Regresion - Entrega 2

2022-10-06

Regresion - MARIAJOSE MURILLO

En el 2021 fueron las elecciones presidenciales, cuya segunda vuelta fue disputada por Keiko Fujimori y Pedro Castillo, donde resulto ganador el segundo, sin embargo, que variables pudieron influenciar en su victoria?

- “dep”: departamento del Peru
- “prov”: provincia del Peru
- “habil”: electores habiles - tambien se usara como variable de control

Variable dependiente:

“casti”: votos a Pedro Castillo

Variables independientes:

“gas”: personas con acceso a gas GLP

“sis”: personas con acceso al seguro de salud

“acsedu”: personas con acceso a la educacion

Primero veamos los contenidos:

```
rm(list = ls())
knitr::knit_hooks$set(inline = as.character)

gitLink="https://github.com/MajoMurillo/Estadistica2---Trabajo/blob/main/dataCGSE.xlsx?raw=true"
dataCGSE=rio::import(gitLink)

library(magrittr)
dataCGSE%>%
    rmarkdown::paged_table()
```

	Dep <chr>	Prov <chr>	casti <dbl>	gas <dbl>	sis <dbl>	acsedu <dbl>	habil <dbl>
1	Amazonas	Chachapoyas	25980	10641	3286	18668	62110
2	Amazonas	Bagua	8374	9917	54088	24452	20917
3	Amazonas	Bongara	15671	4659	18057	7536	40752
4	Amazonas	Condorcanqui	14024	1536	33802	16281	38273
5	Amazonas	Luya	12606	6339	36541	13777	35017
6	Amazonas	Rodriguez de Mendoza	7967	3781	20815	8693	22886
7	Amazonas	Utcubamba	36540	16078	80664	34334	86231
8	Ancash	Huaraz	2325	30307	90834	55893	5817
9	Ancash	Aija	5056	478	4542	1871	10921
10	Ancash	Antonio Raymondi	2860	776	11048	4966	5968

1-10 of 196 rows Previous **1** 2 3 4 5 6 … 20 Next

Vemos que toda las variables son numericas a excepcion de “departamento” y “provincia” por lo que primero se optara por una **regresion lineal**

```
str(dataCGSE)
```

```
## 'data.frame':    196 obs. of  7 variables:
##  $ Dep   : chr  "Amazonas" "Amazonas" "Amazonas" "Amazonas" ...
##  $ Prov  : chr  "Chachapoyas" "Bagua" "Bongara" "Condorcanqui" ...
##  $ casti : num  25980 8374 15671 14024 12606 ...
##  $ gas   : num  10641 9917 4659 1536 6339 ...
##  $ sis   : num  3286 54088 18057 33802 36541 ...
##  $ acsedu: num  18668 24452 7536 16281 13777 ...
##  $ habil : num  62110 20917 40752 38273 35017 ...
```

Nuestra primera hipotesis es: **A nivel provincial, la cantidad de personas que votaron por Pedro Castillo estara afectada por si usan gas GLP**

```
library(knitr)
library(modelsummary)

h1=formula(casti~gas)

rl1=lm(h1, data = dataCGSE)

model1=list('OLS votantes de Castillo (1)'=rl1)
modelsummary(model1, title = "Tabla 1: Resumen de Regresion Lineal",
             stars = TRUE,
             output = "kableExtra")
```

```
## Warning in !is.null(rmarkdown::metadata$output) && rmarkdown::metadata$output
## %in% : 'length(x) = 3 > 1' in coercion to 'logical(1)'
```

Tabla 1: Resumen de Regresion Lineal

	OLS votantes de Castillo (1)
(Intercept)	43534.236***

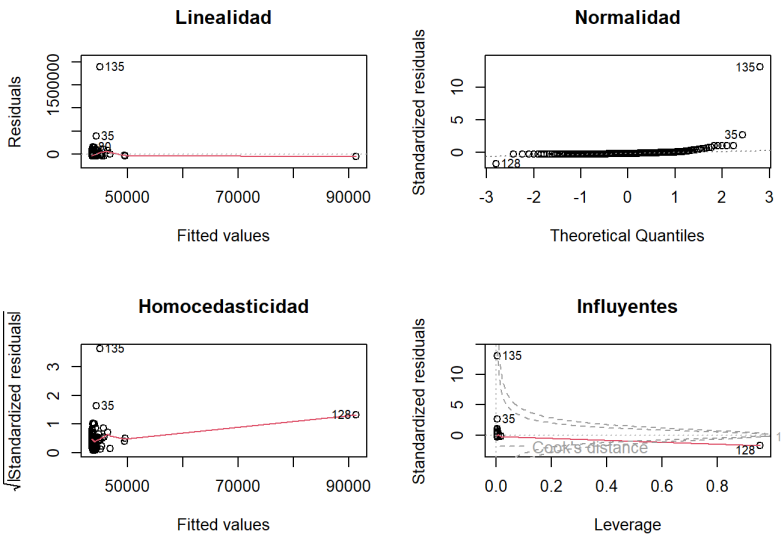
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

	OLS votantes de Castillo (1)
	(10500.202)
gas	0.026
	(0.078)
Num.Obs.	196
R2	0.001
R2 Adj.	−0.005
AIC	5217.7
BIC	5227.5
Log.Lik.	−2605.852
F	0.112
RMSE	143806.91

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Como vemos en la Tabla 1, el predictor covariado (gas) salio con un valor negativo y no significativo (ya que sale sin asteriscos), mientras que su R2 ajustado sale tambien con valores negativos. Sin embargo no nos da un buen ajuste y dificilmente puede ser util :

```
par(mfrow = c(2, 2))
plot(rl1, 1,caption = '');title(main="Linealidad")
plot(rl1, 2, caption = '');title(main="Normalidad")
plot(rl1, 3, caption = '');title(main="Homocedasticidad")
plot(rl1, 5, caption = '');title(main="Influyentes")
```



Para mejorar este modelo, podemos incluir la variable de control (“habil”)

```
library(knitr)
library(modelsummary)

h1control=formula(casti~gas + habil)

rl2=lm(h1control, data = dataCGSE)

modelslm=list('OLS votantes de Castillo (1)'=rl1,'OLS votantes de Castillo (2'=rl2)
modelsummary(modelslm, title = "Regresiones Lineales",
             stars = TRUE,
             output = "kableExtra")
```

Regresiones Lineales

	OLS votantes de Castillo (1)	OLS votantes de Castillo (2
(Intercept)	43534.236***	12076.167***
	(10500.202)	(1601.736)
gas	0.026	−0.001
	(0.078)	(0.012)
habil		0.259***
		(0.003)
Num.Obs.	196	196
R2	0.001	0.978
R2 Adj.	−0.005	0.978

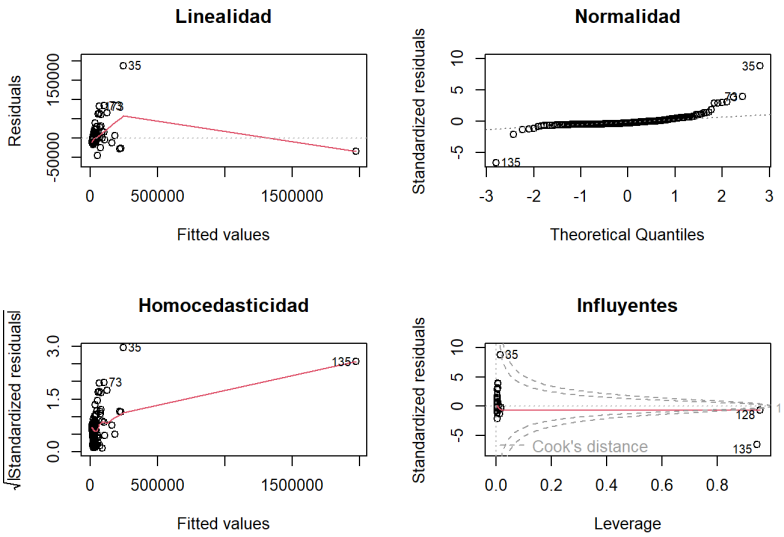
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

	OLS votantes de Castillo (1)	OLS votantes de Castillo (2
AIC	5217.7	4472.6
BIC	5227.5	4485.7
Log.Lik.	−2605.852	−2232.276
F	0.112	4271.803
RMSE	143806.91	21380.24

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

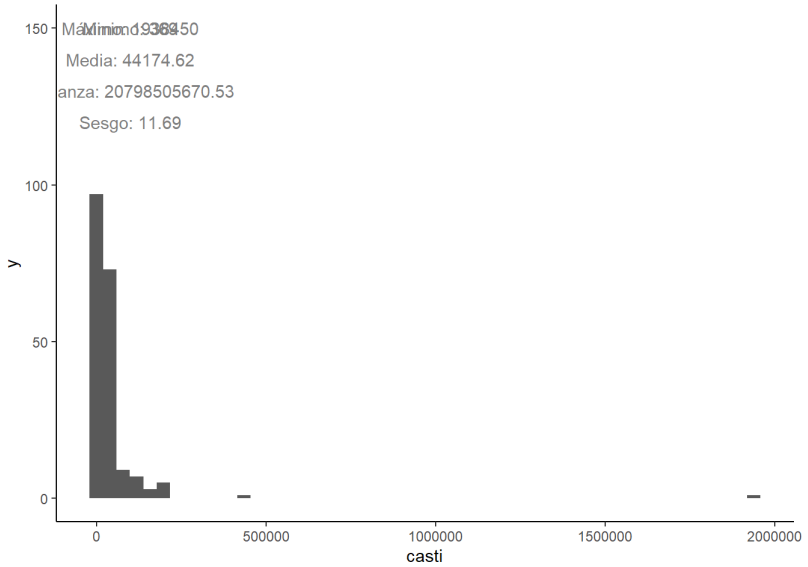
Se presenta una mejora en el R2 ajustado

```
par(mfrow = c(2, 2))
plot(rl2, 1,caption = '');title(main="Linealidad")
plot(rl2, 2, caption = '');title(main="Normalidad")
plot(rl2, 3, caption = '');title(main="Homocedasticidad")
plot(rl2, 5, caption = '');title(main="Influyentes")
```



Para entender mejor la naturaleza de la variable, hacemos un grafico de barras

```
library(ggplot2)
VarProv=dataCGSE$casti
descris=list(min=min(VarProv),
             max=max(VarProv),
             media=round(mean(VarProv),2),
             var=round(var(VarProv),2),
             asim=round(e1071::skewness(VarProv),2))

base=ggplot(data=dataCGSE, aes(x=casti)) + theme_classic()
hist=base + geom_histogram(bins=50)
histInfo=hist + annotate("text", x = 100000, y = 150,
                         color='grey50',
                       label = paste0("Minimo: ",descris$min))
histInfo = histInfo + annotate("text", x = 100000, y = 150,
                       color='grey50',
                       label = paste0("Máximo: ",descris$max))

histInfo = histInfo + annotate("text", x = 100000, y = 140,
                       color='grey50',
                       label = paste0("Media: ",descris$media))

histInfo = histInfo + annotate("text", x = 100000, y = 130,
                       color='grey50',
                       label = paste0("Varianza: ",descris$var))

histInfo = histInfo + annotate("text", x = 100000, y =120,
                       color='grey50',
                       label = paste0("Sesgo: ",descris$asim))

histInfo
```



Nos muestra un distribucion con sesgo positivo, recordandonos que nuestra variable dependiente representa valores enteros positivos

REGRESION POISSON

Comparamos el resultado de la regresion lineal anterior controlada por los electores habiles con la regresion Poisson

```
rp1=glm(h1, data = dataCGSE,
        offset=log(habil),
        family = poisson(link = "log"))

# displaying results
modelslmpoi=list('OLS votantes de Castillo (2)'=rl2,
                 'POISSON votantes de Castillo'=rp1)

modelsummary(modelslmpoi, title = "Regresiones OLS y Poisson",
             stars = TRUE,
             output = "kableExtra")
```

Regresiones OLS y Poisson

	OLS votantes de Castillo (2)	POISSON votantes de Castillo
(Intercept)	12076.167***	−1.017***
	(1601.736)	(0.000)
gas	−0.001	0.000***
	(0.012)	(0.000)
habil	0.259***	
	(0.003)	
Num.Obs.	196	196
R2	0.978	
R2 Adj.	0.978	
AIC	4472.6	160876328.1
BIC	4485.7	160876334.7
Log.Lik.	−2232.276	−80438162.051
F	4271.803	11244.777
RMSE	21380.24	56156.41

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Alteraremos nuestra primera hipotesis, siendo ahora: **A nivel provincial, la cantidad de personas que votaron por Pedro Castillo esta afectada por si cuentan con gas GLP, seguro de salud y acceso a la educacion**

```
h2=formula(casti ~ gas+sis+acsedu)

rp2=glm(h2, data = dataCGSE, offset=log(habil),
        family = poisson(link = "log"))


modelsPois=list('POISSON votantes de Castillo (1)'=rp1,
                'POISSON votantes de Castillo (2)'=rp2)
modelsummary(modelsPois,
             title = "Regresiones Poisson anidadas",
             stars = TRUE,
             output = "kableExtra")
```

Regresiones Poisson anidadas

	POISSON votantes de Castillo (1)	POISSON votantes de Castillo (2)

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

	POISSON votantes de Castillo (1)	POISSON votantes de Castillo (2)
(Intercept)	−1.017***	−1.039***
	(0.000)	(0.000)
gas	0.000***	0.000***
	(0.000)	(0.000)
sis		0.000***
		(0.000)
acsedu		0.000***
		(0.000)
Num.Obs.	196	196
AIC	160876328.1	160790071.6
BIC	160876334.7	160790084.7
Log.Lik.	−80438162.051	−80395031.808
F	11244.777	34452.273
RMSE	56156.41	39294.45

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Equidispersion

Uno de los supuestos en la Regresión Poisson es que la media y la varianza sean iguales.

```
overdispersion=AER::dispersiontest(rp2,alternative='greater')$ p.value<0.05
underdispersion=AER::dispersiontest(rp2,alternative='less')$ p.value<0.05
# tabla
testResult=as.data.frame(rbind(overdispersion,underdispersion))
names(testResult)='Es probable?'
testResult%>%kable(caption = "Test de Equidispersión")%>%kableExtra::kable_styling()
```

Test de Equidispersión

	Es probable?
overdispersion	TRUE
underdispersion	FALSE

Haciendo la equidispersion nos damos cuenta que es improbable que la varianza sea igual a la media, por lo que intentaremos hacer la Quasi Poisson para tratar la presencia de la sobredispersion

REGRESION QUASIPOISSON

```
rqp = glm(h2, data = dataCGSE, offset=log(habil),
          family = quasipoisson(link = "log"))

modelsPQP=list('POISSON votantes de Castillo (2)'=rp2, 'QUASIPOISSON votantes de Castillo'=rqp)

modelsummary(modelsPQP, title = "Regresion Poisson y QuasiPoisson",
             stars = TRUE,
             output = "kableExtra")
```

Regresion Poisson y QuasiPoisson

	POISSON votantes de Castillo (2)	QUASIPOISSON votantes de Castillo
(Intercept)	−1.039***	−1.039***
	(0.000)	(0.029)
gas	0.000***	0.000*
	(0.000)	(0.000)
sis	0.000***	0.000**
	(0.000)	(0.000)
acsedu	0.000***	0.000
	(0.000)	(0.000)
Num.Obs.	196	196
AIC	160790071.6	
BIC	160790084.7	
Log.Lik.	−80395031.808	
F	34452.273	5.949

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

	POISSON votantes de Castillo (2)	QUASIPOISSON votantes de Castillo
RMSE	39294.45	39294.45

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Nos muestra que los coeficientes son los mismos para ambos modelos.

Sin embargo (por recomendacion del profesor), usaremos la regresion Binomial Negativa como otra forma de tratar la sobredispersion, ademas porque es la mas utilizada para estos casos

REGRESION BINOMIAL NEGATIVA

```
h2off=formula(casti~gas + sis + acsedu + offset(log(habil)))
rbn=MASS::glm.nb(h2off,data=dataCGSE)

summary(rbn)
```

```
##
## Call:
## MASS::glm.nb(formula = h2off, data = dataCGSE, init.theta = 8.611336358,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.7852  -0.6969  -0.0637   0.5161   1.8346
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -7.904e-01  3.071e-02 -25.737   <2e-16 ***
## gas          4.441e-07  1.109e-06   0.400    0.689
## sis         -8.028e-07  7.209e-07  -1.114    0.265
## acsedu       2.494e-07  9.529e-07   0.262    0.793
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(8.6113) family taken to be 1)
##
##     Null deviance: 202.42  on 195  degrees of freedom
## Residual deviance: 199.87  on 192  degrees of freedom
## AIC: 4026.7
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  8.611
##           Std. Err.:  0.855
##
##  2 x log-likelihood:  -4016.723
```

```
modelsQP_BN=list('POISSON votantes de Castillo (2)'=rp2,
                 'QuasiPoisson votantes de Castillo (2)'=rqp,
                 'Binomial Negativa votantes de Castillo (2)'=rbn)

f <- function(x) format(x, digits = 4, scientific = FALSE)
modelsummary(modelsQP_BN,fmt=f,
             exponentiate = T,
             statistic = 'conf.int',
             title = "EXP() de la Regresiones Poisson, Quasi Poisson  y Binomial Negativa",
             stars = TRUE,
             output = "kableExtra")
```

EXP() de la Regresiones Poisson, Quasi Poisson y Binomial Negativa

	POISSON votantes de Castillo (2)	QuasiPoisson votantes de Castillo (2)	Binomial Negativa votantes de Castillo (2)
(Intercept)	0.3538***	0.3538***	0.4537***
	[0.3535, 0.3541]	[0.3343, 0.3742]	[0.4271, 0.4822]
gas	1.0000***	1.0000*	1.0000
	[1.0000, 1.0000]	[1.0000, 1.0000]	[1.0000, 1.0000]
sis	1.0000***	1.0000**	1.0000
	[1.0000, 1.0000]	[1.0000, 1.0000]	[1.0000, 1.0000]
acsedu	1.0000***	1.0000	1.0000
	[1.0000, 1.0000]	[1.0000, 1.0000]	[1.0000, 1.0000]
Num.Obs.	196	196	196
AIC	160790071.6		4026.7
BIC	160790084.7		4043.1
Log.Lik.	−80395031.808		−2008.361
F	34452.273	5.949	0.929
RMSE	39294.45	39294.45	117675.53

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Comparacion de modelos

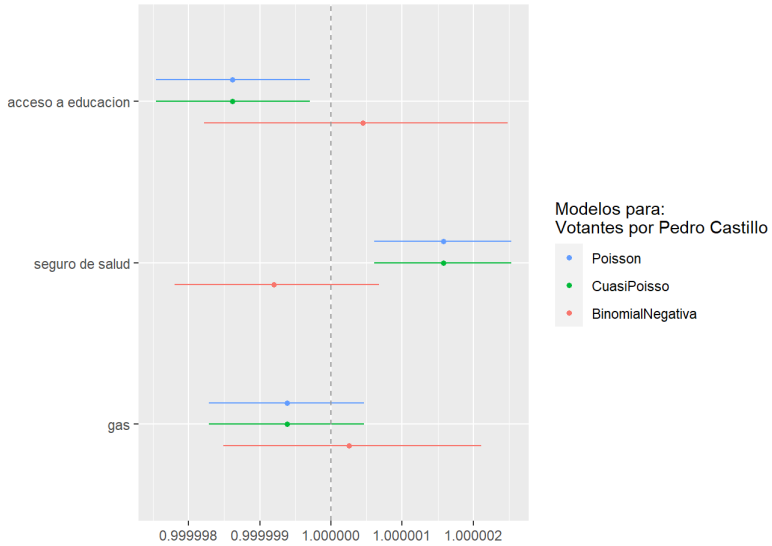
Usamos anova para la comparacion

```
anova(rp2,rqp,rbn, test = "Chisq") %>%
kable(caption = "Tabla ANOVA para comparar modelos")%>%kableExtra::kable_styling(full_width = FALSE)
```

Tabla ANOVA para comparar modelos

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
192	1054200.7853	NA	NA	NA
192	1054200.7853	0	0	NA
192	199.8725	0	1054001	NA

```
library(ggplot2)
dotwhisker::dwplot(list(Poisson=rqp,CuasiPoisso=rqp,BinomialNegativa=rbn),exp=T) + scale_y_discrete(labels=c("gas","seguro d
e salud", "acceso a educacion")) + scale_color_discrete(name="Modelos para:\nVotantes por Pedro Castillo") + geom_vline(
           xintercept = 1,
           colour = "grey60",
           linetype = 2
      )
```



.

.

Regresion - MARIA HERRARA

```
library(rio)
lkXLSX="https://github.com/MajoMurillo/Estadistica2---Trabajo/blob/main/indeppp.xlsx?raw=true"
indps=import(lkXLSX)
```

```
lkXLSY="https://github.com/MajoMurillo/Estadistica2---Trabajo/raw/main/dependiente2.csv"
dep=import(lkXLSY)
```

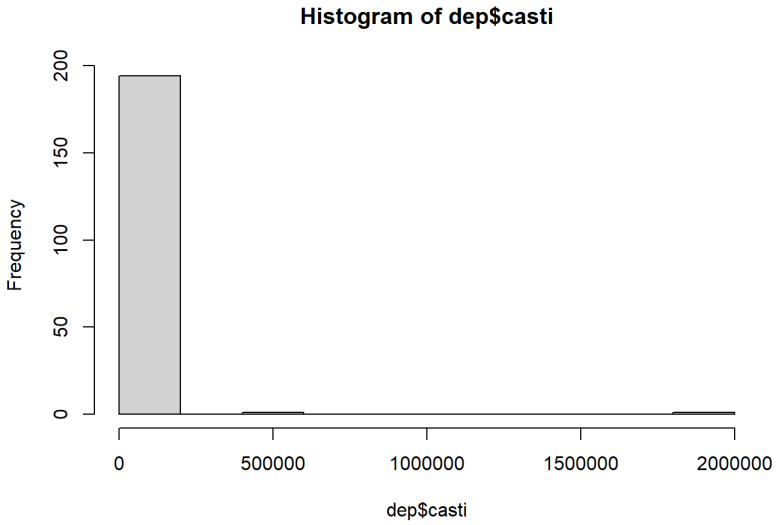
```
str(dep)
```

```
## 'data.frame':    196 obs. of  3 variables:
##  $ Provincia: chr  "Bagua" "Bongara" "Chachapoyas" "Condorcanqui" ...
##  $ casti    : int  25980 8374 15671 14024 12606 7967 36540 2325 5056 2860 ...
##  $ habil    : int  62110 20917 40752 38273 35017 22886 86231 5817 10921 5968 ...
```

```
dep$casti=as.numeric(dep$casti)
```

```
dep$habil=as.numeric(dep$habil)
```

```
hist(dep$casti)
```



Vemos que tiene una asimetría positiva marcada

```
indps$Código=NULL
```

```
library(stringr)
library(magrittr) # para %>%
indps$depar=str_split(string = indps$Provincia,
                      pattern = ', provincia:',
                      simplify = T)[,1]

indps$provin=str_split(string = indps$Provincia,
                       pattern = ', provincia:',
                       simplify = T)[,2]
```

```
indps$Provincia=NULL
```

```
indps=indps[,c(10:16)]
```

```
dep$Provincia =trimws(dep$Provincia,which=c("both"),whitespace = "[\\h\\v]") # el espacio en blanco se determina "\\h\\v", o también "\t\r\n"
```

```
indps$provin =trimws(indps$prov,which=c("both"),whitespace = "[\\h\\v]") # el espacio en blanco se determina "\\h\\v", o tam bién "\t\r\n"
```

```
names(indps)[7]=c("Provincia")
```

```
str(dep)
```

```
## 'data.frame':    196 obs. of  3 variables:
##  $ Provincia: chr  "Bagua" "Bongara" "Chachapoyas" "Condorcanqui" ...
##  $ casti    : num  25980 8374 15671 14024 12606 ...
##  $ habil    : num  62110 20917 40752 38273 35017 ...
```

```
indps=indps[,c(7,6,1:5)]
```

```
basefinal=merge(indps,dep)
```

```
str(basefinal)
```

```
## 'data.frame':    193 obs. of  9 variables:
##  $ Provincia: chr  "Abancay" "Acobamba" "Acomayo" "Aija" ...
##  $ depar    : chr  "Apurímac" "Huancavelica" "Cusco" "Áncash" ...
##  $ Ln       : num  55045 30231 18698 2517 12869 ...
##  $ Luz      : num  29588 8968 5364 1528 20021 ...
##  $ LuzN     : num  3050 2256 1541 413 7946 ...
##  $ AgS      : num  28019 7308 5593 1392 10980 ...
##  $ AgN      : num  2583 1914 472 111 3316 ...
##  $ casti    : num  43244 19060 13 2325 37196 ...
##  $ habil    : num  82538 33498 19229 5817 90033 ...
```

```
names(basefinal)
```

```
## [1] "Provincia" "depar"     "Ln"        "Luz"       "LuzN"      "AgS"
## [7] "AgN"       "casti"     "habil"
```

Número de personas que votaron por Castillo –> VD Personas que tienen Luz electrica en casa –> VI Personas que tienen Agua en casa –> VI Etniticidad –>VI

Regresión:

A nivel provincial, el voto hacia el postulante a presidencia Castillo está afectada por el acceso a bienes básicos como luz

```
library(knitr)
library(modelsummary)

h1=formula(casti~Luz)

rl1=lm(h1, data = basefinal)

model1=list('OLS asegurados (I)'=rl1)
modelsummary(model1, title = "Resumen de Regresion Lineal",
             stars = TRUE,
             output = "kableExtra")
```

Resumen de Regresion Lineal

	OLS asegurados (I)
(Intercept)	11995.834***
	(1501.760)
Luz	0.939***
	(0.010)
Num.Obs.	193
R2	0.980
R2 Adj.	0.980
AIC	4381.5
BIC	4391.3
Log.Lik.	−2187.771
F	9522.276
RMSE	20265.22

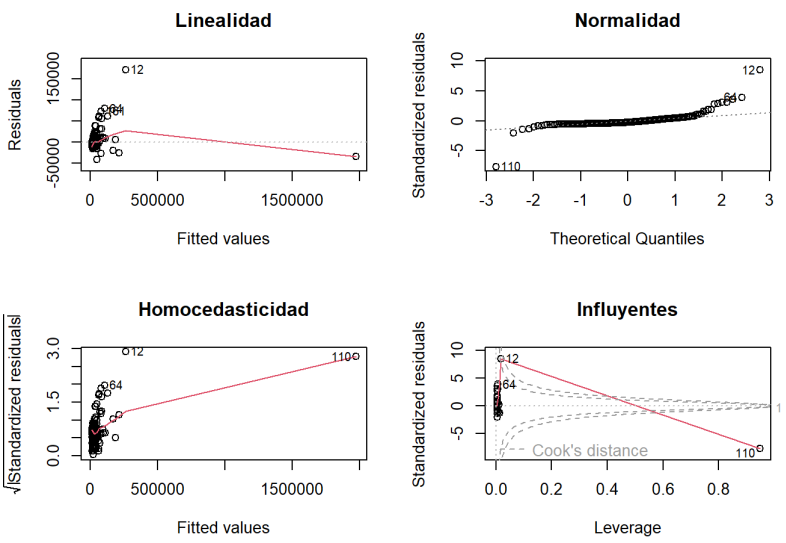
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

```
summary(rl1)
```

```
##
## Call:
## lm(formula = h1, data = basefinal)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -42273  -9441  -5629   3762 171295
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.200e+04  1.502e+03   7.988 1.26e-13 ***
## Luz         9.395e-01  9.628e-03  97.582  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20370 on 191 degrees of freedom
## Multiple R-squared:  0.9803, Adjusted R-squared:  0.9802
## F-statistic:  9522 on 1 and 191 DF,  p-value: < 2.2e-16
```

Vemos que la variable se valida (p-value menor a 0.05) y el R es alto, pero hacemos más pruebas, por ello procedemos a los supuestos:

```
par(mfrow = c(2, 2))
plot(rl1, 1,caption = '');title(main="Linealidad")
plot(rl1, 2, caption = '');title(main="Normalidad")
plot(rl1, 3, caption = '');title(main="Homocedasticidad")
plot(rl1, 5, caption = '');title(main="Influyentes")
```



Vemos que la linealidad esta cayendo, igualmente vemos casos atípicos y la homocedasticidad no cumple la recta. Por lo tanto vemos que los supuestos de la RLM caen.

Agregamos la variable habil(como variable control y hacemos otro RLM):

```
library(knitr)
library(modelsummary)

h1control=formula(casti~Luz + habil)

rl2=lm(h1control, data = basefinal)

modelslm=list('OLS votos casti (I)'=rl1,'OLS votos casti (II)'=rl2)
modelsummary(modelslm, title = "Regresiones Lineales",
             stars = TRUE,
             output = "kableExtra")
```

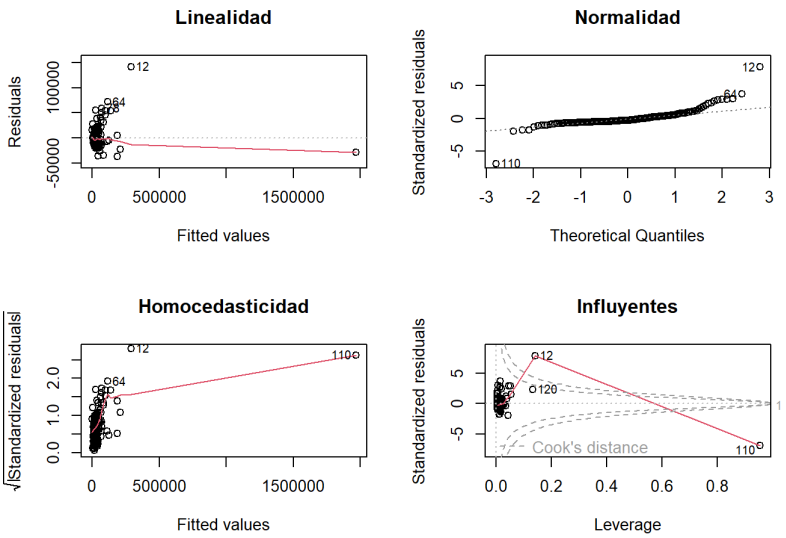
Regresiones Lineales

	OLS votos casti (I)	OLS votos casti (II)
(Intercept)	11995.834***	11768.166***
	(1501.760)	(1437.724)
Luz	0.939***	2.683***
	(0.010)	(0.404)
habil		−0.483***
		(0.112)
Num.Obs.	193	193
R2	0.980	0.982
R2 Adj.	0.980	0.982
AIC	4381.5	4365.4
BIC	4391.3	4378.5
Log.Lik.	−2187.771	−2178.724
F	9522.276	5211.024
RMSE	20265.22	19337.25

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

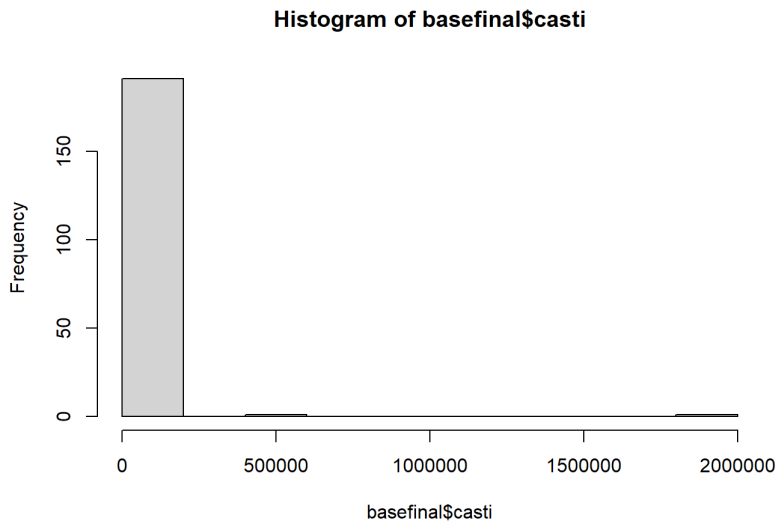
Aunque el poder explicativo no aumento, podemos ver una posible multicolinealidad.

```
par(mfrow = c(2, 2))
plot(rl2, 1,caption = '');title(main="Linealidad")
plot(rl2, 2, caption = '');title(main="Normalidad")
plot(rl2, 3, caption = '');title(main="Homocedasticidad")
plot(rl2, 5, caption = '');title(main="Influyentes")
```



Vemos que aún fallan los supuestos , por lo que procedemos a hacer una revisión gráfica:

```
hist(basefinal$casti)
```




```
summary(basefinal$casti)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##      13    8579   20001   43636   39090 1938450
```

Es histograma de la Figura 1.3 nos muestra una distrubución con asimetría positiva. Ello nos hace reflexionar que nuestra variable dependiente representa conteos, valores enteros positivos. La regresión lineal tendrá problemas pues asume que la variable dependiente tiene valores reales y no acotados. Por lo que procedemos a trabajar con Poisson. Además la media y la mediana no son iguales por lo que tendremos que ver el caso de equidispersión.

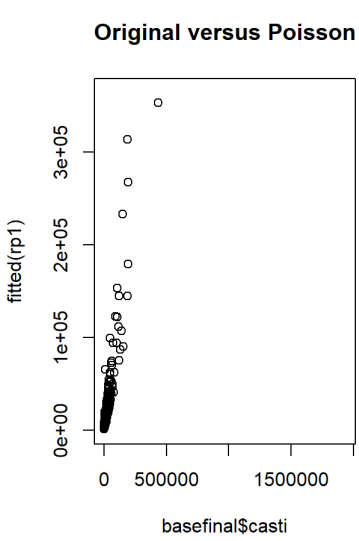
```
rp1=glm(h1, data = basefinal,
        offset=log(habil), #exposure
        family = poisson(link = "log"))

# displaying results
modelslmpoi=list('OLS  (II)'=rl2,
                 'POISSON '=rp1)

modelsummary(modelslmpoi, title = "Regresiones OLS y Poisson",
             stars = TRUE,
             output = "kableExtra")
```

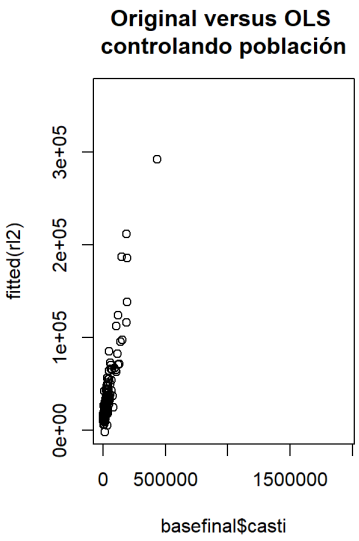
Regresiones OLS y Poisson

	OLS (II)	POISSON
(Intercept)	11768.166***	−0.873***
	(1437.724)	(0.000)
Luz	2.683***	0.000***
	(0.404)	(0.000)
habil	−0.483***	
	(0.112)	
Num.Obs.	193	193
R2	0.982	
R2 Adj.	0.982	
AIC	4365.4	155991169.8
BIC	4378.5	155991176.3
Log.Lik.	−2178.724	−77995582.909
F	5211.024	341931.450
RMSE	19337.25	18696.47

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

```
par(mfrow = c(1, 2))  # divide screen 1 row 2 columns
plot(basefinal$casti,fitted(rp1),ylim=c(0,365000));title(main="Original versus Poisson")
plot(basefinal$casti,fitted(rl2),ylim=c(0,365000));title(main="Original versus OLS \ncontrolando población")
```






```
names(basefinal)
```

```
## [1] "Provincia" "depar"     "Ln"        "Luz"       "LuzN"      "AgS"
## [7] "AgN"       "casti"     "habil"
```

```
h2=formula(casti~Luz + AgS + Ln)

rp2=glm(h2, data = basefinal, offset=log(habil),
        family = poisson(link = "log"))


modelsPois=list('POISSON votos casti (I)'=rp1,
                'POISSON votos casti (II)'=rp2)
modelsummary(modelsPois,
             title = "Regresiones Poisson anidadas",
             stars = TRUE,
             output = "kableExtra")
```

Regresiones Poisson anidadas

	POISSON votos casti (I)	POISSON votos casti (II)
(Intercept)	−0.873***	−0.923***
	(0.000)	(0.001)
Luz	0.000***	0.000***
	(0.000)	(0.000)
AgS		0.000***
		(0.000)
Ln		0.000***
		(0.000)
Num.Obs.	193	193
AIC	155991169.8	155601895.2
BIC	155991176.3	155601908.3
Log.Lik.	−77995582.909	−77800943.603
F	341931.450	258227.365
RMSE	18696.47	10949.49

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

```
overdispersion=AER::dispersiontest(rp2,alternative='greater')$ p.value<0.05
underdispersion=AER::dispersiontest(rp2,alternative='less')$ p.value<0.05
# tabla
testResult=as.data.frame(rbind(overdispersion,underdispersion))
names(testResult)='Es probable?'
testResult%>%kable(caption = "Test de Equidispersión")%>%kableExtra::kable_styling()
```

Test de Equidispersión

	Es probable?
overdispersion	TRUE
underdispersion	FALSE

Hay sobredispersión por lo que procedemos a usar la binomial negativa.

```
rqp = glm(h2, data = basefinal, offset=log(habil),
          family = quasipoisson(link = "log"))

modelsPQP=list('POISSON votos casti (II)'=rp2,'QUASIPOISSON votos casti (II)'=rqp)

modelsummary(modelsPQP, title = "Regresiones Poisson y QuasiPoisson",
             stars = TRUE,
             output = "kableExtra")
```

Regresiones Poisson y QuasiPoisson

	POISSON votos casti (II)	QUASIPOISSON votos casti (II)
(Intercept)	−0.923***	−0.923***
	(0.001)	(0.023)
Luz	0.000***	0.000***
	(0.000)	(0.000)
AgS	0.000***	0.000**
	(0.000)	(0.000)
Ln	0.000***	0.000***
	(0.000)	(0.000)
Num.Obs.	193	193
AIC	155601895.2	
BIC	155601908.3	
Log.Lik.	−77800943.603	
F	258227.365	143.980
RMSE	10949.49	10949.49

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

```
library(arm)
```

```
## Loading required package: MASS
```

```
## Loading required package: Matrix
```

```
## Loading required package: lme4
```

```
##
## Attaching package: 'lme4'
```

```
## The following object is masked from 'package:rio':
##
##     factorize
```

```
##
## arm (Version 1.13-1, built: 2022-8-25)
```

```
## Working directory is C:/Users/jenni/Downloads/PUCP/Estadistica 2/TODO JUNTO AAAA
```

```
cbind(coefPoi=coef(rp2),coefQuasiPoi=coef(rqp))
```

```
##                   coefPoi  coefQuasiPoi
## (Intercept) -9.234752e-01 -9.234752e-01
## Luz         -7.552308e-06 -7.552308e-06
## AgS          6.562992e-06  6.562992e-06
## Ln           4.727818e-06  4.727818e-06
```

```
cbind(sePoi=se.coef(rp2),seQuasiPoi=se.coef(rqp))
```

```
##                    sePoi   seQuasiPoi
## (Intercept) 5.313366e-04 2.250191e-02
## Luz         4.760398e-08 2.016011e-06
## AgS         5.461933e-08 2.313109e-06
## Ln          8.436558e-09 3.572852e-07
```

```
summary(rqp)$dispersion; summary(rp2)$dispersion
```

```
## [1] 1793.492
```

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

```
modelsQPexp=list('QuasiPoisson votos casti (II) exponenciado'=rqp)

f <- function(x) format(x, digits = 4, scientific = FALSE)
modelsummary(modelsQPexp,fmt=f,
             exponentiate = T,
             statistic = 'conf.int',
             title = "EXP() de la Regresión Quasi Poisson (II) para Interpretación",
             stars = TRUE,
             output = "kableExtra")
```

EXP() de la Regresión Quasi Poisson (II) para Interpretación

	QuasiPoisson votos casti (II) exponenciado
(Intercept)	0.3971***
	[0.3799, 0.4149]
Luz	1.0000***
	[1.0000, 1.0000]
AgS	1.0000**
	[1.0000, 1.0000]
Ln	1.0000***
	[1.0000, 1.0000]
Num.Obs.	193
Log.Lik.	
F	143.980
RMSE	10949.49

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

```
h2off=formula(casti~ + Luz+ Ln+ AgS + offset(log(habil)))
rbn=glm.nb(h2off,data=basefinal)

modelsQP_BN=list('Poisson asegurados (II)'=rp2,
                 'QuasiPoisson asegurados (II)'=rqp,
                 'Binomial Negativa asegurados (II)'=rbn)

f <- function(x) format(x, digits = 4, scientific = FALSE)
modelsummary(modelsQP_BN,fmt=f,
             exponentiate = T,
             statistic = 'conf.int',
             title = "EXP() de la Regresiones Poisson, Quasi Poisson  y Binomial Negativa",
             stars = TRUE,
             output = "kableExtra")
```

EXP() de la Regresiones Poisson, Quasi Poisson y Binomial Negativa

	Poisson asegurados (II)	QuasiPoisson asegurados (II)	Binomial Negativa asegurados (II)
(Intercept)	0.3971***	0.3971***	0.4109***
	[0.3967, 0.3976]	[0.3799, 0.4149]	[0.386, 0.4378]
Luz	1.0000***	1.0000***	1.0000*
	[1.0000, 1.0000]	[1.0000, 1.0000]	[1.000, 1.0000]
AgS	1.0000***	1.0000**	1.0000+
	[1.0000, 1.0000]	[1.0000, 1.0000]	[1.000, 1.0000]
Ln	1.0000***	1.0000***	1.0000***
	[1.0000, 1.0000]	[1.0000, 1.0000]	[1.000, 1.0000]
Num.Obs.	193	193	193
AIC	155601895.2		3972.5
BIC	155601908.3		3988.9
Log.Lik.	−77800943.603		−1981.271
F	258227.365	143.980	20.182
RMSE	10949.49	10949.49	22582.72

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

```
anova(rp2,rqp,rbn, test = "Chisq") %>%
kable(caption = "Tabla ANOVA para comparar modelos")%>%kableExtra::kable_styling(full_width = FALSE)
```

Tabla ANOVA para comparar modelos

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
189	367015.3178	NA	NA	NA

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
189	367015.3178	0	0.0	NA
189	200.1307	0	366815.2	NA

.

.

Regresion - ANGELO PALOMINO

```
library(rio)
link="https://github.com/MajoMurillo/Estadistica2---Trabajo/blob/main/base%20de%20datos-Anelo.xlsx?raw=true"
data=import(link)
```

```
library(summarytools)
library(kableExtra)
dfSummary(data,
          plain.ascii  = FALSE,
          varnumbers = FALSE,
          style        = "grid",
          graph.col=F,
          na.col    = FALSE) %>%
    kable(caption = "Descriptivos Univariados")%>%
    kable_styling(full_width = F)
```

```
## data was converted to a data frame
```

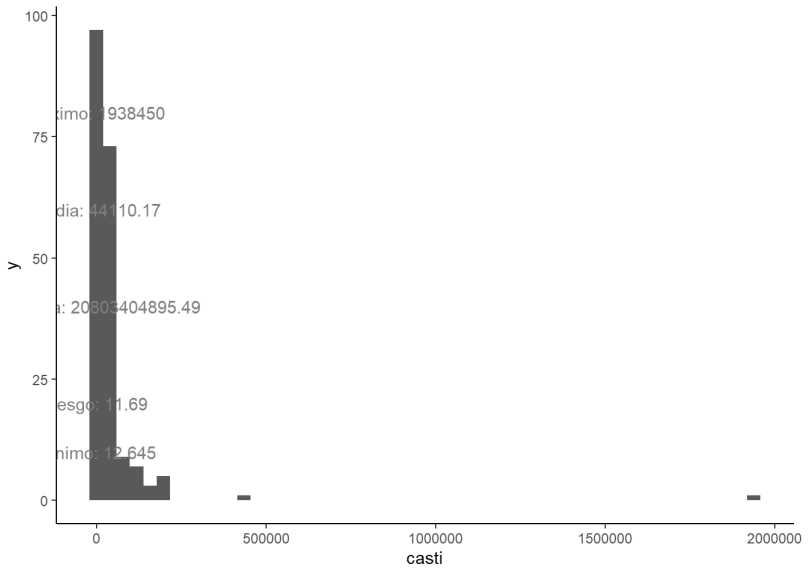
```
## Warning in seq_len(ncol(x)): first element used of 'length.out' argument
```

Descriptivos Univariados

Variable	Stats / Values	Freqs (% of Valid)	Valid
Provincia [character]	1. Abancay 2. Acobamba 3. Acomayo 4. Aija 5. Alto Amazonas 6. Ambo 7. Andahuaylas 8. Angaraes 9. Anta 10. Antabamba [186 others]	1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) 1 (0.5%) \186 (94.9%)	196 (100.0%)
casti [numeric]	Mean (sd) : 44110.2 (144233.9) min < med < max: 12.6 < 19896 < 1938450 IQR (CV) : 30502 (3.3)	196 distinct values	196 (100.0%)
habil [numeric]	Mean (sd) : 123891.5 (549863.7) min < med < max: 3041 < 45547 < 7558581 IQR (CV) : 70165.2 (4.4)	196 distinct values	196 (100.0%)
Jóvenes (18 - 29 años) [numeric]	Mean (sd) : 29935 (133310.2) min < med < max: 489 < 9859.5 < 1823609 IQR (CV) : 16398.8 (4.5)	196 distinct values	196 (100.0%)
Sí tiene conexión a internet [numeric]	Mean (sd) : 11807.1 (84911.1) min < med < max: 3 < 659.5 < 1171306 IQR (CV) : 3352.8 (7.2)	184 distinct values	196 (100.0%)
No tiene conexión a internet [numeric]	Mean (sd) : 30296.4 (87639.1) min < med < max: 685 < 14894.5 < 1182644 IQR (CV) : 20926.8 (2.9)	196 distinct values	196 (100.0%)
Ninguna [numeric]	Mean (sd) : 6022.2 (33240) min < med < max: 38 < 1424 < 458304 IQR (CV) : 3663.2 (5.5)	187 distinct values	196 (100.0%)
reliion [numeric]	Mean (sd) : 112326.7 (482136.9) min < med < max: 1856 < 39675.5 < 6602456 IQR (CV) : 60216.2 (4.3)	196 distinct values	196 (100.0%)

```
data$Provincia=as.factor(data$Provincia)
```

```
library(ggplot2)
VarDep=data$casti
descris=list(min=min(VarDep),
             max=max(VarDep),
             media=round(mean(VarDep),2),
             var=round(var(VarDep),2),
             asim=round(e1071::skewness(VarDep),2))
base=ggplot(data=data, aes(x=casti)) + theme_classic()
hist=base + geom_histogram(bins=50)
histInfo=hist + annotate("text", x = 100 , y = 10 ,
                         color='grey50',
                        label = paste0("Minimo: ",descris$min))
histInfo = histInfo + annotate("text", x = 100 , y = 80 ,
                        color='grey50',
                        label = paste0("Máximo: ",descris$max))
histInfo = histInfo + annotate("text", x = 100 , y = 60 ,
                        color='grey50',
                        label = paste0("Media: ",descris$media))
histInfo = histInfo + annotate("text", x = 100 , y = 40 ,
                        color='grey50',
                        label = paste0("Varianza: ",descris$var))
histInfo = histInfo + annotate("text", x = 100 , y = 20 ,
                        color='grey50',
                        label = paste0("Sesgo: ",descris$asim))
histInfo
```




```
h1=formula(data$casti~data$`Jóvenes (18 - 29 años)`+data$`Sí tiene conexión a internet`+data$`No tiene conexión a internet`+
data$Ninguna+data$reliion)
rp1=glm(h1, data=data, offset = log(data$habil),
        family = poisson(link = "log"))
```

```
## Warning in dpois(y, mu, log = TRUE): non-integer x = 12.645000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 12.645000
```

```
modelsPois=list('POISSON asegurados (I)'=rp1)
library(modelsummary)
modelsummary(modelsPois,
             title = "Regresiones Poisson anidadas",
             stars = TRUE,
             output = "kableExtra")
```

```
## Warning in dpois(y, mu, log = TRUE): non-integer x = 12.645000

## Warning in dpois(y, mu, log = TRUE): non-integer x = 12.645000
```

Regresiones Poisson anidadas

(Intercept)
data$Jóvenes (18 - 29 años)$ < /td >< td style =" text - align : center;"> 0.000 * ** < /td >< /tr >< tr >< td style =" text - align : left;">< /td ><
conexión a internet
data$No tiene conexión a internet$ < /td >< td style =" text - align : center;"> 0.000 * ** < /td >< /tr >< tr >< td style =" text - align : left;">< /t
data$reliion
+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Num.Obs.
F
RMSE

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

```
overdispersion=AER::dispersiontest(rp1,alternative='greater')$ p.value<0.05
underdispersion=AER::dispersiontest(rp1,alternative='less')$ p.value<0.05
# tabla
testResult=as.data.frame(rbind(overdispersion,underdispersion))
names(testResult)='Es probable?'
testResult%>%kable(caption = "Test de Equidispersión")%>%kableExtra::kable_styling()
```

Test de Equidispersión

	Es probable?
overdispersion	TRUE
underdispersion	FALSE

#binomial

```
h2off=formula(data$casti~data$`Jóvenes (18 - 29 años)`+data$`Sí tiene conexión a internet`+data$`No tiene conexión a interne
t`+data$Ninguna+data$reliion+ offset(log(habil)))
library(MASS)
rbn=MASS::glm.nb(h2off,data=data)
```

```
## Warning in dpois(y, mu, log = TRUE): non-integer x = 12.645000
```

```
summary(rbn)
```

```
##
## Call:
## MASS::glm.nb(formula = h2off, data = data, init.theta = 6.654621408,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.2233  -0.5464   0.0357   0.4712   1.9813
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -8.460e-01  4.872e-02 -17.363  < 2e-16 ***
## data$`Jóvenes (18 - 29 años)`     4.490e-06  1.252e-05   0.359  0.71980
## data$`Sí tiene conexión a internet`  4.272e-05  9.485e-06   4.504 6.67e-06 ***
## data$`No tiene conexión a internet`  3.311e-05  7.901e-06   4.191 2.78e-05 ***
## data$Ninguna                     -2.014e-05  1.007e-05  -1.999  0.04556 *
## data$reliion                     -1.342e-05  4.136e-06  -3.246  0.00117 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(6.6546) family taken to be 1)
##
##     Null deviance: 225.54  on 195  degrees of freedom
## Residual deviance: 203.16  on 190  degrees of freedom
## AIC: 4072.2
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  6.655
##           Std. Err.:  0.664
##
##  2 x log-likelihood:  -4058.219
```

```
modelsQP_BN=list('Binomial Negativa asegurados (II)'=rbn)
f <- function(x) format(x, digits = 4, scientific = FALSE)
modelsummary(modelsQP_BN,fmt=f,
             exponentiate = T,
             statistic = 'conf.int',
             title = "EXP() de la Regresiones Poisson, Quasi Poisson  y Binomial Negativa",
             stars = TRUE,
             output = "kableExtra")
```

EXP() de la Regresiones Poisson, Quasi Poisson y Binomial Negativa

(Intercept)

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

data$Jóvenes(18-29años) < /td >< tdstyle =" text-align : center;"> 1.0000 < /td >< /tr >< tr >< tdstyle =" text-align : left;"><< /td >< tds
internet
data$Notieneconexiónainternet < /td >< tdstyle =" text-align : center;"> 1.0000 ∗ ∗∗ < /td >< /tr >< tr >< tdstyle =" text-align : left;"><< /
data$reliion
Num.Obs.
AIC
BIC
Log.Lik.
F
RMSE

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001