Regresión Binomial Negativa

DATOS: Galapagos Islands dataset

Hay 30 inslas en Galapagos. Se estudia la relación entre el número de especies de plantas y varias variables geográficas de interés.

El conjunto de datos contiene las siguientes variables:

Species: el número de especies que se encuentrasn en las islas.

Endemics: número de especies endémicas

Area: el área de la isla (km^2)

Elevation: la mayor elevación de la isla(m)

Nearest: la distancia a la isla más cercana (km)

Scruz: la distancia a la isla Santa Cruz(km)

Adjacent: el área de la isla adyacente (km^2)

```
library(faraway)
library(tidyverse)
```

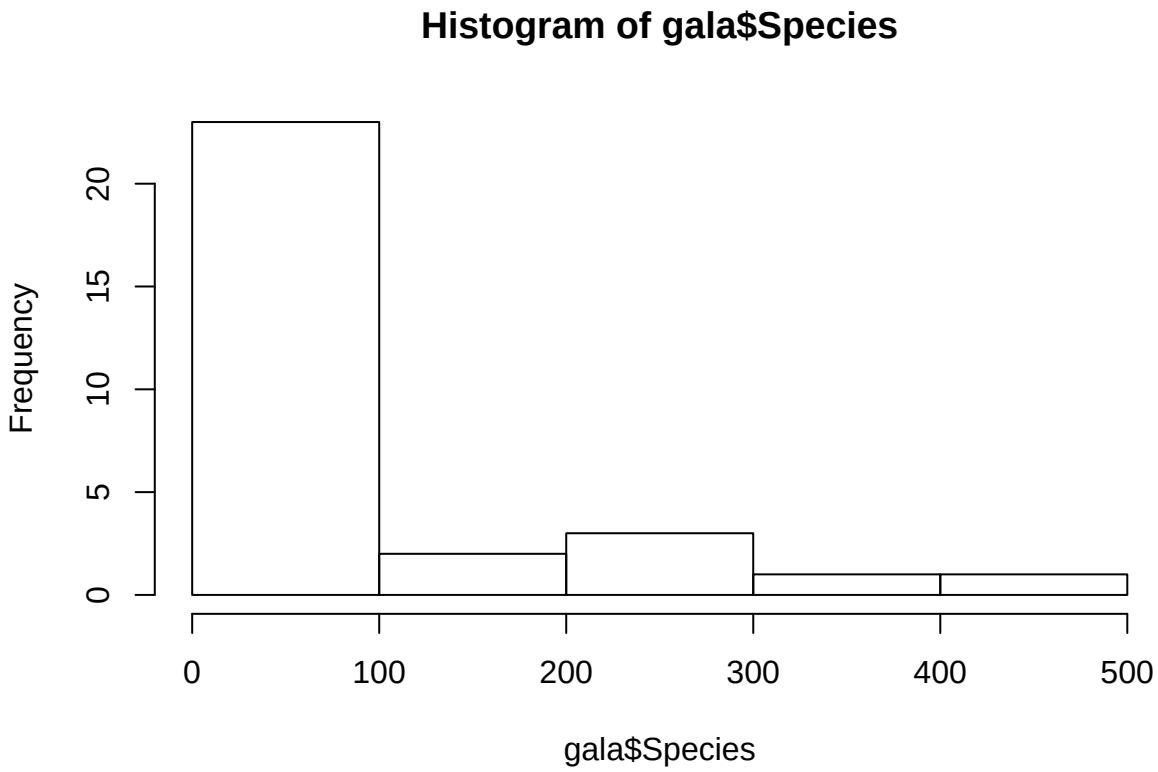
```
## -- Attaching packages --------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.2.0     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(visreg)
library(broom)

# histograma del número de especies en Galapagos
hist(gala$Species,breaks=6)
```




```
#?gala
gala <- faraway::gala %>%
  as_tibble() %>%
  mutate(Island = rownames(faraway::gala))
gala
```

```
## # A tibble: 30 x 8
##    Species Endemics  Area Elevation Nearest Scruz Adjacent Island
##      <dbl>    <dbl> <dbl>     <dbl>   <dbl> <dbl>    <dbl> <chr>
##  1      58       23 25.1        346     0.6   0.6     1.84 Baltra
##  2      31       21  1.24       109     0.6  26.3   572.   Bartolome
##  3       3        3  0.21       114     2.8  58.7     0.78 Caldwell
##  4      25        9  0.1         46     1.9  47.4     0.18 Champion
##  5       2        1  0.05        77     1.9   1.9   904.   Coamano
##  6      18       11  0.34       119     8     8       1.84 Daphne.Major
##  7      24        0  0.08        93     6    12       0.34 Daphne.Minor
##  8      10        7  2.33       168    34.1 290.      2.85 Darwin
##  9       8        4  0.03        71     0.4   0.4    18.0  Eden
## 10       2        2  0.18       112     2.6  50.2     0.1  Enderby
## # ... with 20 more rows
```

MLG POISSON

```
modp <- glm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
            family=poisson, gala)
summary(modp)
```

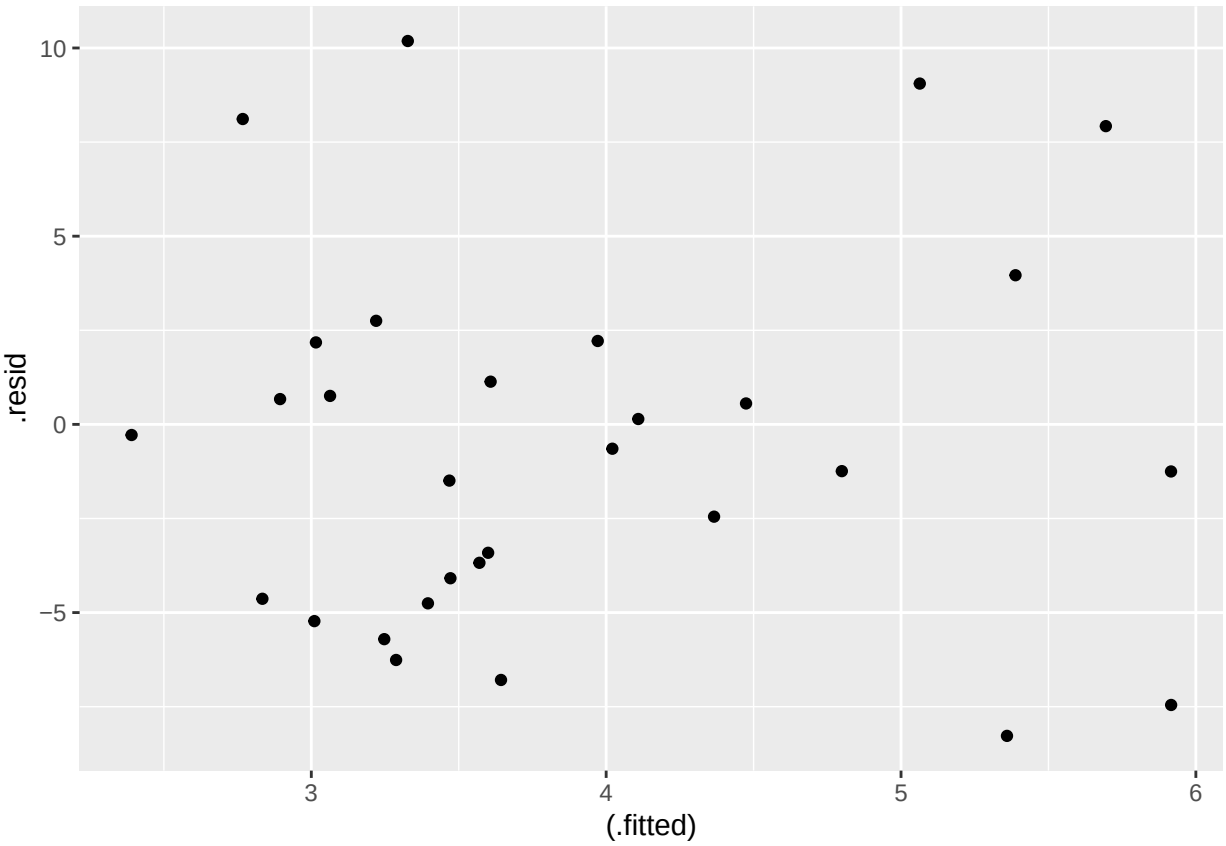
```
##
## Call:
```

```
## glm(formula = Species ~ Area + Elevation + Nearest + Scruz +
##     Adjacent, family = poisson, data = gala)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.2752  -4.4966  -0.9443   1.9168  10.1849
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.155e+00  5.175e-02  60.963  < 2e-16 ***
## Area        -5.799e-04  2.627e-05 -22.074  < 2e-16 ***
## Elevation    3.541e-03  8.741e-05  40.507  < 2e-16 ***
## Nearest      8.826e-03  1.821e-03   4.846 1.26e-06 ***
## Scruz       -5.709e-03  6.256e-04  -9.126  < 2e-16 ***
## Adjacent    -6.630e-04  2.933e-05 -22.608  < 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: 3510.73  on 29  degrees of freedom
## Residual deviance:  716.85  on 24  degrees of freedom
## AIC: 889.68
##
## Number of Fisher Scoring iterations: 5
```

```
1- pchisq(deviance(modp), df.residual(modp), lower.tail=TRUE)
```

```
## [1] 0
```

```
augment(modp) %>%
  ggplot(aes(x=(.fitted), y=.resid)) +
  geom_point()
```



El modelo no tiene buen ajuste a los datos.

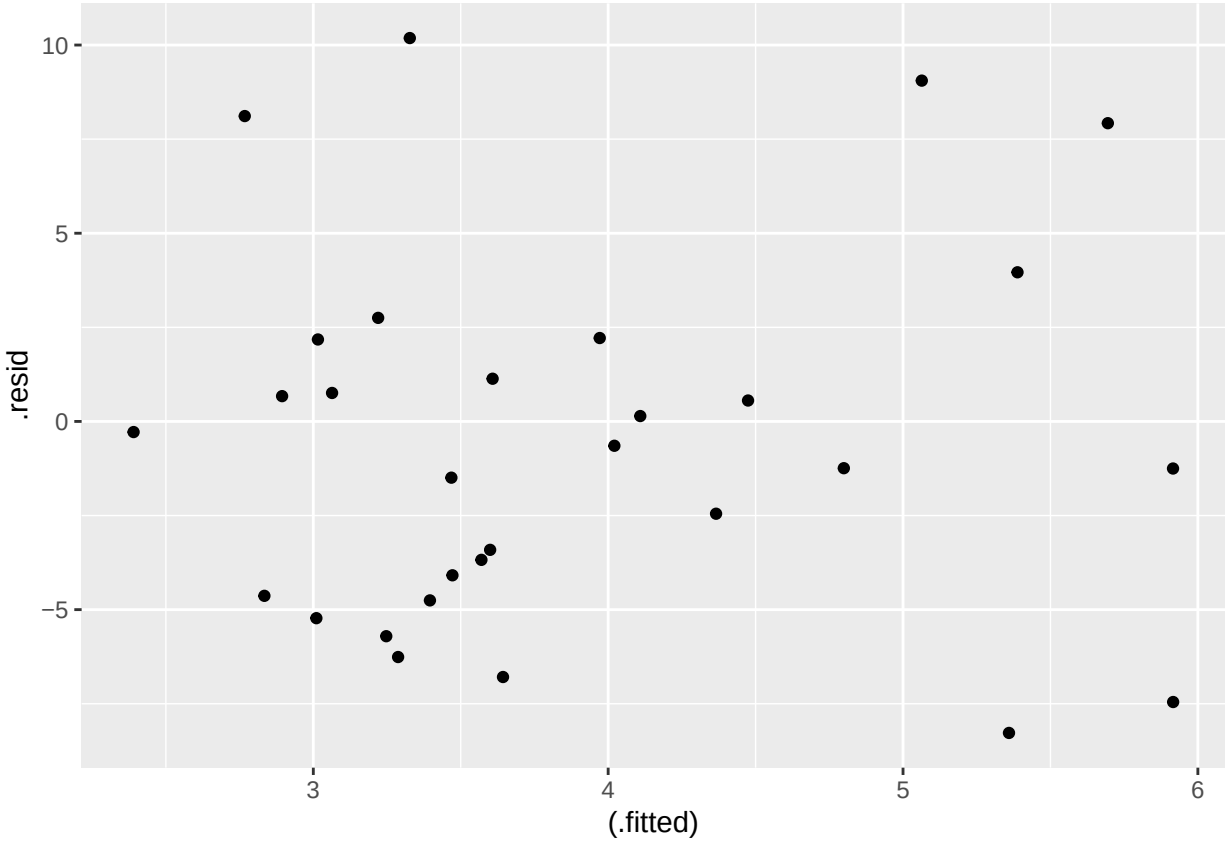
MLG QUASI-POISSON

```
pm2 <- glm(Species ~ Area + Elevation + Nearest + Scruz + Adjacent, family = quasipoisson,
    data = gala)
summary(pm2)
```

```
##
## Call:
## glm(formula = Species ~ Area + Elevation + Nearest + Scruz +
##     Adjacent, family = quasipoisson, data = gala)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -8.2752  -4.4966  -0.9443   1.9168  10.1849
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.1548079  0.2915901  10.819 1.03e-10 ***
## Area        -0.0005799  0.0001480  -3.918 0.000649 ***
## Elevation    0.0035406  0.0004925   7.189 1.98e-07 ***
## Nearest      0.0088256  0.0102622   0.860 0.398292
## Scruz       -0.0057094  0.0035251  -1.620 0.118380
## Adjacent    -0.0006630  0.0001653  -4.012 0.000511 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 31.74921)
##
##     Null deviance: 3510.73  on 29  degrees of freedom
## Residual deviance:  716.85  on 24  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

```
augment(modp) %>%
  ggplot(aes(x=(.fitted), y=.resid)) +
  geom_point()
```



El parámetro de dispersión $\phi = 31.75 >> 1$!! Existe sobredispersión en los datos.

Modelo de regresión binomial negativa

```
modnb <- MASS::glm.nb(Species ~ Area + Elevation + Nearest + Scruz + Adjacent,
                      gala)
summary(modnb)
```

```
##
## Call:
## MASS::glm.nb(formula = Species ~ Area + Elevation + Nearest +
##     Scruz + Adjacent, data = gala, init.theta = 1.674602286,
##     link = log)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1344  -0.8597  -0.1476   0.4576   1.8416
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.9065247  0.2510344  11.578  < 2e-16 ***
## Area        -0.0006336  0.0002865  -2.211 0.027009 *
## Elevation    0.0038551  0.0006916   5.574 2.49e-08 ***
## Nearest      0.0028264  0.0136618   0.207 0.836100
## Scruz       -0.0018976  0.0028096  -0.675 0.499426
## Adjacent    -0.0007605  0.0002278  -3.338 0.000842 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.6746) family taken to be 1)
##
##     Null deviance: 88.431  on 29  degrees of freedom
## Residual deviance: 33.196  on 24  degrees of freedom
## AIC: 304.22
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.675
##           Std. Err.:  0.442
##
##  2 x log-likelihood:  -290.223
```

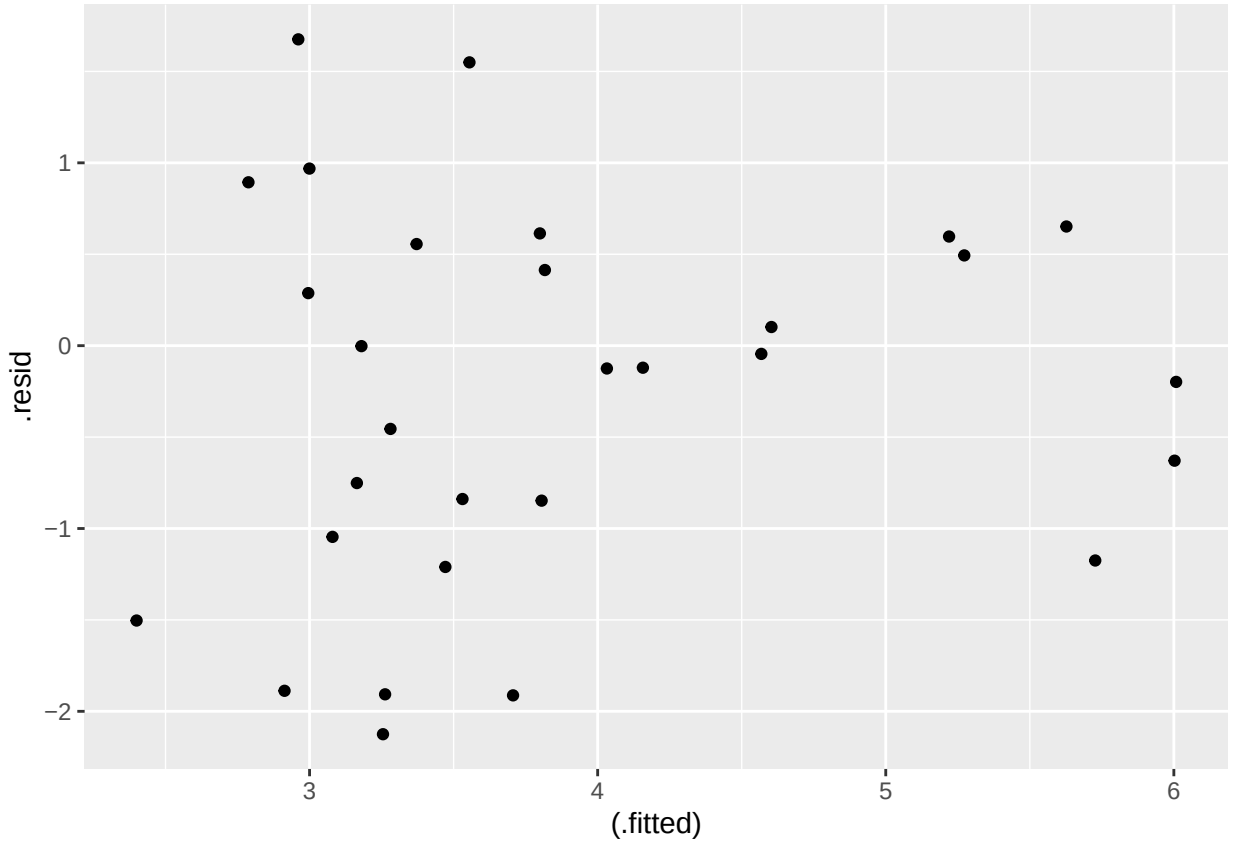
Modelo de regresión binomial negativa reducido (con variables explicativas significativas. . .)

```
modnb1 <- MASS::glm.nb(Species ~ (Area) + (Elevation)  + (Adjacent),
                       gala)
summary(modnb1)
```

```
##
## Call:
## MASS::glm.nb(formula = Species ~ (Area) + (Elevation) + (Adjacent),
##     data = gala, init.theta = 1.651522946, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1251  -0.9963  -0.1226   0.5403   1.6755
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.8148989  0.2231456  12.615  < 2e-16 ***
## Area        -0.0006449  0.0002804  -2.300 0.021459 *
## Elevation    0.0039299  0.0006761   5.812 6.16e-09 ***
## Adjacent    -0.0007943  0.0002196  -3.616 0.000299 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.6515) family taken to be 1)
##
##     Null deviance: 87.279  on 29  degrees of freedom
## Residual deviance: 33.155  on 26  degrees of freedom
## AIC: 300.59
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.652
##           Std. Err.:  0.434
##
##  2 x log-likelihood:  -290.593
```

```
augment(modnb1) %>%
  ggplot(aes(x=(.fitted), y=.resid)) +
  geom_point()
```



El parámetro de dispersión = Theta = 1.6515= 1.652 es el κ en nuestra definición.

Interpretación: elevation

exp(0.0039299)= 1.003938

La media esperada del número de especies de plantas encontradas en la isla se incrementa en (0.003)*100% =

0.3% por cada m adicional de altitud (elevation), (manteniendo las otras variables explicativas constantes).

Quasi-Gamma

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
gm0 <- glm(Days + 0.1 ~ Age*Eth*Sex*Lrn,family=Gamma(link=log), data=quine,
          start = c(3, rep(0,31)))

summary(gm0)
```

```
##
## Call:
## glm(formula = Days + 0.1 ~ Age * Eth * Sex * Lrn, family = Gamma(link = log),
##     data = quine, start = c(3, rep(0, 31)))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0385  -0.7164  -0.1532   0.3863   1.3087
##
## Coefficients: (4 not defined because of singularities)
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.06105    0.39162   7.816 2.52e-12 ***
## AgeF1                 -0.61870    0.52541  -1.178 0.241343
## AgeF2                 -2.31911    0.87569  -2.648 0.009196 **
## AgeF3                 -0.37623    0.47067  -0.799 0.425690
## EthN                  -0.13789    0.55384  -0.249 0.803814
## SexM                  -0.48844    0.52541  -0.930 0.354462
## LrnSL                 -1.92965    0.87569  -2.204 0.029496 *
## AgeF1:EthN             0.10249    0.72916   0.141 0.888460
## AgeF2:EthN            -0.50874    1.23841  -0.411 0.681966
## AgeF3:EthN             0.06314    0.66049   0.096 0.924003
## AgeF1:SexM             0.40695    0.83993   0.484 0.628930
## AgeF2:SexM             3.06173    0.98852   3.097 0.002441 **
## AgeF3:SexM             1.10841    0.65716   1.687 0.094310 .
## EthN:SexM             -0.74217    0.72916  -1.018 0.310834
## AgeF1:LrnSL            2.60967    0.97513   2.676 0.008505 **
## AgeF2:LrnSL            4.78434    1.20706   3.964 0.000127 ***
## AgeF3:LrnSL                 NA         NA      NA       NA
## EthN:LrnSL             2.22936    1.23841   1.800 0.074388 .
## SexM:LrnSL             1.56531    1.04595   1.497 0.137182
## AgeF1:EthN:SexM       -0.30235    1.17050  -0.258 0.796620
## AgeF2:EthN:SexM        0.29742    1.39064   0.214 0.831014
## AgeF3:EthN:SexM        0.82215    0.91458   0.899 0.370517
## AgeF1:EthN:LrnSL      -3.50803    1.36957  -2.561 0.011685 *
## AgeF2:EthN:LrnSL      -3.33529    1.70454  -1.957 0.052744 .
## AgeF3:EthN:LrnSL            NA         NA      NA       NA
```

```
## AgeF1:SexM:LrnSL      -2.39791    1.33764  -1.793 0.075592 .
## AgeF2:SexM:LrnSL      -4.12161    1.42308  -2.896 0.004502 **
## AgeF3:SexM:LrnSL            NA         NA      NA       NA
## EthN:SexM:LrnSL       -0.15305    1.47227  -0.104 0.917384
## AgeF1:EthN:SexM:LrnSL  2.13480    1.84803   1.155 0.250352
## AgeF2:EthN:SexM:LrnSL  2.11886    2.01804   1.050 0.295883
## AgeF3:EthN:SexM:LrnSL       NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Gamma family taken to be 0.6134666)
##
##     Null deviance: 190.40  on 145  degrees of freedom
## Residual deviance: 128.36  on 118  degrees of freedom
## AIC: 1103.7
##
## Number of Fisher Scoring iterations: 7
```

```
gm <- glm(Days + 0.1 ~ Age*Eth*Sex*Lrn,
          quasi(link=log, variance="mu^2"), data=quine,
          start = c(3, rep(0,31)))

summary(gm)
```

```
##
## Call:
## glm(formula = Days + 0.1 ~ Age * Eth * Sex * Lrn, family = quasi(link = log,
##     variance = "mu^2"), data = quine, start = c(3, rep(0, 31)))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.0385  -0.7164  -0.1532   0.3863   1.3087
##
## Coefficients: (4 not defined because of singularities)
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.06105    0.39162   7.816 2.52e-12 ***
## AgeF1                 -0.61870    0.52541  -1.178 0.241343
## AgeF2                 -2.31911    0.87569  -2.648 0.009196 **
## AgeF3                 -0.37623    0.47067  -0.799 0.425690
## EthN                  -0.13789    0.55384  -0.249 0.803814
## SexM                  -0.48844    0.52541  -0.930 0.354462
## LrnSL                 -1.92965    0.87569  -2.204 0.029496 *
## AgeF1:EthN             0.10249    0.72916   0.141 0.888460
## AgeF2:EthN            -0.50874    1.23841  -0.411 0.681966
## AgeF3:EthN             0.06314    0.66049   0.096 0.924003
## AgeF1:SexM             0.40695    0.83993   0.484 0.628930
## AgeF2:SexM             3.06173    0.98852   3.097 0.002441 **
## AgeF3:SexM             1.10841    0.65716   1.687 0.094310 .
## EthN:SexM             -0.74217    0.72916  -1.018 0.310834
## AgeF1:LrnSL            2.60967    0.97513   2.676 0.008505 **
## AgeF2:LrnSL            4.78434    1.20706   3.964 0.000127 ***
## AgeF3:LrnSL                 NA         NA      NA       NA
## EthN:LrnSL             2.22936    1.23841   1.800 0.074388 .
## SexM:LrnSL             1.56531    1.04595   1.497 0.137182
## AgeF1:EthN:SexM       -0.30235    1.17050  -0.258 0.796620
```

```
## AgeF2:EthN:SexM         0.29742    1.39064   0.214 0.831014
## AgeF3:EthN:SexM         0.82215    0.91458   0.899 0.370517
## AgeF1:EthN:LrnSL       -3.50803    1.36957  -2.561 0.011685 *
## AgeF2:EthN:LrnSL       -3.33529    1.70454  -1.957 0.052744 .
## AgeF3:EthN:LrnSL             NA         NA      NA       NA
## AgeF1:SexM:LrnSL       -2.39791    1.33764  -1.793 0.075592 .
## AgeF2:SexM:LrnSL       -4.12161    1.42308  -2.896 0.004502 **
## AgeF3:SexM:LrnSL             NA         NA      NA       NA
## EthN:SexM:LrnSL        -0.15305    1.47227  -0.104 0.917384
## AgeF1:EthN:SexM:LrnSL   2.13480    1.84803   1.155 0.250352
## AgeF2:EthN:SexM:LrnSL   2.11886    2.01804   1.050 0.295883
## AgeF3:EthN:SexM:LrnSL        NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasi family taken to be 0.6134666)
##
##     Null deviance: 190.40  on 145  degrees of freedom
## Residual deviance: 128.36  on 118  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 7
```