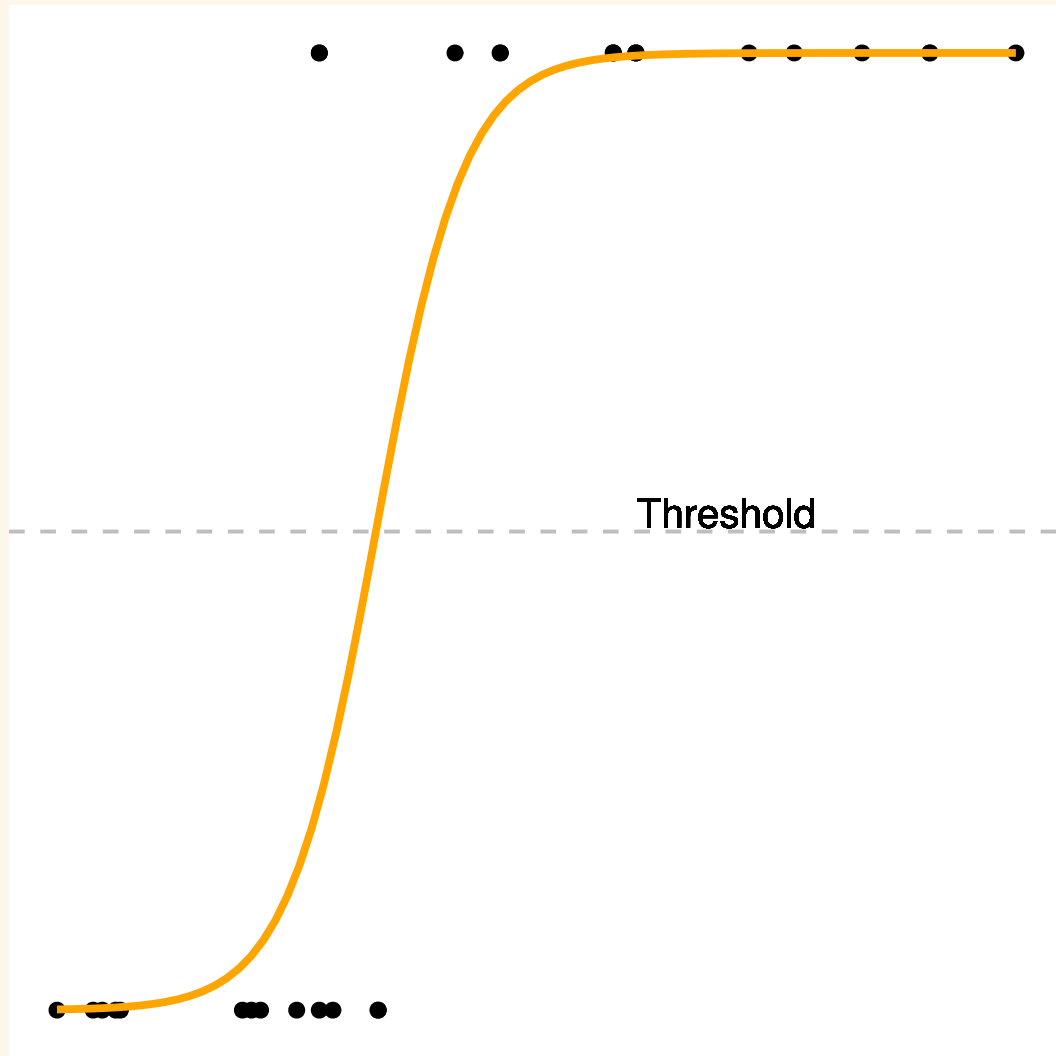


# Binomial Logistic regression: deviance

Stat 230

May 25 2022

# Overview



Today:

Deviance

Assumptions

Residuals and case influence

## Review: A Logistic Regression Model for Binomial Count Data

For all  $i = 1, \dots, n$ ,

$$y_i \sim \text{binomial}(m_i, \pi_i),$$

where  $m_i$  is a known number of trials for observation  $i$ ,

$$\pi_i = \frac{\exp(\mathbf{x}_i' \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i' \boldsymbol{\beta})},$$

and  $y_1, \dots, y_n$  are independent.

# Overview: Binomial Distribution

Recall that for  $y_i \sim \text{binomial}(m_i, \pi_i)$ , the probability mass function of  $y_i$  is

$$P(y_i = y) = \begin{cases} \binom{m_i}{y} \pi_i^y (1 - \pi_i)^{m_i - y} & \text{for } y \in \{0, \dots, m_i\} \\ 0 & \text{otherwise} \end{cases}$$
$$E(y_i) = m_i \pi_i, \quad \text{and } \text{Var}(y_i) = m_i \pi_i (1 - \pi_i).$$

The binomial log likelihood function is

$$\ell(\boldsymbol{\beta} \mid \mathbf{y}) = \sum_{i=1}^n \left[ y_i \log\left(\frac{\pi_i}{1 - \pi_i}\right) + m_i \log(1 - \pi_i) \right] \\ + \text{constant}$$

# Deviance for Binomial responses

- With Binomial responses, the likelihood function is

$$L(\beta) = \prod_{i=1}^n \binom{m_i}{y_i} \pi(X_i)^{y_i} (1 - \pi(X_i))^{m_i - y_i}$$

- and the deviance is

$$\begin{aligned} G^2 &= 2[\ln L(\bar{\pi}) - \ln L(\hat{\pi}(X))] \\ &= 2 \sum_{i=1}^n \left[ y_i \ln \left( \frac{y_i}{m_i \hat{\pi}(X_i)} \right) + (m_i - y_i) \ln \left( \frac{m_i - y_i}{m_i - m_i \hat{\pi}(X_i)} \right) \right] \end{aligned}$$

- $L(\hat{\pi}(X))$  : likelihood of the data that plugs in estimates  $\hat{\pi}(X_i)$  from the logistic model.
- $L(\bar{\pi})$  : likelihood of the data that plugs in estimates  $\bar{\pi}_i = y_i/m_i$

$$L(\bar{\pi}) \geq L(\hat{\pi}(X))$$

# Logistic Regression Model vs Saturated model

## Logistic Regression Model

- $y_i \sim \text{binomial}(m_i, \pi_i)$
- $y_1, \dots, y_n$  independent

$$\pi_i = \frac{\exp(\mathbf{x}_i' \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i' \boldsymbol{\beta})}$$

- $p + 1$   $\boldsymbol{\beta}$  parameters
- MLE:  $\hat{\pi}_i = \frac{\exp(\mathbf{x}_i' \hat{\boldsymbol{\beta}})}{1 + \exp(\mathbf{x}_i' \hat{\boldsymbol{\beta}})}$

## Saturated Model

- $y_i \sim \text{binomial}(m_i, \pi_i)$
- $y_1, \dots, y_n$  independent
- $\pi_i \in [0, 1]$  for  $i = 1, \dots, n$  with no other restrictions
- $n$  parameters
- MLE:  $y_i / m_i$

## Deviance for Binomial responses

Deviance for binomial models can be used for two types of hypothesis tests:

1. **Drop-in-deviance:** Used to compare two models, just like in binary logistic models.
2. **Goodness-of-fit:** Used to test binomial response model adequacy.

# Goodness-of-fit test

Our hypotheses for the GOF test are:

$H_0$  : logistic model

$H_A$  : saturated model

$H_0 : Y_i \mid X_i \sim \text{Binom}(m_i, \pi(X_i))$

- $\pi(X_i)$  equals the logistic function of the  $p$  predictor terms.

$H_A$  : the saturated "model"

- uses the  $n$  empirical proportion of successes  $\bar{\pi} = y_i/m_i$  for each case as the probability of success for all  $m_i$  trials.



# Goodness-of-fit test

- The test statistic is the residual deviance of the logistic model

$$G^2 = 2[\ln L(\bar{\pi}) - \ln L(\hat{\pi}(X))]$$

- If the "fit" (likelihood) of the logistic model is "close" then  $G^2$  is "close" to 0 and we can claim that the logistic model is adequate.
- If  $H_0$  is true and  $m_i$  's are large,  $G^2$  will have an approximate chi-square distribution with  $n - (p + 1)$  (model) degrees of freedom. The **p-value** is the probability of getting residual deviance values larger than the observed value:

$$p - \text{value} = 1 - P(\chi^2 > G^2) = 1 - \text{pchisq}(G^2, df = n - (p + 1))$$

The suggested rule of thumb for "large  $m$  " is that we want most  $m_i$  's to be at least 5 .

## Goodness-of-fit test conclusions

### **Do not reject the null:** (large p-value)

- Your logistic model is adequate.
- You don't have a large enough sample size  $n$  to have the power to detect inadequacies in your model.

### **Reject the null:** (small p-value)

- You have outlier(s) that are inflating the residual deviance.
- Your logistic model is inadequate.

# Goodness-of-fit test conclusions

## Why might a model be inadequate?

- Your log-odds model is inadequate, it is ill-fitting and transformations are needed
- Extra-binomial variation: your response counts aren't well modeled by a Binomial model\*.

\*This could mean:

- trials are not **independent** for each case
- probability of success is **not** constant across trials for each case
- your choice of predictors **isn't** sufficient (i.e. you are missing key explanatory variables)

# Case study 21.1: Krunnit Island

- **GOF** test hypotheses are

$$H_0 : \log(\text{odds}) = \beta_0 + \beta_1 \log(\text{area})$$

$$H_A : \log(\text{odds}) = \alpha_i (\text{saturated model})$$

We can conduct a GOF test because all our  $m_i$ 's (AtRisk), are above 5

```
island <- case2101  
island$AtRisk
```

```
[1] 75 67 66 51 28 20 43 31 28 32 30 20 31 16 15 33 40 6
```

# Case study 21.1: Krunit Island

Test stat: residual deviance of  $G^2 = 12.062$

```
krunit_glm <- glm(Extinct/AtRisk ~ log(Area), family="binomial", weights=AtRisk, data = island)
summary(krunit_glm)
```

Call:

```
glm(formula = Extinct/AtRisk ~ log(Area), family = "binomial",
     data = island, weights = AtRisk)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.71726	-0.67722	0.09726	0.48365	1.49545

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.19620	0.11845	-10.099	< 2e-16 ***
log(Area)	-0.29710	0.05485	-5.416	6.08e-08 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 45.338 on 17 degrees of freedom  
Residual deviance: 12.062 on 16 degrees of freedom  
AIC: 75.394

Number of Fisher Scoring iterations: 4

## Case study 21.1: Krunnit Island

- **p-value:** Using  $18 - 2 = 16$  degrees of freedom

$$p\text{-value} = 1 - P(\chi^2 > 12.062) = 1 - pchisq(12.062, df = 16) = 0.7397$$

- The **large** p-value means that we **do not reject** the null hypothesis.
- Our model for the probability of extinction given log-area looks to be adequate.

# Case study 21.1: Krunit Island

```
krunit_glm_nolog <- glm(Extinct/AtRisk ~ Area, family="binomial", weights=AtRisk, data = island)
summary(krunit_glm_nolog)
```

Call:

```
glm(formula = Extinct/AtRisk ~ Area, family = "binomial", data = island,
     weights = AtRisk)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-1.6526	-1.0661	-0.1877	1.0038	2.1860

Coefficients:

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	-1.305957	0.117339	-11.130	< 2e-16 ***
Area	-0.010121	0.002684	-3.771	0.000163 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 45.338 on 17 degrees of freedom  
Residual deviance: 24.661 on 16 degrees of freedom  
AIC: 87.993

Number of Fisher Scoring iterations: 4

## Case study 21.1: Krunnit Island

- Regression of extinction on area (not logged) has residual deviance of 24.661.
- The GOF p-value for this model is 0.076, which suggests that this model may not be adequate
  - agrees with our EDA for the log-odds suggests

```
1-pchisq(24.661, df=16)
```

```
[1] 0.07602884
```



# Checking Assumptions

Binomial logistic model as the **same assumptions** as binary models:

- **Independence** of cases takes an understanding of how the data was collected.
- Log-odds **linearity** can be checked with an empirical log-odds plot against quantitative predictors and residual plots.

A third assumption is that the counts of successes  $Y_i$  has a **binomial distribution**:

- the  $m_i$  trials are independent events, and a success or failure for one trial doesn't affect the outcome of another trial, and
- the probability of success  $\pi(X_i)$  is the same for all  $m_i$  trials.

## Checking Assumptions

- If one, or both, of these assumptions is violated, then it often induces extra-binomial variation (a.k.a. **over dispersion**):
- the actual variation  $SD(Y | X_i)$  is larger than the binomial SD of  $\sqrt{m_i \pi(X_i) (1 - \pi(X_i))}$
- making our reported standard errors smaller and p-values too small
- **Check:** use the goodness-of-fit test, when  $m_i$  are large enough, to check our binomial distribution assumption.

# Checking Assumptions

If we do find evidence of lack-of-fit in our binomial model, then you should

- Check deviance residuals as case influence stats to see if an outlier(s) is affecting GOF results.
- Check the log odds form, change model structure, see if transformations of quantitative predictors are needed

If outliers and transformations aren't a concern, then consider an alternative model:

- binary logistic model if trial-level predictors are available
- quasi-binomial logistic model
- a model that allows for correlated trials (like a mixed-effects logistic model)

## Case study 21.1: Krunnit Island

How might the Krunnit Island extinction counts violate the binomial counts model assumptions?

### **Independence:**

- This assumption implies that the extinction, or not, of all at risk species on an island are independent events.
- This could be violated if the extinction of one species makes the extinction of a second more likely.

## Case study 21.1: Krunnit Island

### Probability:

- This assumption implies that the probability  $\pi(\text{area}_{\{i\}})$  of extinction on island  $i$  is the same for all at risk species on island  $i$ .
- This could be violated if, for example, species living primarily on the interior of the island had a lower chance of extinction than species living on the coastal region.

# Residuals

Pearson residuals are basically response residuals standardized based on the binomial SD:

$$pr_i = \frac{y_i - m_i \hat{\pi}(X_i)}{\sqrt{m_i \hat{\pi}(X_i) (1 - \hat{\pi}(X_i))}}$$

- `resid(my_glm, type = "pearson")`
- `augment(my_glm, type.residuals = "pearson")`

# Residuals

Deviance residuals are each case's contribution to the residual deviance, with a  $\pm$  based on whether we over- or under-estimate a case's response (the  $\pm$  is denoted by  $\text{sign}(y_i - m_i \hat{\pi}(X_i))$ ):

$$\text{Dres}_i = \text{sign}(y_i - m_i \hat{\pi}(X_i)) \sqrt{2 \left[ y_i \ln \left( \frac{y_i}{m_i \hat{\pi}(X_i)} \right) + (m_i - y_i) \ln \left( \frac{m_i - y_i}{m_i - m_i \hat{\pi}(X_i)} \right) \right]}$$

- `resid(my_glm, type = "deviance")`
- `augment(my_glm, type.residuals = "deviance")`

## Residuals

- Pearson residuals are "easy" to interpret
- Deviance residuals are good to check if you find significant results in a GOF test.
- When  $m_i$  's are large (at least 5), both types of residuals should be similar in value and have a  $N(0, 1)$  distribution (approximately).
- Regardless of size of  $m_i$ , we should plot residuals vs. quantitative predictors to assess linearity of the log odds.



# Case study 21.1: Krunnit Island

- use the `augment` command to get both sets of residuals

```
island_aug <- augment(krunnit_glm, type.residual="pearson")
```

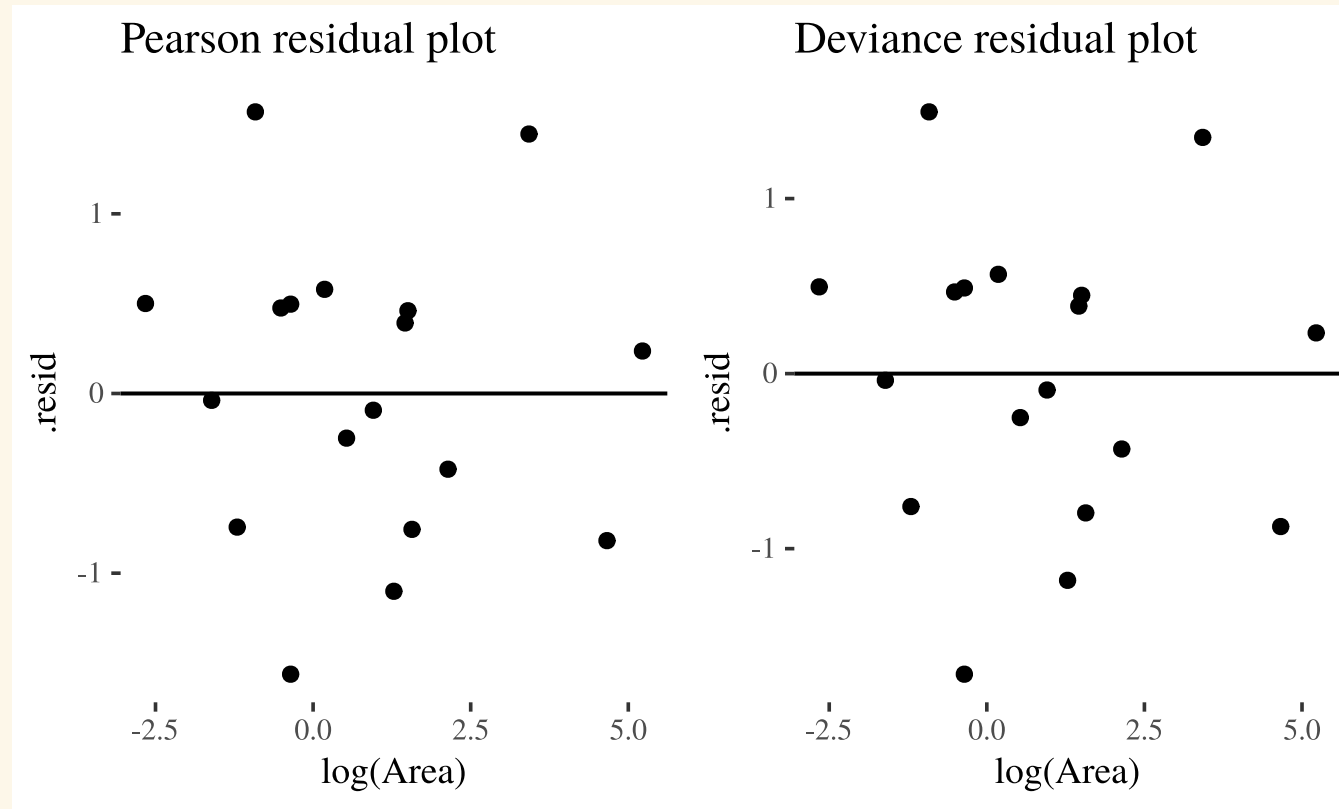
```
plotA <- ggplot(island_aug, aes(x=`log(Area)`, y=.resid)) +  
  geom_point() +  
  geom_hline(yintercept = 0) +  
  labs(title="Pearson residual plot")
```

```
island_aug <- augment(krunnit_glm, type.residual="deviance")
```

```
plotB <- ggplot(island_aug, aes(x=`log(Area)`, y=.resid)) +  
  geom_point() +  
  geom_hline(yintercept = 0) +  
  labs(title="Deviance residual plot")
```

# Case study 21.1: Krunnit Island

```
library(gridExtra)
grid.arrange(plotA, plotB, ncol=2)
```



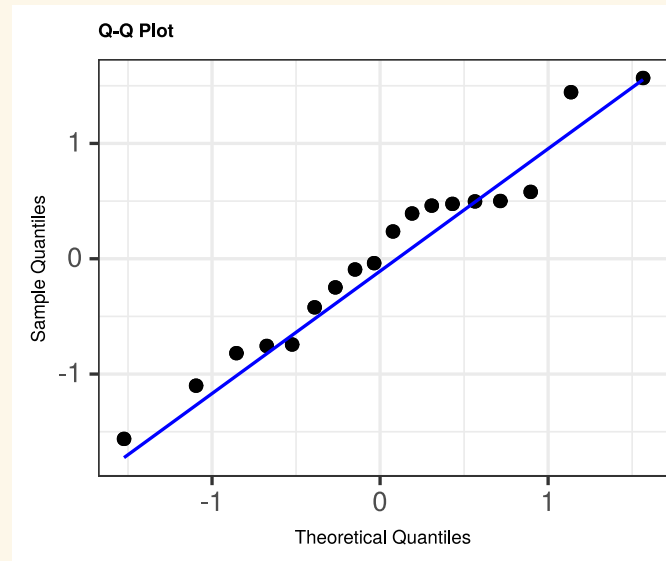
## Residuals using ggResidpanel

- `resid_xpanel(my_glm, type = )`: where type could be pearson or deviance (or response)
- `resid_panel(my_glm, plots = "qq", type = )`: qq plot of residuals given by type

# Case study 21.1: Krunit Island

At risk counts  $m_i$  are rather large (all cases are 6 or larger) so residuals should be approximately  $N(0, 1)$ .

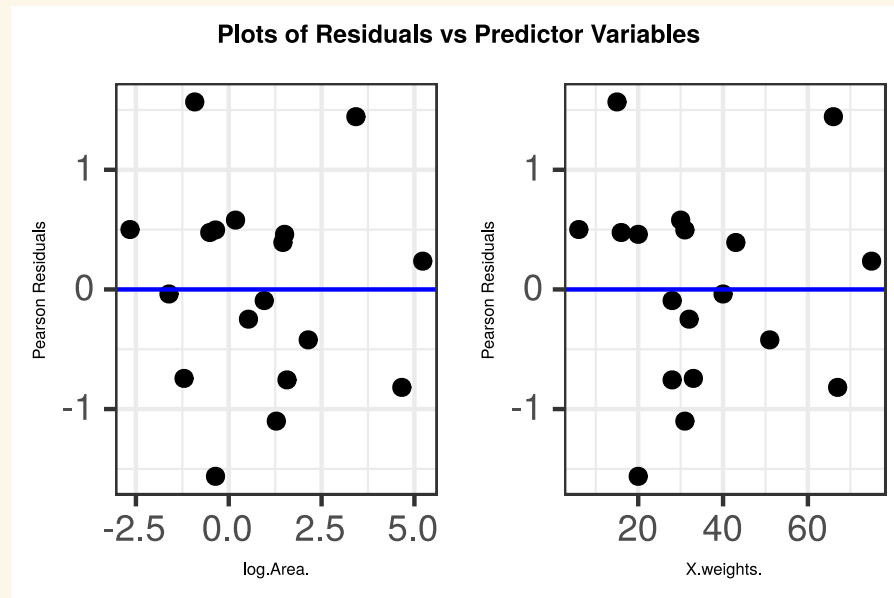
```
resid_panel(krunnit_glm, plots = "qq", type = "pearson", axis.text.size = 6, title.text.size = 6)
```



# Case study 21.1: Krunnit Island

at risk counts  $m_i$  are the `X.weights` variable (we don't want residual values to depend on  $m_i$ )

```
resid_xpanel(krunnit_glm, type = "pearson", axis.text.size = 4, title.text.size = 6)
```



## Case influence stats

In a GLM, leverage measures

- both a cases's "extremeness" in terms of it's predictor values and
  - the size of a case's weight  $m_i$
- 
- cases with high values of  $m_i$  are given more weight, and hence higher leverage, in the fitted model
  - Cook's distance also takes into account a cases leverage (measured both by predictor values and by  $m_i$  size) and a case's residual value.

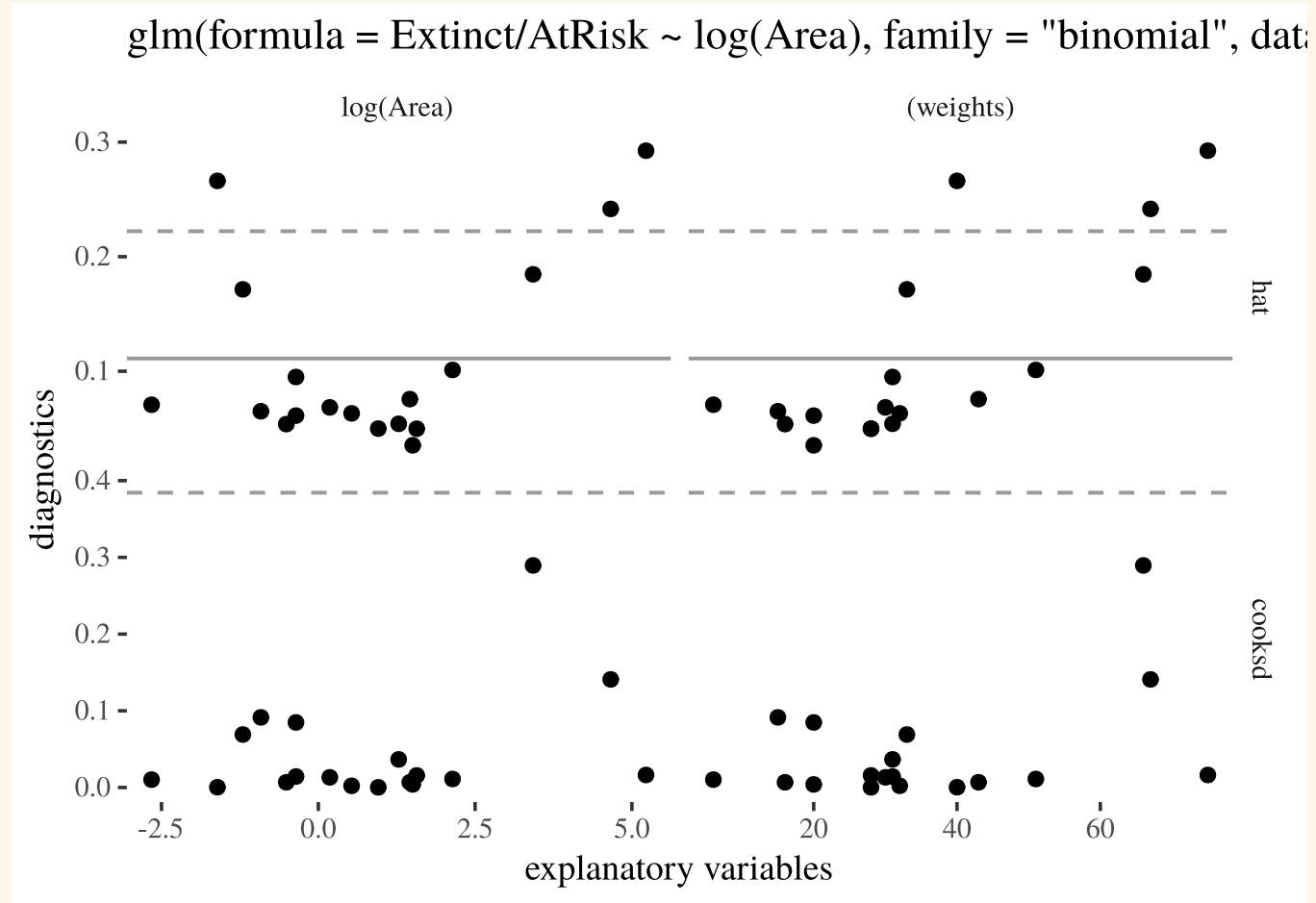
## Case study 21.1: Krunnit Island

- rows 1,2 and 17 have largest leverage while case 3 looks to have the highest Cook's distance value.

```
plot(krunnit_glm, which=5, id.n=18)
```

# Case study 21.1: Krunit Island

```
library(GGally)
ggnostic(krunnit_glm, columnsY = c(".hat", ".cooksd"))
```





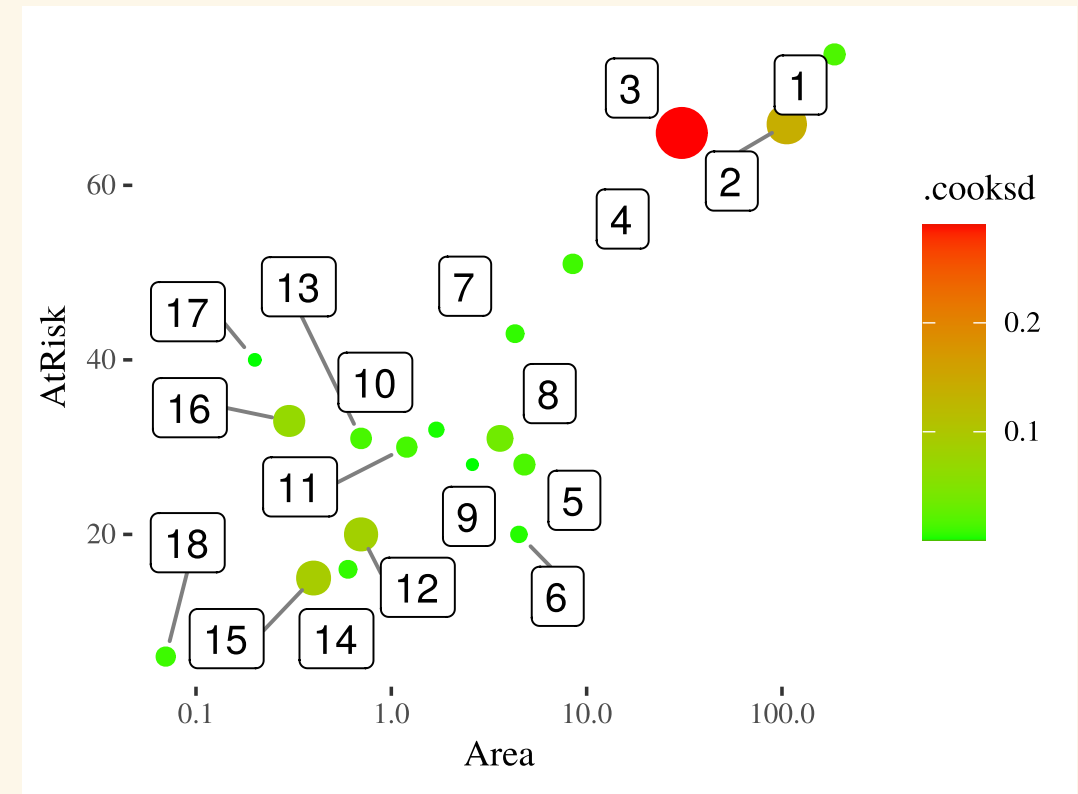
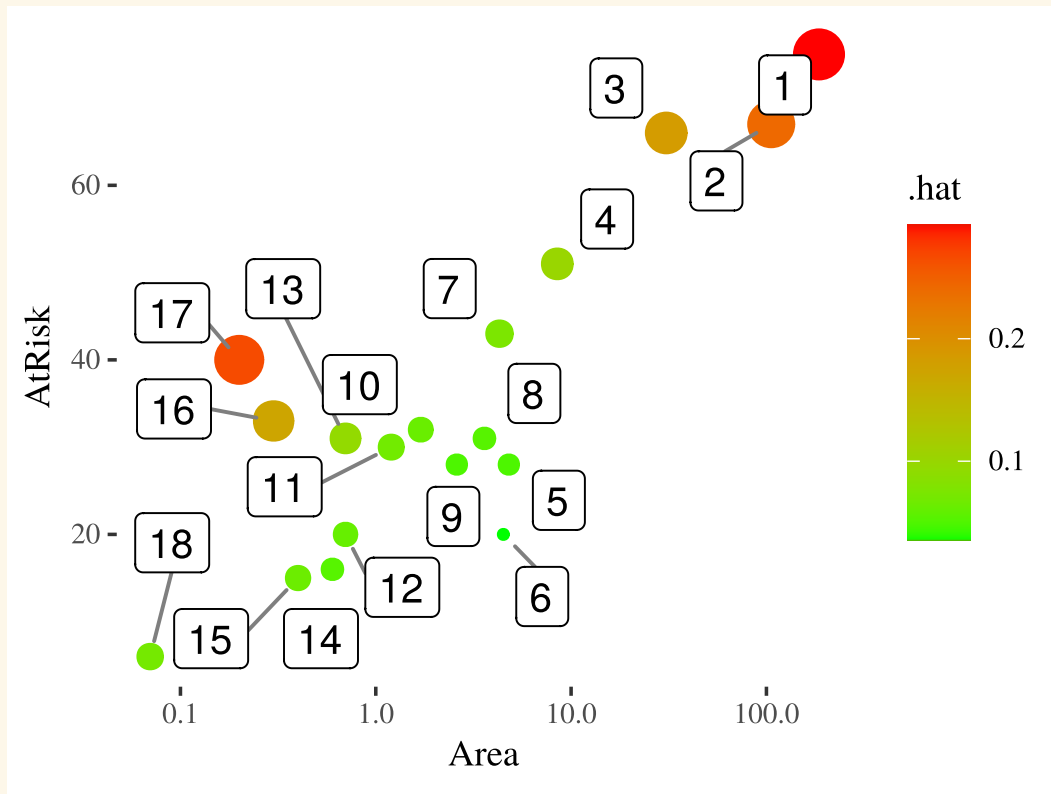
## Case study 21.1: Krunnit Island

```
island_aug <- augment(krunnit_glm, data=island, type.predict = "response")
island_aug <- island_aug %>% mutate(ID = row_number())
island_aug %>% slice(1,2,3,17) %>% select(-.sigma)
# A tibble: 4 × 10
  Island      Area AtRisk Extinct .fitted .resid .std.resid .hat .cooksd   ID
  <fct>    <dbl>  <int>  <int>  <dbl>  <dbl>    <dbl>  <dbl>  <dbl> <int>
1 Ulkokrunni 186.    75     5  0.0602  0.233    0.277  0.293 1.63e-2     1
2 Maakrunni 106.    67     3  0.0704 -0.874   -1.00  0.242 1.41e-1     2
3 Ristikari  30.7    66    10  0.0985  1.35     1.49  0.185 2.89e-1     3
4 Tiirakari   0.2    40    13  0.328  -0.0381 -0.0445 0.266 3.59e-4    17
```

```
summary(select(island_aug, Area, AtRisk))
```

Area		AtRisk	
Min. :	0.070	Min. :	6.00
1st Qu.:	0.625	1st Qu.:	22.00
Median :	2.150	Median :	31.00
Mean :	19.804	Mean :	35.11
3rd Qu.:	4.725	3rd Qu.:	42.25
Max. :	185.800	Max. :	75.00

# Case study 21.1: Krunnit Island



# Case study 21.1: Krunnit Island

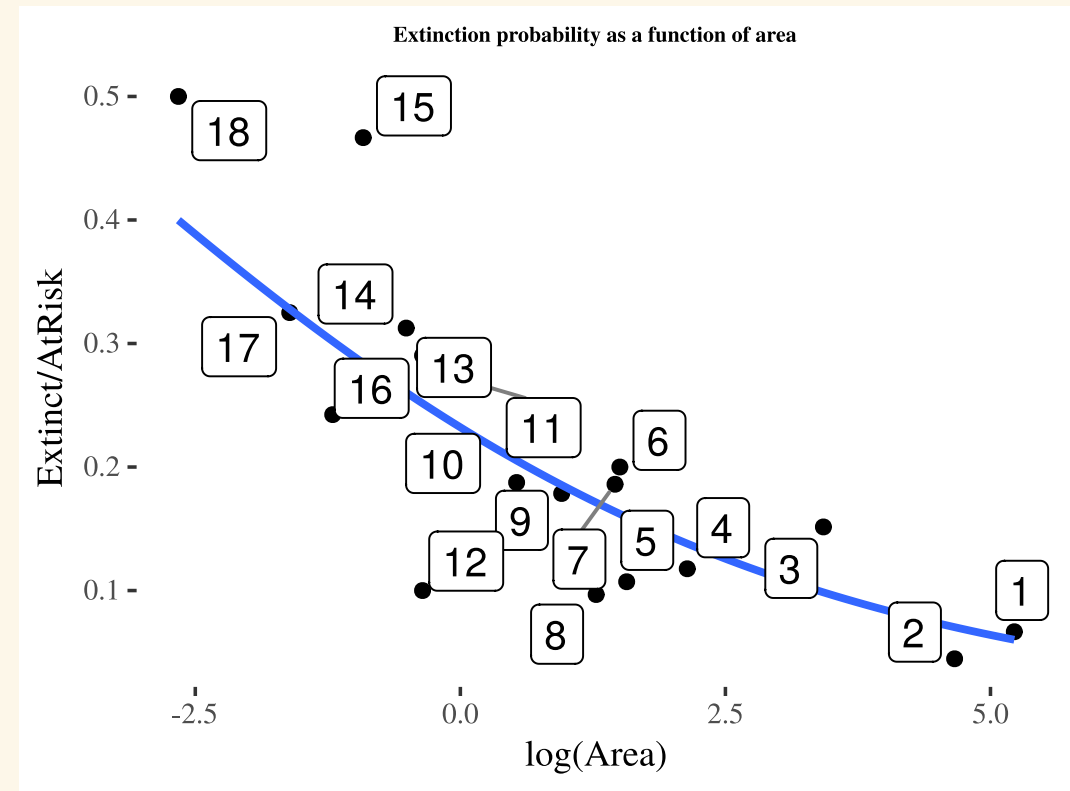
## Case 1 (Ulkokrunni):

- largest area and the largest number at risk  $m_i$ .
- has largest leverage but it doesn't have an large residual value so it doesn't have high Cook's distance.

```
library(ggplot2)
island <- island %>% mutate(ID = row_number())
plot <- ggplot(island, aes(x=log(Area),
                           y = Extinct/AtRisk,
                           weight = AtRisk)) +

  geom_point() +
  geom_smooth(method="glm", se=FALSE,
              method.args = list(family="binomial")) +
  labs(title="Extinction probability as a function of area")
theme(plot.title = element_text(hjust=0.5, size=12,
                                face='bold'))

plot + geom_label_repel(aes(label = ID),
                        box.padding = 0.15,
                        point.padding = 0.3,
                        segment.color = 'grey50')
```



# Case study 21.1: Krunnit Island

## Case 17 (Tiirakari)

- second smallest area but a large number of at risk species given its small size.
- has larger leverage than case 18 which has the smallest area but smaller number at risk.
- has a small residual and low Cook's distance.

## Case 3 (Ristikari)

- third largest area (30.7) but it's number at risk (66) is only one smaller than case 2 (Maakrunni) which is the second largest area (106).
- has a much larger residual than case 2, which results in it having the highest Cook's distance value in the data set.

- None of these cases is overly influential in the model it and removal of case 3 , the highest Cook's distance, changes the estimate of  $\beta_1$  from 0.30 to 0.33 and it's significance doesn't change.