianza de los residuales, sin embargo aún persisten ciertos valores influyentes.
- Se observa que aún siguen persistiendo valores influyentes.