Models for count data

QERM 514 - Homework 8 Answer key

22 May 2020

R Markdown file

You can find the R Markdown file used to create this answer key here.

Background

This week's homework assignment focuses on fitting and evaluating models for count data. One of your colleagues is interested in the theory of island biogeography and has acquired a data set with which to examine how species richness varies with the area of an island, the island's elevation, the distance to the nearest island, and the area of the nearest island. In particular, her expectation is that the number of plant species should increase with island area, and a plot of the data suggests this to indeed be the case, but her initial modeling effort has yielded the opposite result. Recognizing that she does not have much experience with this type of data analysis, she has turned to you for assistance.

Her data are contained in the accompanying file `plant_richness.csv`, which has the following columns of information:

- `island`: name of the island
- `species`: number of plant species on the island
- `area` the area of the island (km^2)
- `elevation`: the highest elevation of the island (m)
- `distance`: the distance to the nearest island (km)
- `adjacent` the area of the nearest island (km^2)

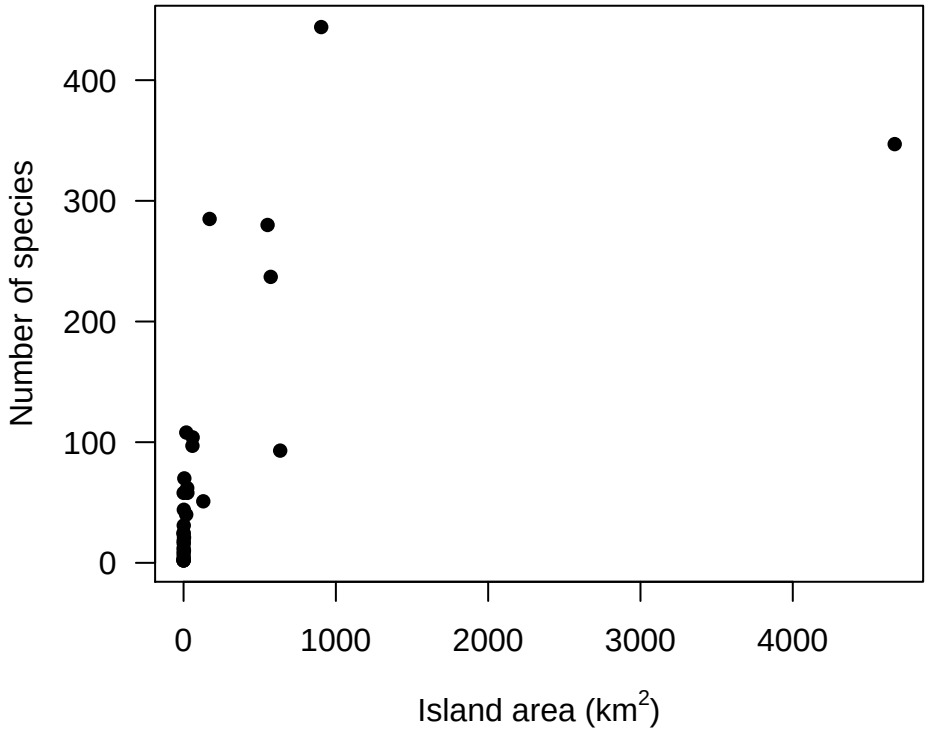
Questions

a) Plot the number of species versus island area and describe any patterns you observe. Does your colleague's assumption of a positive relationship between richness and area seem to hold?

```
## get the data
dat <- read.csv("plant_richness.csv")

## plot the richness vs area
par(mai = c(0.9, 0.9, 0.6, 0.1),
    omi = c(0, 0, 0, 0))
```

```r
plot(dat$area, dat$species, las = 1, pch = 16,
     ylab = "Number of species", xlab = expression(paste("Island area (",km^2,")")))
```



Yes, there does appear to be a positive relationship between the number of species and island area. There is also one apparent outlier that seems to suggest a possible nonlinear, saturating relationship between the number of species and island area.

---

b) Your colleague explains that she fit the following model, which yielded the surprising result. Fit the model for yourself and verify if there is indeed a negative effect of `area` on `species`. Do the signs of the other coefficients seem to make sense from an ecological perspective? Why or why not?

$$\text{species}_i = \beta_0 + \beta_1 \text{area}_i + \beta_2 \text{elevation}_i + \beta_3 \text{nearest}_i + \beta_4 \text{adjacent}_i + e_i$$

```r
## fit the model
fit_lm <- lm(species ~ area + elevation + nearest + adjacent, data = dat)
summary(fit_lm)
```

```
##
## Call:
## lm(formula = species ~ area + elevation + nearest + adjacent,
##     data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -115.84  -32.56  -11.34   29.19  186.00
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.17925   18.10976  -0.010 0.992181
## area        -0.02489    0.02252  -1.105 0.279599
## elevation    0.32540    0.05366   6.064 2.46e-06 ***
## nearest     -0.72732    0.82637  -0.880 0.387167
## adjacent    -0.07858    0.01746  -4.501 0.000136 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61.28 on 25 degrees of freedom
## Multiple R-squared:  0.7537, Adjusted R-squared:  0.7143
## F-statistic: 19.12 on 4 and 25 DF,  p-value: 2.58e-07
```

Yes, the effect of area is approximately -0.025, although its p-value would suggest the effect is non-significant. The effect of `elevation` is positive, which seems reasonable as greater variability in elevation should allow for greater niches. The effect of `nearest` is negative, although insignificant, which makes sense because greater distances between islands should decrease the dispersal of seeds between them. The effect of `adjacent` is negative, which doesn't make sense because larger neighboring islands should have more species and hence a greater probability that one of them would make it to the island in question.

---

c) Offer one explanation for the unexpected effect of `area` given the apparent relationship in (a). Based on this evaluation, offer a possible suggestion for estimating the effect of `area` on `species`.

We have discussed how collinearity among predictor variables can causes non-identifiability problems if/when including 2+ covariates in the same model if they are highly correlated. Here are the correlations among the four covariates.

```
## correlation among predictors
round(cor(dat[,-(1:2)]), 2)
```

```
##            area elevation nearest adjacent
## area       1.00      0.75   -0.11     0.18
## elevation  0.75      1.00   -0.01     0.54
## nearest   -0.11     -0.01    1.00    -0.12
## adjacent   0.18      0.54   -0.12     1.00
```

The correlation between `area` and `elevation` is quite high ($\rho \approx 0.75$), and this colinearity is causing non-identifiability problems when fitting the model. One simple solution is to fit models that include *either* `area` or `elevation`, but not both of them in the same model.

---

d) Does it seem reasonable to use `species` as a response variable in a linear model like the one your colleague fit initially? Why or why not? What would be a more appropriate response variable in a linear model like this?

No, using `species` itself as a response variable is not a good idea because it's a count, which means a simple linear model could predict negative counts. Also, the diagnostics from the linear model in (b) suggest problems with the assumptions of IID errors. Two possible solutions would be to model the log(`species`) or the log(`counts`/`area`) as a function of the covariates.

---

e) Based upon your knowledge of models for count data, offer a *simple* alternative regression model that models `species` as a function of `area`, `nearest`, and `adjacent`. What are the important components to this model?

An obvious choice would be a Poisson regression model (GLM) with the following three components:

1) data distribution: $y_i \sim \text{Poisson}(\lambda_i)$
2) link function: $\log(\lambda_i) = \eta_i$
3) linear predictor: $\eta_i = \beta_0 + \beta_1 \text{area}_i + \beta_2 \text{nearest}_i + \beta_3 \text{adjacent}_i$

---

f) Fit the model you recommended in (e) and examine the summary information. Does the effect of `area` seem more reasonable in this model? Do you see any problems with this model?

```
## fit Poisson regression model
fit_glm <- glm(species ~ area + nearest + adjacent, data = dat,
               family = poisson(link = "log"))
## model summary
summary(fit_glm)
```

```
##
## Call:
## glm(formula = species ~ area + nearest + adjacent, family = poisson(link = "log"),
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -10.739   -8.069   -5.004    2.205   25.653
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.180e+00  2.918e-02 143.283  < 2e-16 ***
## area         4.302e-04  1.199e-05  35.879  < 2e-16 ***
## nearest      5.889e-03  1.434e-03   4.106 4.03e-05 ***
## adjacent    -8.064e-05  2.793e-05  -2.887  0.00389 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 3510.7  on 29  degrees of freedom
## Residual deviance: 2587.2  on 26  degrees of freedom
## AIC: 2756.1
```

```
##
## Number of Fisher Scoring iterations: 6
```

The effect of `area` is positive (~4.3×10^{-4}), which is what we would expect based on theory and the plot in (a).

The deviance D of this model is ~2587 based upon 26 degrees of freedom. Thus, the estimated dispersion ($\hat{c}$) is

$$\hat{c} = \frac{D}{n-k}$$

```
## sample size
nn <- nrow(dat)
##
k <- length(coef(fit_glm))
## estimated dispersion
(c_hat <- fit_glm$deviance / (nn - k))
```

```
## [1] 99.5086
```

The estimated overdispersion from this model is *very* large, suggesting that we need to account for it when estimating the variance of the parameters and any test statistics associated with them.

---

g) Based on your assessment of the model in (f), identify three possible alternatives for estimating the model parameters and their associated uncertainty, and show how you would do so in **R**. How do the these alternative models compare to the estimates in (f).

Option 1: Poisson model with overdispersion using the `c_hat` estimated above.

```
## Poisson with overdispersion
summary(fit_glm, dispersion = c_hat)
```

```
##
## Call:
## glm(formula = species ~ area + nearest + adjacent, family = poisson(link = "log"),
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -10.739   -8.069   -5.004    2.205   25.653
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.180e+00  2.910e-01  14.364  < 2e-16 ***
## area         4.302e-04  1.196e-04   3.597 0.000322 ***
## nearest      5.889e-03  1.431e-02   0.412 0.680629
## adjacent    -8.064e-05  2.787e-04  -0.289 0.772295
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for poisson family taken to be 99.5086)
##
##     Null deviance: 3510.7  on 29  degrees of freedom
## Residual deviance: 2587.2  on 26  degrees of freedom
## AIC: 2756.1
##
## Number of Fisher Scoring iterations: 6
```

Option 2: Quasi-Poisson model.

```
## fit quasi-Poisson regression model
fit_glm_quasi <- glm(species ~ area + nearest + adjacent,
                     data = dat,
                     family = quasipoisson(link = "log"))

## model summary
summary(fit_glm_quasi)
```

```
##
## Call:
## glm(formula = species ~ area + nearest + adjacent, family = quasipoisson(link = "log"),
##     data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -10.739   -8.069   -5.004    2.205   25.653
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.180e+00  3.247e-01  12.877 8.65e-13 ***
## area         4.302e-04  1.334e-04   3.224  0.00339 **
## nearest      5.889e-03  1.596e-02   0.369  0.71512
## adjacent    -8.064e-05  3.108e-04  -0.259  0.79736
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 123.8197)
##
##     Null deviance: 3510.7  on 29  degrees of freedom
## Residual deviance: 2587.2  on 26  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

Option 3: Negative binomial model.

```
## fit negative binomial regression model
```

```
fit_glm_NB <- MASS::glm.nb(species ~ area + nearest + adjacent, data = dat,
                           link = "log", maxit = 50)
## model summary
summary(fit_glm_NB)
```

```
##
## Call:
## MASS::glm.nb(formula = species ~ area + nearest + adjacent, data = dat,
##     maxit = 50, link = "log", init.theta = 0.8235594176)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.9144  -1.1960  -0.3954   0.3149   1.8517
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.915e+00  2.701e-01  14.493  < 2e-16 ***
## area         1.107e-03  2.419e-04   4.574  4.8e-06 ***
## nearest      5.801e-03  1.458e-02   0.398    0.691
## adjacent    -8.161e-05  2.432e-04  -0.336    0.737
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.8236) family taken to be 1)
##
##     Null deviance: 44.821  on 29  degrees of freedom
## Residual deviance: 34.818  on 26  degrees of freedom
## AIC: 324.23
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.824
##           Std. Err.:  0.190
##
##  2 x log-likelihood:  -314.226
```

As expected, all of these alternative models that account for overdispersion have the same point estimates of the parameters as the model in (f), but the uncertainty in the parameters is much greater, leading to smaller z values and non-significant effects of `nearest` and `adjacent`.

---

h) For one of your alternatives in (g), evaluate whether a model that includes only `area` as a predictor is better than a model with all three predictors. Show the **R** code necessary to estimate the model and any test(s) or comparison(s) you might use.

Option 1: overdispersed Poisson

If we use an overdispersed Poisson model, we can compare the two models via QAIC.

```
## fit reduced Poisson model
fit_glm_r <- glm(species ~ area, data = dat,
                 family = poisson(link = "log"))

## log-likelihoods
LL_f <- logLik(fit_glm)
LL_r <- logLik(fit_glm_r)

## QAIC for full Poisson model
QAIC_f <- 2*4 - as.numeric(logLik(fit_glm) / (fit_glm$deviance / (nn - 4)))
## QAIC for reduced Poisson model
QAIC_r <- 2*2 - as.numeric(logLik(fit_glm_r) / (fit_glm_r$deviance / (nn - 2)))
## delta-QAIC
QAIC_r - QAIC_f
```

```
## [1] -2.947275
```

The reduced model has an QAIC that is ~3 units less than the full model, indicating more data support for the model with `area` only.

Option 2: Quasi-Poisson model

For the quasi-Poisson model, we do not have a likelihood from which to estimate QAIC, but we can use an F-test (recall that the χ^2 test is not appropriate for Poisson models with overdispersion). The null hypothesis is that the model with `area` only provides a better fit to the data.

```
## fit reduced models
## quasi-Poisson regression model
fit_glm_quasi_r <- glm(species ~ area, data = dat,
                       family = quasipoisson(link = "log"))

## F test with df = 2
anova(fit_glm_quasi_r, fit_glm_quasi, test = "F")
```

```
## Analysis of Deviance Table
##
## Model 1: species ~ area
## Model 2: species ~ area + nearest + adjacent
##   Resid. Df Resid. Dev Df Deviance      F Pr(>F)
## 1        28     2615.6
## 2        26     2587.2  2   28.368 0.1146 0.8922
```

This p-value is ~0.9 so we cannot reject the null hypothesis and therefore we conclude that the full model does not provide an improvement over the reduced model with `area` only.

Option 3: Negative binomial

We can use a likelihood ratio test or compare AIC values for both forms of the negative binomial models.

```
## fit reduced negative binomial model
fit_glm_NB_r <- MASS::glm.nb(species ~ area, data = dat,
                             link = "log", maxit = 50)

## LRT with df = 2
anova(fit_glm_NB_r, fit_glm_NB)
```

```
## Likelihood ratio tests of Negative Binomial Models
##
## Response: species
##                        Model     theta Resid. df    2 x log-lik.   Test
## 1                       area 0.8163231        28       -314.5613
## 2 area + nearest + adjacent 0.8235594        26       -314.2256 1 vs 2
##      df  LR stat.   Pr(Chi)
## 1
## 2     2 0.3357233 0.8454708
```

The p-value is ~0.8 so we cannot reject the null hypothesis and therefore we conclude that the full model does not provide an improvement over the reduced model with `area` only.

```
## compare AIC values
## AIC for full Poisson model with overdispersion
AIC_3_f <- summary(fit_glm_NB)$aic
## AIC for reduced Poisson model with overdispersion
AIC_3_r <- summary(fit_glm_NB_r)$aic
## delta-AIC
AIC_3_r - AIC_3_f
```

```
## [1] -3.664277
```

The AIC value for the model with `area` only is ~3.7 units less than the model with three predictors, so we should select the more simple model with only `area`.

---

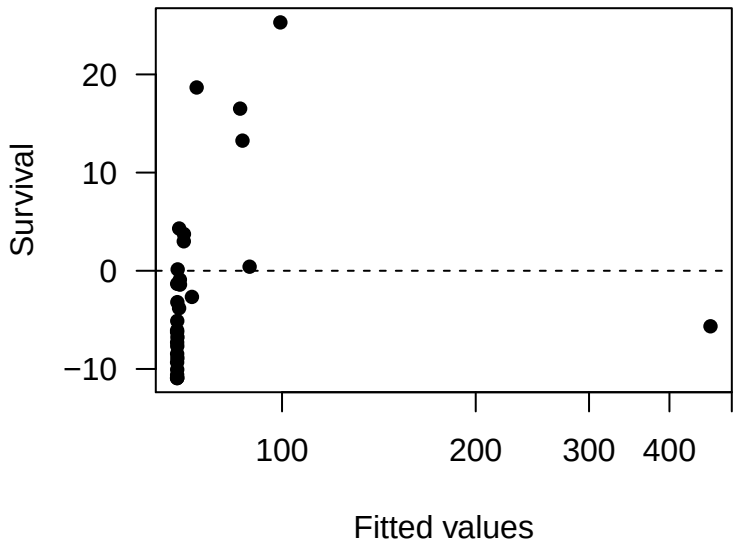
i) Evaluate the diagnostics for your model from (h) with `species` as a function of `area` only. Do you see any problems with this model?

Option 1: overdispersed Poisson

There is no way to evaluate diagnostics for the overdispersed Poisson model, in that the residuals, etc come from the model fitted without accounting for overdispersion (i.e., we only use the dispersion parameter in computing the SE's, z-values, and p-values). Here I examine a plot of the residuals versus the fitted values, and I check the leverages and Cook's Distance.

```
## set up plot region
par(mai = c(0.9, 0.9, 0.1, 0.1),
    omi = c(0, 0, 0, 0))

## residuals vs fitted
plot(fitted(fit_glm_r), residuals(fit_glm_r), las = 1, pch = 16, log = "x",
     ylab = "Survival", xlab = "Fitted values")
abline(h = 0, lty = "dashed")
```




```
## leverages
hat_values <- hatvalues(fit_glm_r)
names(hat_values) <- dat$island
## threshold value
(h_crit <- 2 * length(coef(fit_glm_r)) / nn)
```

```
## [1] 0.1333333
```

```
## check if any h_i > b_crit
hat_values[hat_values > h_crit]
```

```
##   Isabela
## 0.9857684
```

```
## Cook's D
CD <- cooks.distance(fit_glm_r)
names(CD) <- dat$island
## Threshold value
(CD_crit <- qf(0.5, nn, nn - length(coef(fit_glm_r))))
```

```
## [1] 1.00161
```

```
## check if any CD_i > CD_crit
CD[CD > CD_crit]
```

```
##     Caldwell      Coamano      Enderby     Gardner2      Isabela
##     1.172550     1.208443     1.208499     1.102505 71030.146401
##       Onslow SanCristobal  SanSalvador    SantaCruz   SantaMaria
##     1.208425     8.319748     4.966860    25.160060    11.227301
```

The plot of the residuals against the fitted values does not appear to be too disturbing, but there does appear to be one fitted value that is much larger than the rest.

There is one island with a leverage greater than the expected value of ~0.13.

There are 10 islands with Cook's D greater than the expected value of ~1.

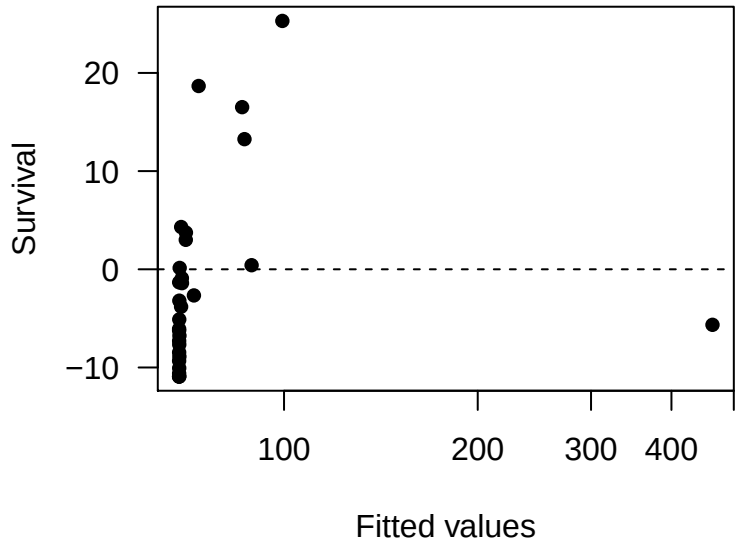
These diagnostics suggest we perhaps have some work to do in further refining our model.

Option 2: Quasi-Poisson model

Here are some diagnostic checks for the quasi-Poisson model.

```
## set up plot region
par(mai = c(0.9, 0.9, 0.1, 0.1),
    omi = c(0, 0, 0, 0))

## residuals vs fitted
plot(fitted(fit_glm_quasi_r), residuals(fit_glm_quasi_r), las = 1, pch = 16, log = "x",
     ylab = "Survival", xlab = "Fitted values")
abline(h = 0, lty = "dashed")
```




```
## leverages
hat_values <- hatvalues(fit_glm_quasi_r)
names(hat_values) <- dat$island
## threshold value
(h_crit <- 2 * length(coef(fit_glm_quasi_r)) / nn)
```

```
## [1] 0.1333333
```

```
## check if any h_i > b_crit
hat_values[hat_values > h_crit]
```

```
##   Isabela
## 0.9857684
```

```
## Cook's D
CD <- cooks.distance(fit_glm_quasi_r)
names(CD) <- dat$island
## Threshold value
(CD_crit <- qf(0.5, nn, nn - length(coef(fit_glm_quasi_r))))
```

```
## [1] 1.00161
```

```
## check if any CD_i > CD_crit
CD[CD > CD_crit]
```

```
##  Isabela
## 609.6077
```

The plot of the residuals against the fitted values does not appear to be too disturbing, but there does appear to be one fitted value that is much larger than the rest.

There is one island with a leverage greater than the expected value of ~0.13.

There is one island with a Cook's D greater than the expected value of ~1.

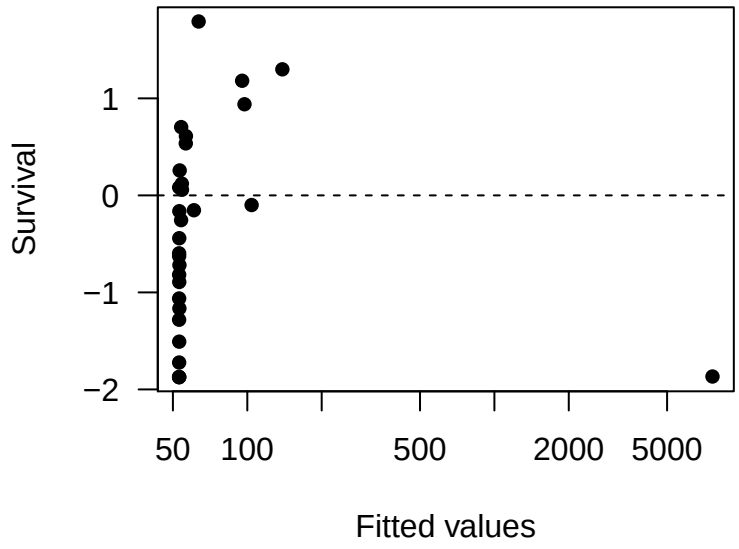
These diagnostics suggest we perhaps have some work to do in further refining our model.

Option 3: Negative binomial

Here are some diagnostic checks for the negative binomial model.

```
## set up plot region
par(mai = c(0.9, 0.9, 0.1, 0.1),
    omi = c(0, 0, 0, 0))

## residuals vs fitted
plot(fitted(fit_glm_NB_r), residuals(fit_glm_NB_r), las = 1, pch = 16, log = "x",
     ylab = "Survival", xlab = "Fitted values")
abline(h = 0, lty = "dashed")
```




```
## leverages
hat_values <- hatvalues(fit_glm_NB_r)
names(hat_values) <- dat$island
## threshold value
(h_crit <- 2 * length(coef(fit_glm_NB_r)) / nn)
```

```
## [1] 0.1333333
```

```
## check if any h_i > b_crit
hat_values[hat_values > h_crit]
```

```
##   Isabela
## 0.9310076
```

```
## Cook's D
CD <- cooks.distance(fit_glm_NB_r)
names(CD) <- dat$island
## Threshold value
(CD_crit <- qf(0.5, nn, nn - length(coef(fit_glm_NB_r))))
```

```
## [1] 1.00161
```

```
## check if any CD_i > CD_crit
CD[CD > CD_crit]
```

```
## Isabela
## 72.7315
```

The plot of the residuals against the fitted values does not appear to be too disturbing, but there does appear to be one fitted value that is much larger than the rest.

There is one island with a leverage greater than the expected value of ~0.13.

There is one island with a Cook's D greater than the expected value of ~1.

These diagnostics suggest we perhaps have some work to do in further refining our model.