Modelos para Datos de Conteo

Modelos Estadísticos Avanzados

Santiago Benitez-Vieyra

Distribución de Poisson

Proceso Poisson

Proceso Poisson

Número de parejas, hijos, semillas, plántulas, granos de polen, etc. Número de individuos observados en un lapso de tiempo. Número de individuos observados en un área determinada.

- Registro de sucesos que ocurren en un lapso determinado de tiempo y que son independientes unos de otros.
- Aproximación a eventos de naturaleza binomial (de probabilidad muy pequeña) o a sucesos de naturaleza multinomial (categóricos).

Distribución de Poisson

- Tiene un único parámetro, μ (suele aparecer como λ).
- La media es igual a la varianza.
- Al aumentar la media… aumenta la varianza.

La variable respuesta se distribuye como Poisson

$$Y_i \sim P(\mu_i)$$

$$E(Y_i) = \mu_i \; var(Y_i = \mu_i)$$

El enlace canónico es logaritmo

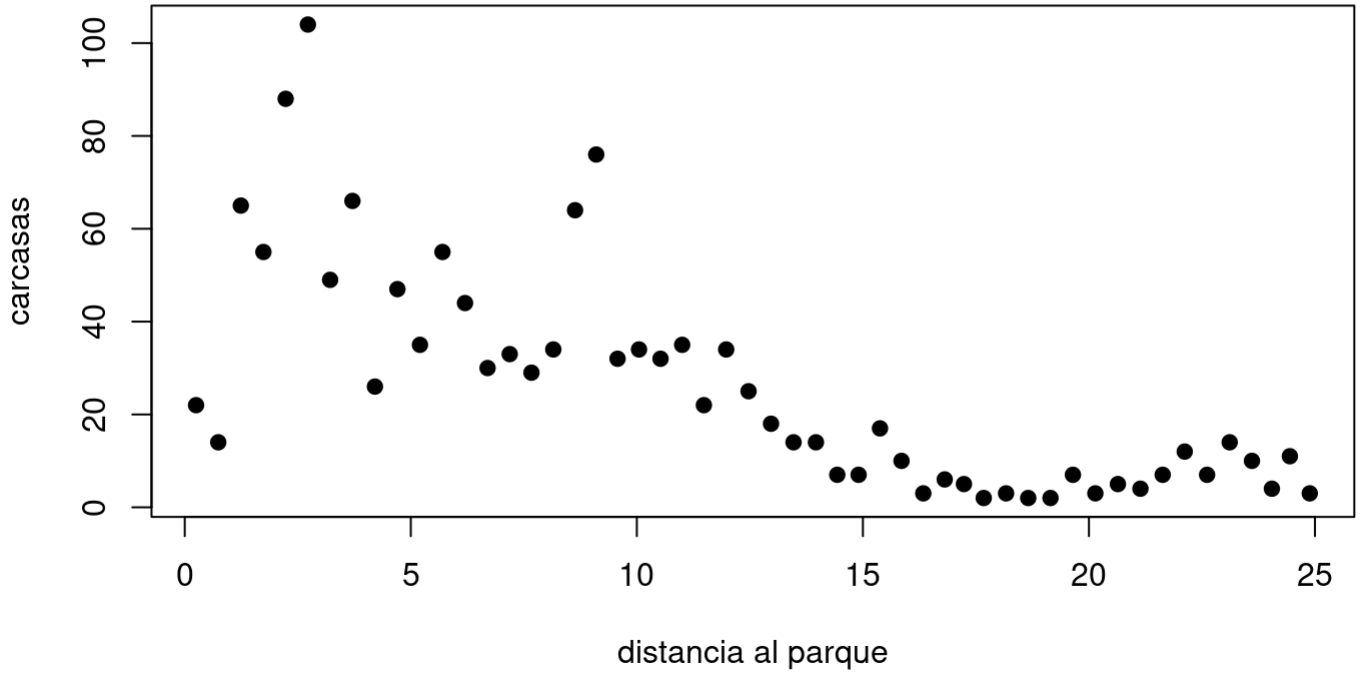
$$log(\mu_i) = \eta$$

Por lo que el valor esperado es …

$$\mu_i = e^{\eta}$$

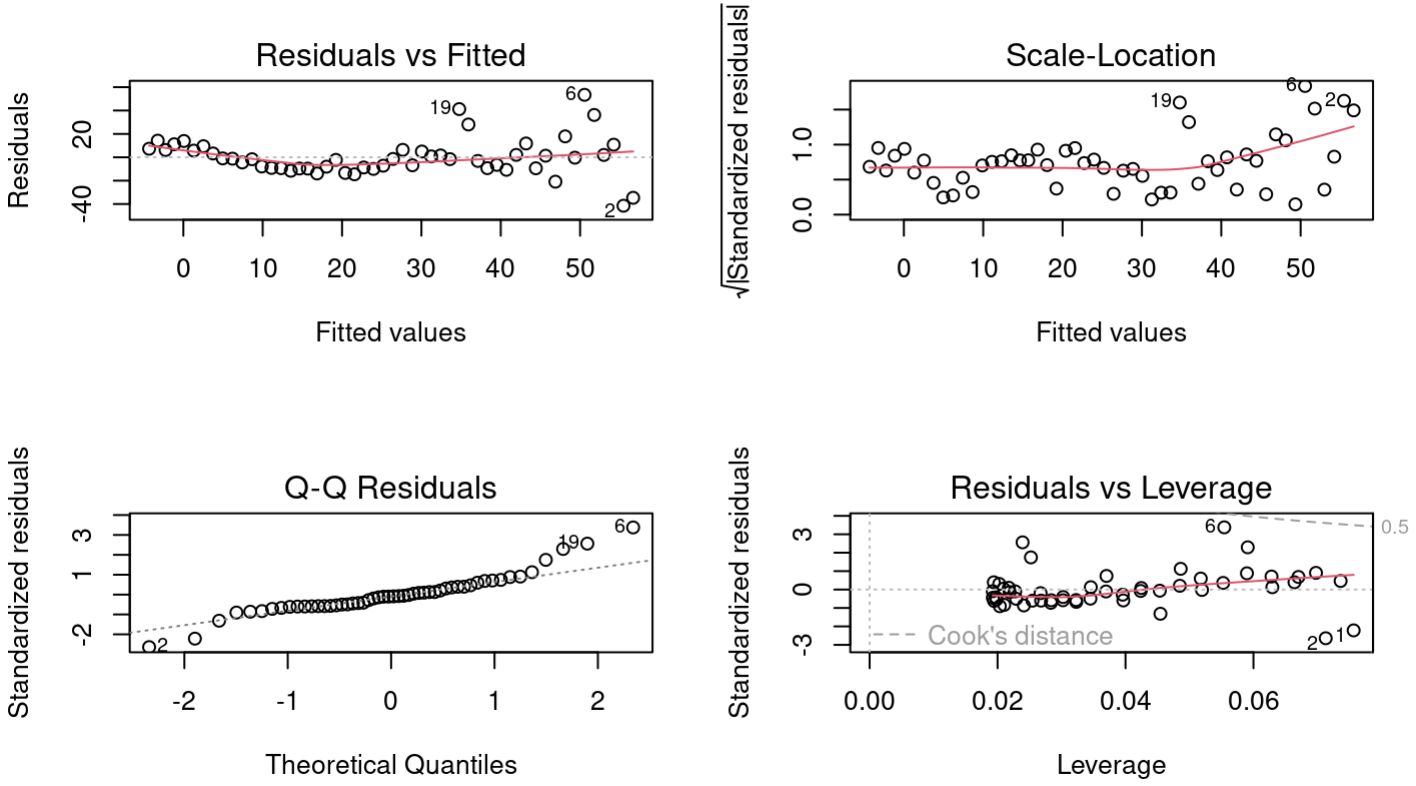
Colocando las variables predictoras en la parte sistemática…

$$\mu_i = e^{\beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki}}$$


carcasas
0
20
40
60
80
100
0
5
10
15
20
25
distancia al parque

```
Call:
lm(formula = TOT.N ~ D.PARK, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-41.470  -9.316  -1.511   6.269  53.441

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  57.3049     4.5479  12.600  < 2e-16 ***
D.PARK       -2.4763     0.3113  -7.955 1.95e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.29 on 50 degrees of freedom
Multiple R-squared:  0.5586,    Adjusted R-squared:  0.5498
F-statistic: 63.28 on 1 and 50 DF,  p-value: 1.952e-10
```

Los horribles diagnósticos de un modelo lineal general



```
fit <- glm(TOT.N ~ D.PARK, data = dat, family = poisson); summary(fit)


Call:
glm(formula = TOT.N ~ D.PARK, family = poisson, data = dat)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.316485   0.043220   99.87   <2e-16 ***
D.PARK      -0.105851   0.004387  -24.13   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1071.4  on 51  degrees of freedom
Residual deviance:  390.9  on 50  degrees of freedom
AIC: 634.29

Number of Fisher Scoring iterations: 4
```

¿Es significativa la devianza residual?

Esto indicaría un mal ajuste del modelo, si bien es un test vago.

```
pchisq(390.9, 50, lower.tail=F)
```

```
[1] 2.321894e-54
```

```
fit0 <- glm(TOT.N ~ 1, data = dat, family = poisson)
fit1 <- glm(TOT.N ~ D.PARK, data = dat, family = poisson)
anova(fit0, fit1, test = "Chisq")
```

```
Analysis of Deviance Table

Model 1: TOT.N ~ 1
Model 2: TOT.N ~ D.PARK
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1        51     1071.4
2        50      390.9  1   680.55 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La diferencia de devianzas se distribuye como χ^2 con sus grados de libertad igual a la diferencia de grados de libertad entre los modelos.

Interpretación de parámetros

```
fit$coefficients
```

```
(Intercept)      D.PARK
  4.3164850  -0.1058506
```

```
exp(fit$coefficients)
```

```
(Intercept)      D.PARK
  74.924803    0.899559
```

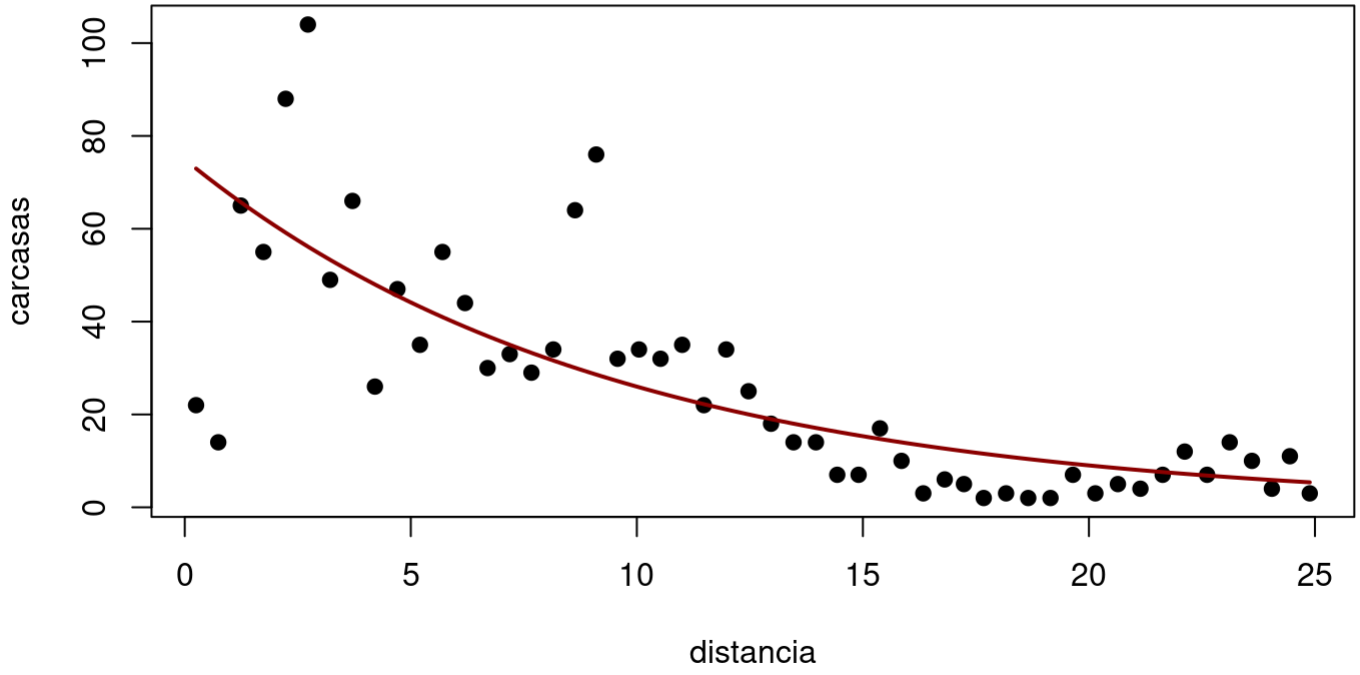
```
1/exp(fit$coefficients)
```

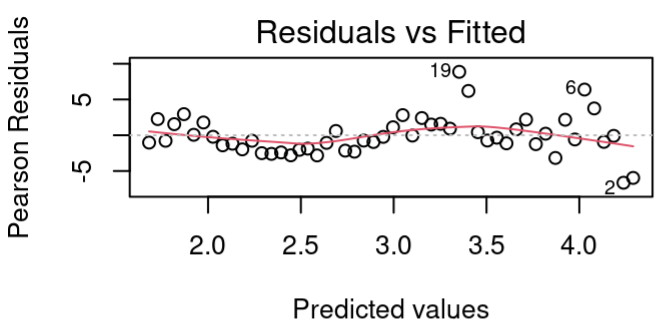
```
(Intercept)      D.PARK
 0.01334672  1.11165580
```

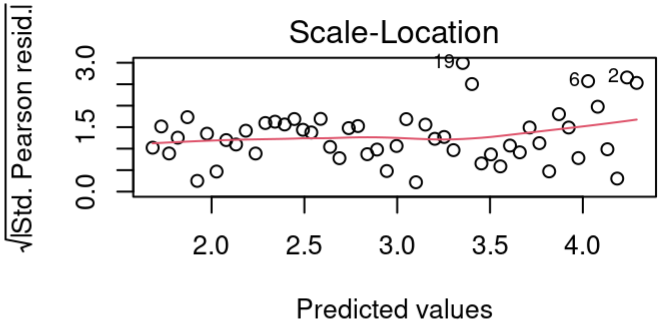
El número de muertos disminuye un 11.17% por cada km de distancia al parque.

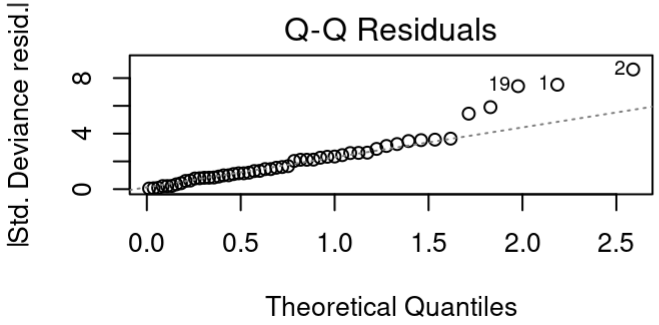
Interpretación de parámetros

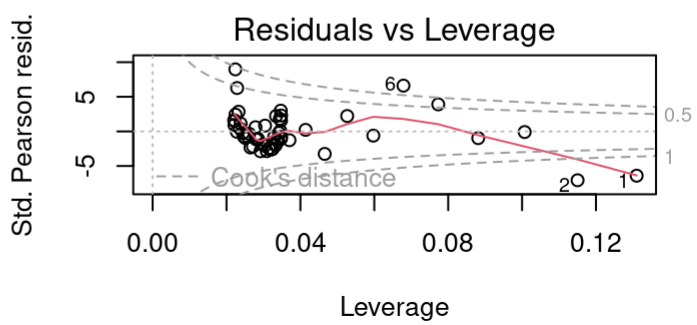
- Si β es cercano a 0, exp(β) es cercano a 1. El efecto es pequeño.
- Si β es 0.13976, exp(β) es 1.15. Por cada aumento de una unidad en la variable X, la variable Y se incrementa un 15%.
- Si β es -0.2231, exp(β) es 0.80, 1/exp(β) es 1.25. Por cada disminución de una unidad en la variable X, la variable Y disminuye un 25%.

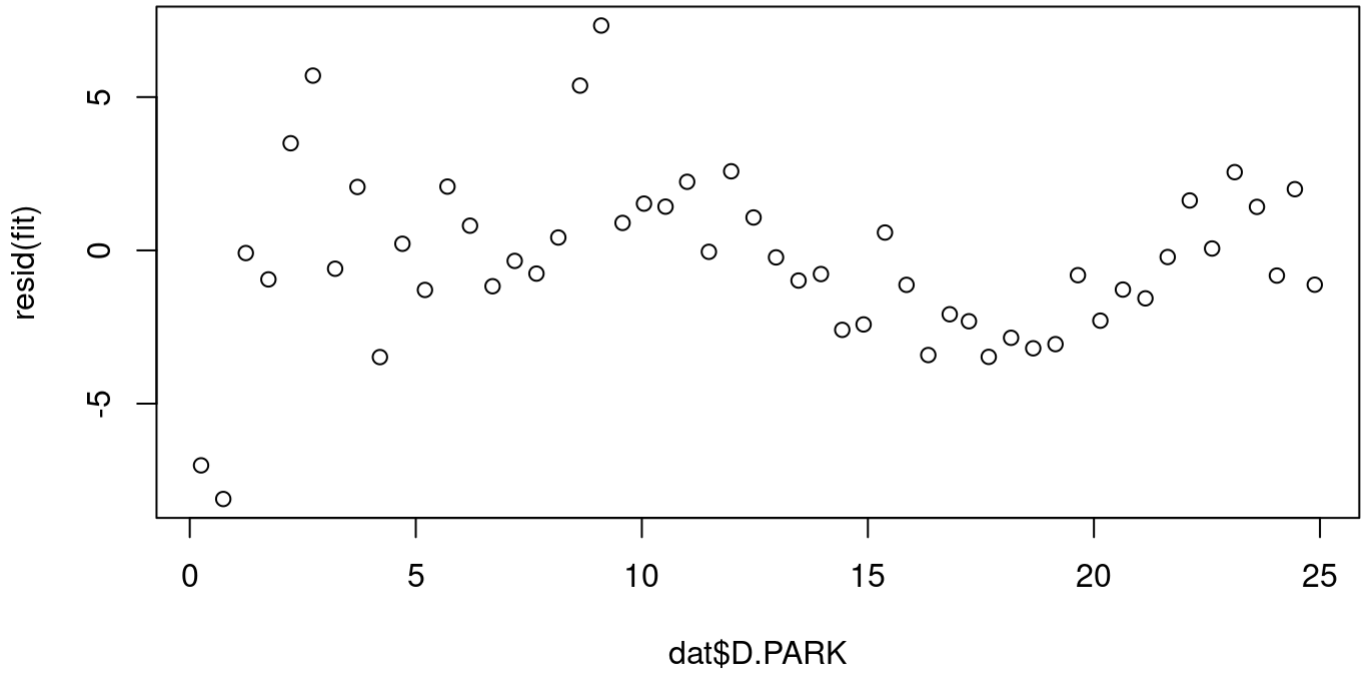

carcasas
0
20
40
60
80
100
0
5
10
15
20
25
distancia


Residuals vs Fitted
Pearson Residuals
Predicted values
19
6
2


Scale-Location
√|Std. Pearson resid.|
Predicted values
19
6
2


Q-Q Residuals
|Std. Deviance resid.|
Theoretical Quantiles
19
1
2


Residuals vs Leverage
Std. Pearson resid.
Leverage
Cook's distance
0.5
1
6
2
1



Ver la asociación entre los residuos y las variables independientes. (notar que existe más de un tipo de residuos)

Sobredispersión

- La sobredispersión es provocada por variación al azar **NO EXPLICADA** en la variable respuesta.
- Habitualmente se observa una importante devianza residual, que permanece incluso si se incorporan variables al modelo.
- Esta variabilidad provoca que la relación media-varianza esperada para la familia no se cumpla.
- Debe distinguirse de la sobredispersión aparente , debida a variables o interacciones faltantes, efectos no lineales no considerados, outliers en la variable respuesta o errores en la elección del enlace.

Uno de los supuestos de un MLG es que la variable dependiente se ajusta a una de las funciones de la familia exponencial. Cada una de estas familias está caracterizada por una relación media-varianza específica.

Distribución	Posición	Dispersión
Poisson	$E(Y) = \mu$	$var(Y) = \mu$
Binomial	$E(Y) = N\pi$	$var(Y) = N\pi(1 - \pi)$

Sin embargo, a veces la variabilidad observada es mayor en una proporción ϕ respecto a lo esperado. Si $\phi > 1$ entonces existe sobredispersión

Distribución	Posición	Dispersión
Poisson	$E(Y) = \mu$	$var(Y) = \phi\mu$
Binomial	$E(Y) = N\pi$	$var(Y) = \phi N\pi(1-\pi)$

La sobredispersión provoca que los errores estándar de los parámetros sean la raíz cuadrada de ϕ veces más chicos que lo que deberían (y aumenta por tanto el error de tipo I)

JAMÁS pueden presentar sobredispersión:

1- Datos binarios (Binomiales no agregados: 0 y 1).
2- Cuando el modelo en consideración es el saturado (ej. modelos loglineales para tablas de contingencia).

Si nuestro modelo presenta un buen ajuste y aún así permanece mucha devianza residual no explicada debemos estimar ϕ para probar si existe sobredispersión.

- El parámetro ϕ es aproximadamente igual a la suma de los residuos Pearson sobre sus grados de libertad (o a la devianza residual sobre los grados de libertad). Para evitar hacer la cuenta podemos ajustar un modelo de cuasiverosimilitud (*quasibinomial* o *quasipoisson* en R)
- En caso de ser significativa la sobre dispersión, nuestro modelo final será el de quasiverosimilitud.

¿Cuál es la magnitud de sobredispersión tolerable? Podemos usar una aproximación práctica (examinar si los valores *P* cambian significativamente), o usar el límite arbitrario de $\phi > 2$.

- No existen la “familia quasipoisson” o “quasibinomial”, sino que especificamos una relación media-varianza y ajustamos el modelo mediante cuasiverosimilitud.
- Ventaja: Controlan la sobredispersión, por lo que disminuyen el error del valor P, los valores estimados de los parámetros (y por lo tanto la interpretación biológica) no cambian.
- Desventaja: Dificultan la selección e modelos, se utiliza QAIC en vez de AIC.

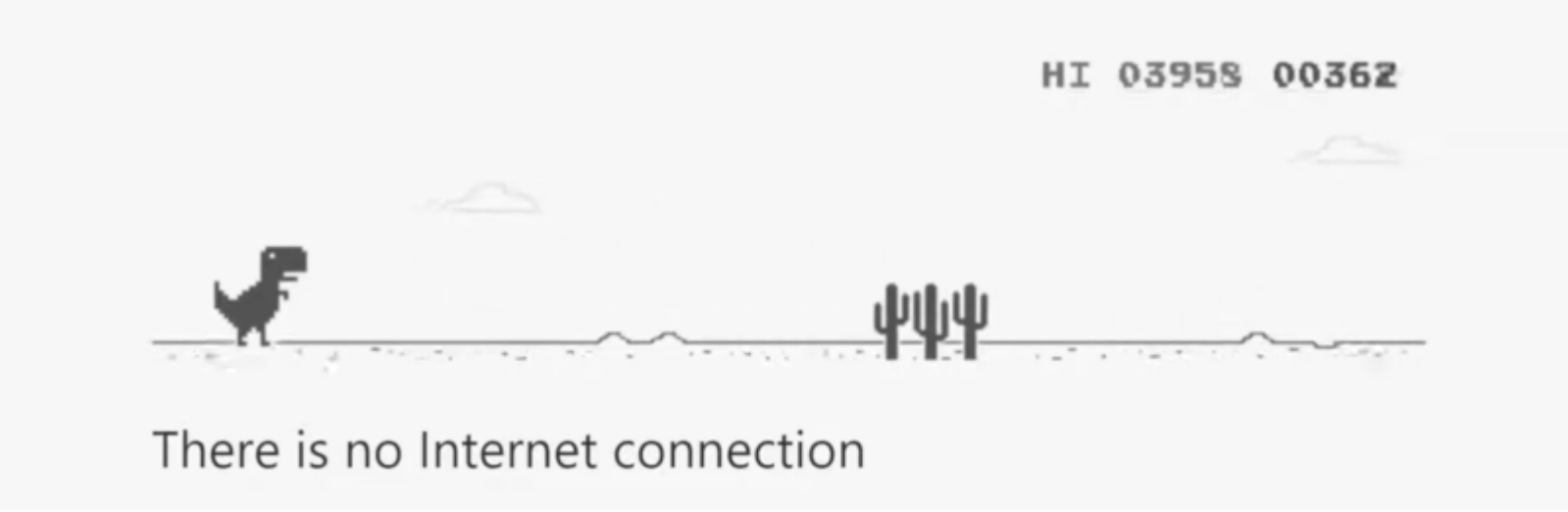
Modelos Binomiales Negativos.

Distribución	Posición	Dispersión
Poisson	$E(Y) = \mu$	$var(Y) = \mu$
Binomial Negativa	$E(Y) = \mu$	$var(Y) = \mu + \theta\mu^2$

El parámetro θ regula la sobredispersión.

La distribución binomial negativa es la distribución de los experimentos de Bernoulli independientes hasta la consecución del éxito.

Volviendo… https://www.youtube.com/watch?v=TktiBhPdZYE

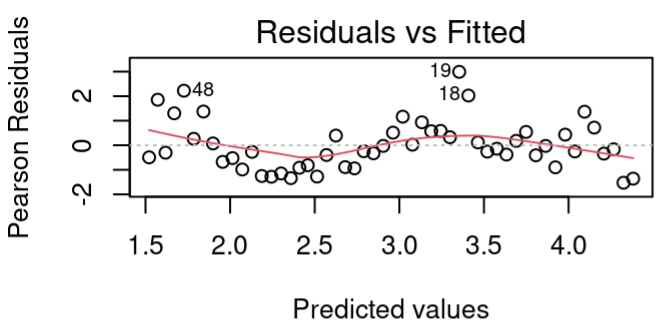


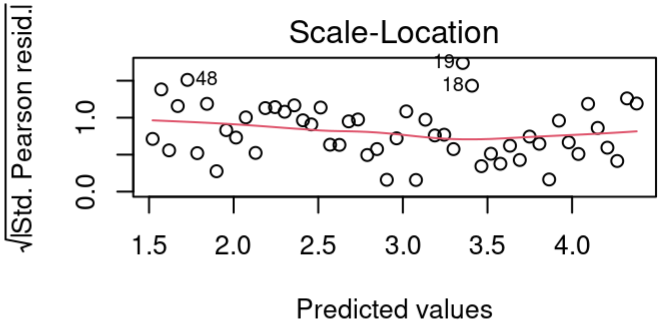
```
library(MASS)
nbfit <- glm.nb(TOT.N ~ D.PARK, data=dat)
summary(nbfit)


Call:
glm.nb(formula = TOT.N ~ D.PARK, data = dat, init.theta = 3.681040094,
    link = log)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.41072    0.15476   28.50   <2e-16 ***
D.PARK      -0.11612    0.01137  -10.21   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(3.681) family taken to be 1)

    Null deviance: 155.445  on 51  degrees of freedom
Residual deviance:  54.742  on 50  degrees of freedom
AIC: 393.09

Number of Fisher Scoring iterations: 1


              Theta:  3.681
          Std. Err.:  0.891
```

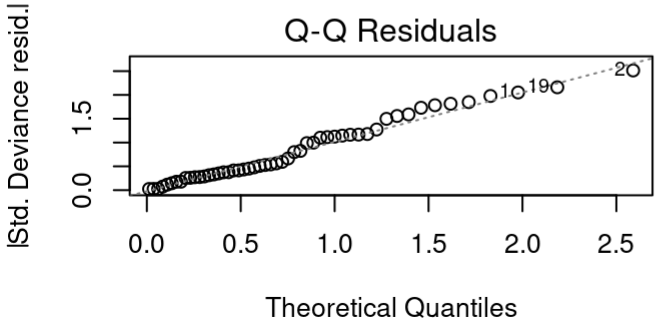
```
anova(nbfit, test = "Chisq")
```

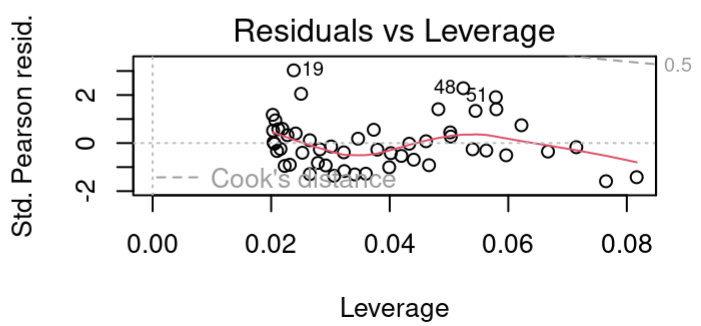
```
Warning in anova.negbin(nbfit, test = "Chisq"): tests made without
re-estimating 'theta'
```

```
Analysis of Deviance Table

Model: Negative Binomial(3.681), link: log

Response: TOT.N

Terms added sequentially (first to last)

       Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                      51    155.445
D.PARK  1    100.7        50     54.742 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Residuals vs Fitted
Pearson Residuals
Predicted values
48
19
18
-2
0
2
1.5
2.0
2.5
3.0
3.5
4.0


Scale-Location
√|Std. Pearson resid.|
Predicted values
48
19
18
0.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0


Q-Q Residuals
|Std. Deviance resid.|
Theoretical Quantiles
1
19
20
0.0
1.5
0.0
0.5
1.0
1.5
2.0
2.5

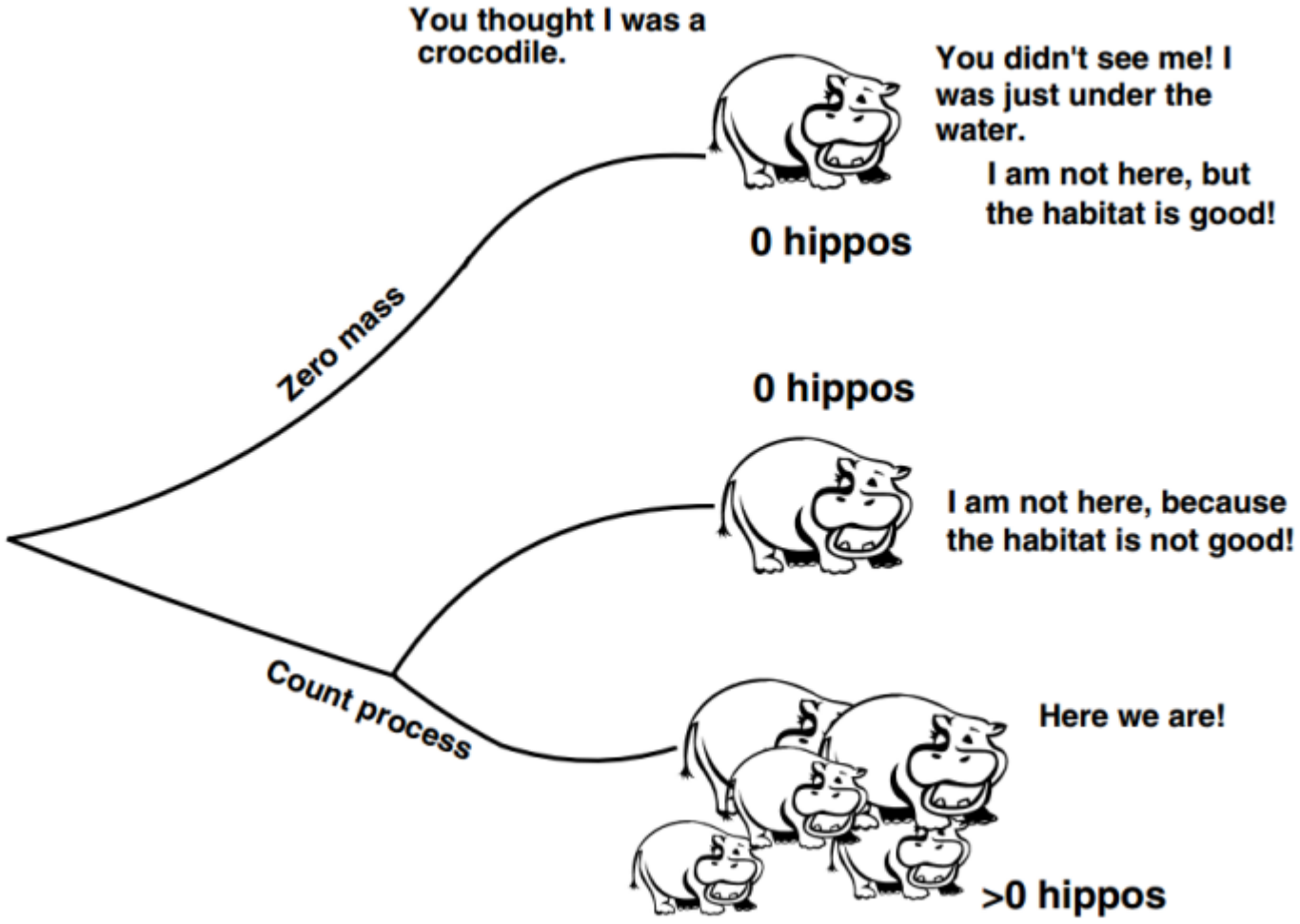

Residuals vs Leverage
Std. Pearson resid.
Leverage
Cook's distance
0.5
19
48
51
-2
0
2
0.00
0.02
0.04
0.06
0.08

- Los modelos con la familia binomial negativa pueden estar sobredispersos (examinar la devianza residual). Pero no existen métodos de cuasi-verosimilitud para ellos.
- La familia binomial negativa permite calcular el AIC, lo que facilita la elección de modelos.
- Los modelos poisson están anidados dentro de un modelo binomial negativo, ya que solamente difieren en un parámetro (θ). Puede testearse su adecuación con una cociente de verosimilitudes.

```
library(MASS)
nbfit <- glm.nb(TOT.N ~ D.PARK, data = dat)
pofit <- glm(TOT.N ~ D.PARK, data = dat, family = poisson)
L.NB = logLik(nbfit)
L.Po = logLik(pofit)
d <- 2 * (L.NB - L.Po); attributes(d) <- NULL
pchisq(d, df = 1, lower.tail = FALSE)/2
```

```
[1] 3.956301e-55
```

Poisson inflado en ceros (ZIP)



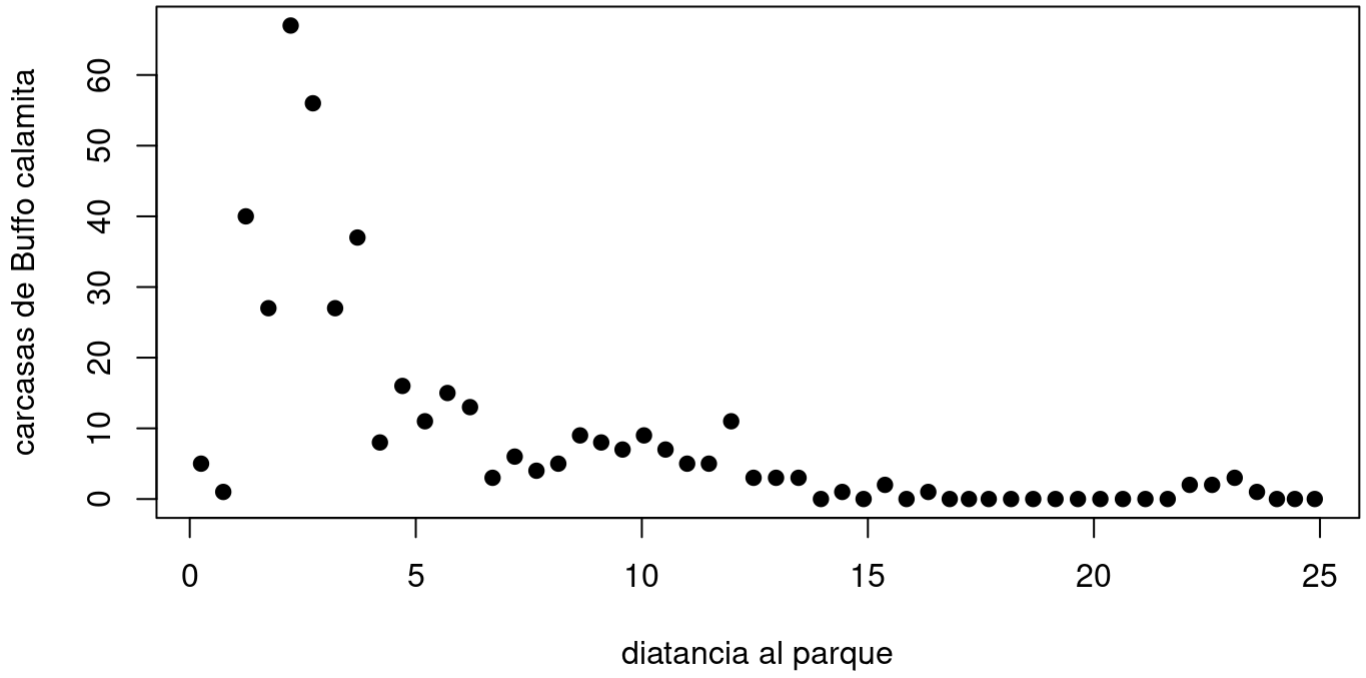
* de Zuur et al. 2009

$$P(y_i = 0) = P(0_{falso}) + P(0_{verdadero})$$

$$P(y_i = 0) = P(0_{falso}) + (1 - P(0_{falso})) * P(conteo = 0)$$

$$P(y_i \neq 0) = (1 - P(0_{falso})) * P(conteo \neq 0)$$

proceso **BINOMIAL** regula la aparición de ceros falsos
proceso **POISSON** (o **BN**) regula el conteo (que incluye ceros)


carcasas de Buffo calamita
0
10
20
30
40
50
60
0
5
10
15
20
25
diatancia al parque

```
library(pscl)
zip1 <- zeroinfl(BufoCalamita~D.PARK, dist = "poisson",
                 link = "logit", data = dat)
summary(zip1)
```

```
Call:
zeroinfl(formula = BufoCalamita ~ D.PARK, data = dat, dist = "poisson",
    link = "logit")

Pearson residuals:
    Min      1Q  Median      3Q     Max
-4.8872 -0.7967 -0.2920  0.2498  7.0707

Count model coefficients (poisson with log link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.63582    0.07706   47.18   <2e-16 ***
D.PARK      -0.17088    0.01288  -13.27   <2e-16 ***

Zero-inflation model coefficients (binomial with logit link):
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.88051    1.44747  -3.372 0.000747 ***
D.PARK       0.25622    0.08647   2.963 0.003043 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of iterations in BFGS optimization: 9
Log-likelihood: -202.1 on 4 Df
```

END