

Práctica 2 - Modelos Lineales 2

Juan Pablo Moreano,

10/26/2019

Pregunta 1

Sea Y una variable aleatoria discreta con distribución binomial negativa μ y parámetro de dispersión ϕ , 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 ϕ 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 ϕ 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 ϕ 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

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:

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

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

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:

$$\phi = 1$$

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

Pregunta 4

La base de datos utilizada para el presente informe consiste en <>. Esta contiene las siguientes variables:

- **Variable respuesta**
 - nsiniestros: Cantidad de siniestros ocurridos.
- **Covariables**
 - Asegurados_Total: Cantidad de asegurados para la presente póliza.

- Planilla_total: <>
- nivel_riesgo: <>

La presente base de datos contiene 14,064 observaciones. Estas observaciones fueron recabadas desde el año «» al año «».

El objetivo principal del estudio es modelar la cantidad de siniestros ocurridos en base a la cantidad total de asegurados, planilla total y nivel de riesgo. Para ello, 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("~/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

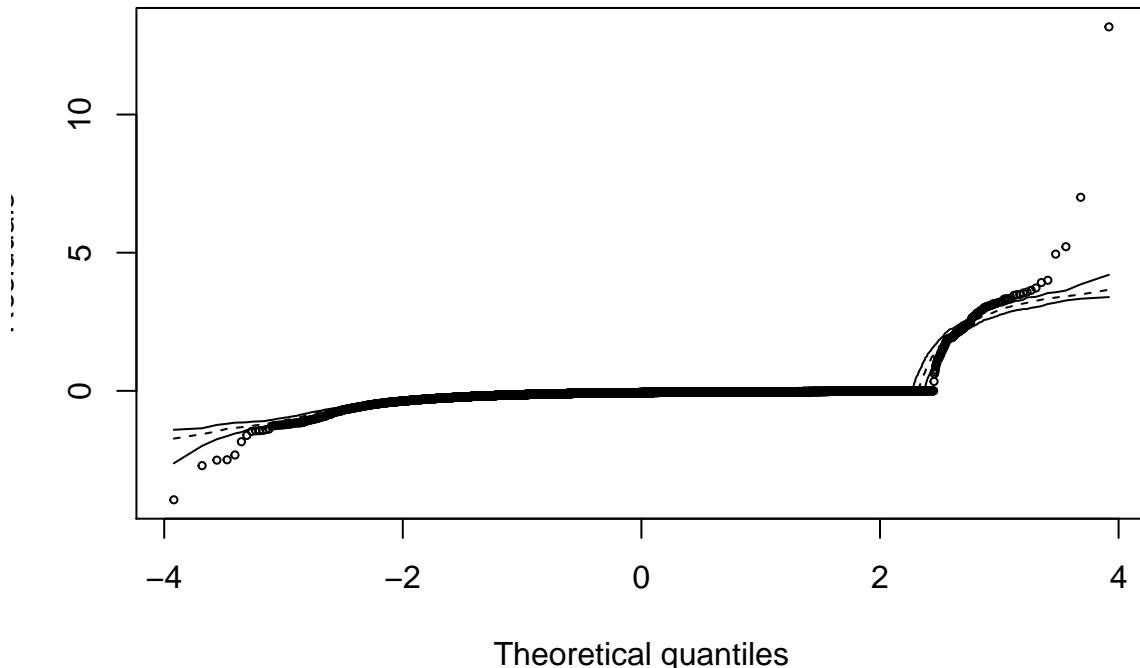
La interpretación anteriormente indicada se sostiene en la medida que los supuestos del modelo se cumplan. Para ello, se realizó el siguiente análisis de diagnóstico:

Residuales

Ver a continuación el diagnóstico de residuales:

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

```
## Poisson model
```

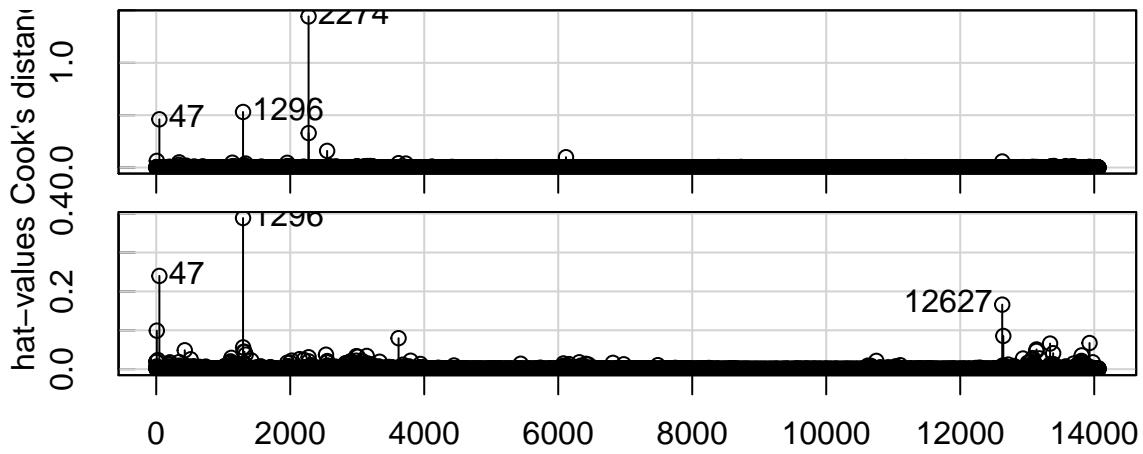


Se observa, en el diagnóstico de residuales, que los residuos se ajustan adecuadamente a las bandas de confianza. Sin embargo, en las colas existe mayor dispersión. Esto podría deberse a los valores atípicos que existen dentro de la muestra (esto se observó en el análisis exploratorio).

Visualización de puntos influyentes

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

Diagnóstico 10.3



Index

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 | nivel_riesgo3 | nivel_riesgo4 |
| ## | 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.

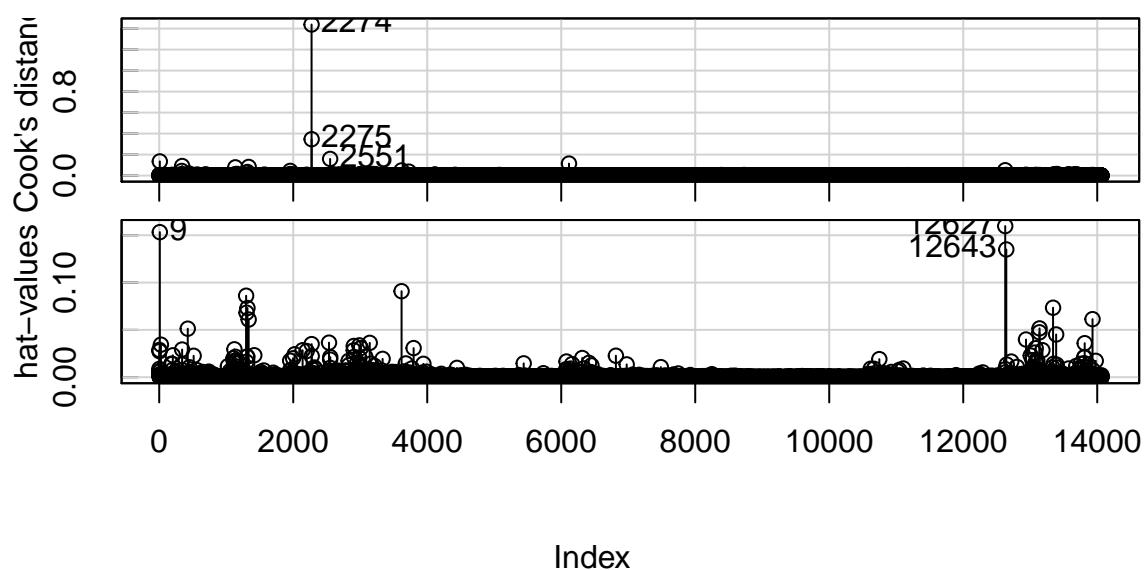
Asimismo, se han hecho los siguientes diagnósticos.

```

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

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

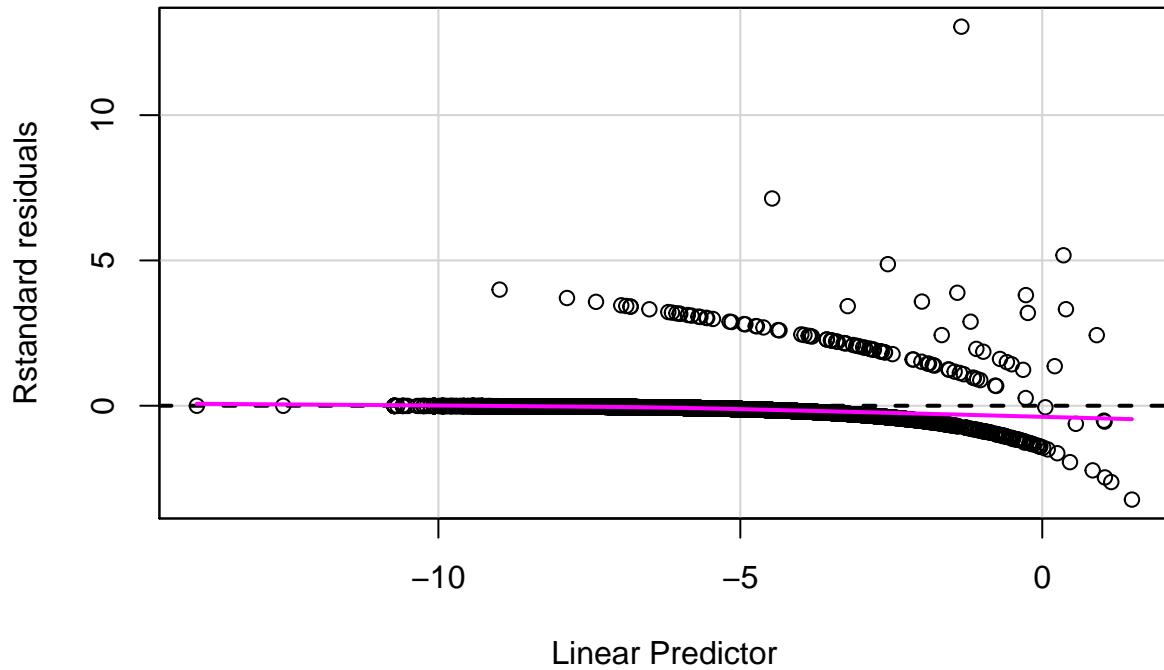
```



```

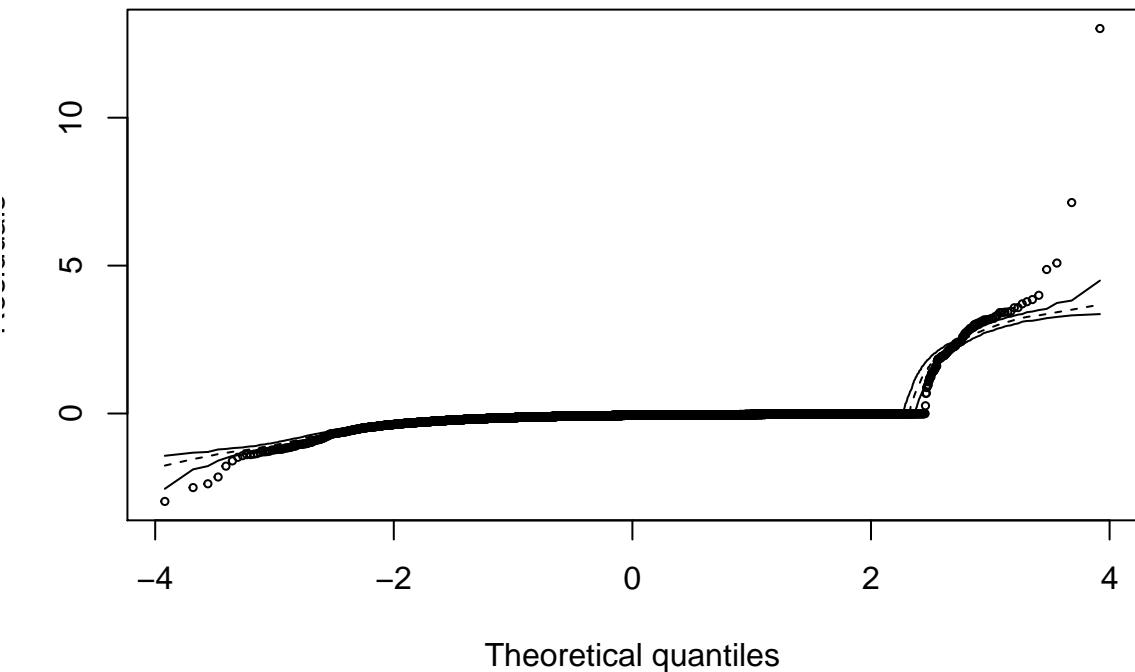
###Grafico de residuos versus valores ajustados
residualPlot(modelo2,type="rstandard")

```



```
### Gráfico 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.