Logistic Regression with a Binomial Response

Tate Jacobson

Department of Statistics
Oregon State University, Corvallis, Oregon

Spring 2026

Reading: Ch 21 of *The Sleuth*

Outline

1. Logistic Regression for Binomial Counts
   - Case Study 21.1.1
   - Counts versus Proportions
   - The Model
   - Model Checking
   - Interpretation

Logistic Regression for Binomial Counts

In the last chapter, we introduced logistic regression for binary data. In this chapter, we'll look at an extension of the logistic regression model for **binomial count** data.

Case Study 21.1.1

Island Size and Bird Extinctions data

- Experimental units: 18 islands in the Krunnit Archipelago in the Baltic Sea
- Variables:
  - island area in km^2
  - number of bird species observed in 1949 (called “species at risk”)
  - number out of those “at risk” that were NOT observed in 1959 (and therefore labeled “extinct”)

Question: Is there a relationship between island size and number of “extinctions?”

Island area, number of bird species present in 1949, and number of these not present in 1959; from the Krunnit Islands study

Island	Area (km^2)	Species at risk	Extinctions
Ulkokrunni	185.80	75	5
Maakrunni	105.80	67	3
Ristikari	30.70	66	10
Isonkivenletto	8.50	51	6
Hietakraasukka	4.80	28	3
Kraasukka	4.50	20	4
Länsiletto	4.30	43	8
Pihlajakari	3.60	31	3
Tyni	2.60	28	5
Tasasenletto	1.70	32	6
Raiska	1.20	30	8
Pohjanletto	0.70	20	2
Töro	0.70	31	9
Luusiletto	0.60	16	5
Vatunginletto	0.40	15	7
Vatunginnokka	0.30	33	8
Tiirakari	0.20	40	13
Ristikarenletto	0.07	6	3

Binomial Responses

We've already seen that if $Y_1, \ldots, Y_m$ are a sample of m independent Bernoulli(π) (binary) observations, then

$$W = \sum_{i=1}^{m} Y_i$$

is a Binomial random variable of size m and with probability (or binomial proportion) π.

We write: $W \sim Bin(m, \pi)$.

Binomial Response in Krunnit Island Example

In the Island Size and Bird Extinction data, each island has its own binomial count of extinctions, its own sample size, and its own binomial proportion.

Therefore we write $W_i \sim Bin(m_i, \pi_i)$ for $i = 1, \ldots, n$, where

- n is the number of islands
- m_i is the number of species at risk for island i
- π_i is the probability of a species going extinct on island i

Island	Area (km^2)	Species at risk	Extinctions
Ulkokrunni	185.80	75	5
Maakrunni	105.80	67	3
Ristikari	30.70	66	10

For the first few observations, what is the distribution of the count W_i and what is the observed count (w_i)?

-
-
-

Why analyze the Binomial counts?

It's often tempting to think of Binomial count data as proportion data instead, but this isn't a great idea.

- That is, we could just take W_i/m_i for each island, and analyze the n **proportions of extinction**.
- One problem with this is that in doing so, we lose track of the individual sample sizes.
  - That is, we would treat $1/10 = 0.1$ the same as we would treat $10/100 = 0.1$.
  - However, we have far more information if our observed count is based on 100 trials rather than 10 trials, so some information is lost in only using the proportions.

When to analyze the proportion itself

There are certainly instances when we might collect response data that are proportions:

- The proportion of land-cover that is pine.
- The proportion of total fatty acid mass represented by a particular fatty acid.
- The proportion of ozone in an air sample.

None of these proportions comes from binomial counts (there is no notion of a “sample size” here). These kinds of proportions should be analyzed using multiple linear regression, NOT logistic regression.

The Binomial Logistic Regression Model

There are two components to this model: the **distribution of the response**

$$Y_i \sim Bin(m_i, \pi_i)$$

for $i = 1, 2, \ldots, n$, and the Y_i are independent, and the **link**

$$\text{logit}(\pi_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}.$$

This is very similar to the Binary logistic regression model (really, binary logistic regression is just a special case of binomial logistic regression with $m_i = 1$ for all i), and *we'll interpret the regression coefficients in the same way as before*.

Plotting the Empirical Logits

With binomial counts, we can look at scatterplots of the empirical logits versus the explanatory variables (**no need for binning**!).

- For a binomial count X out of a sample size of n, the empirical logit is just

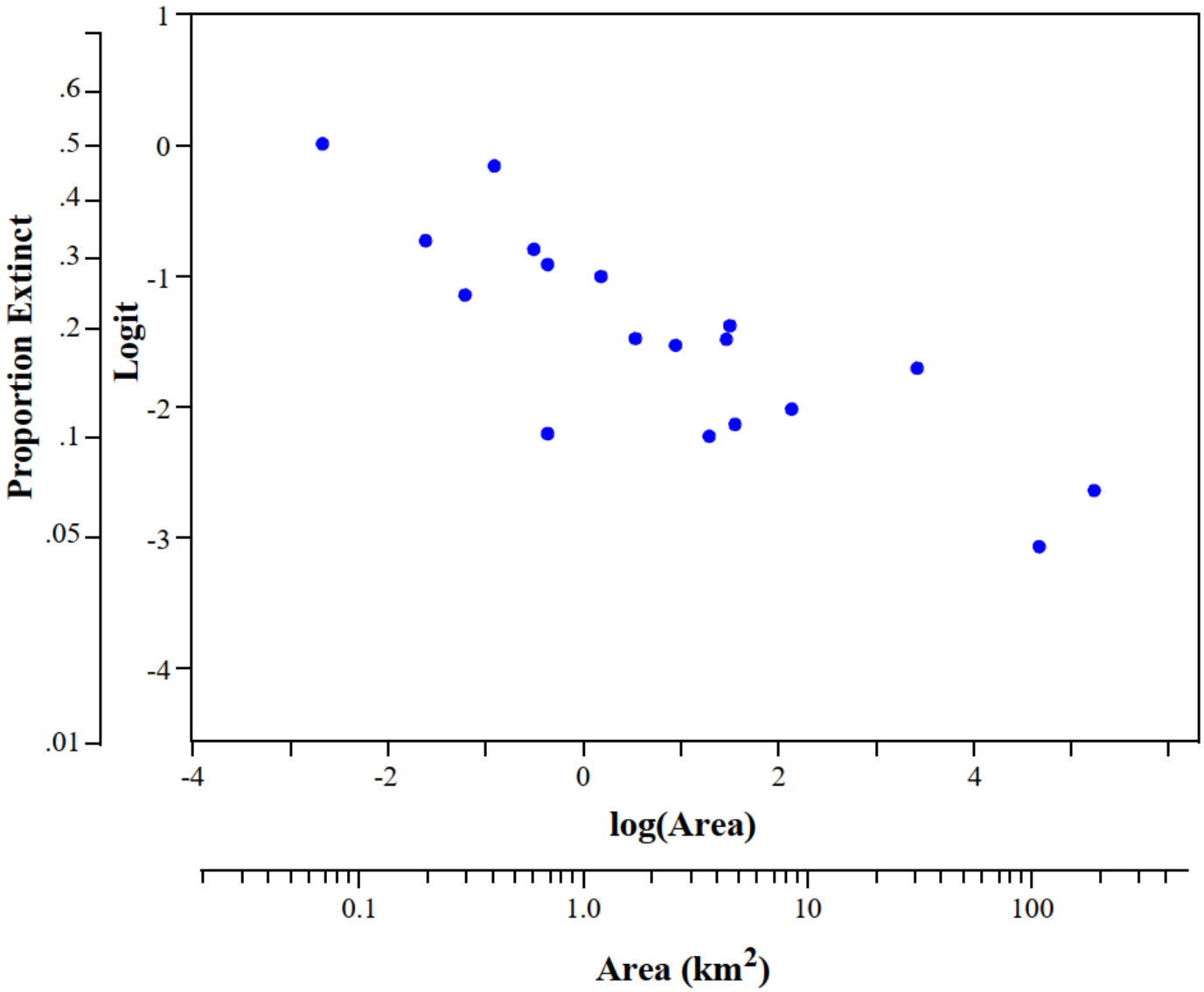
$$\log\left(\frac{X}{n-X}\right).$$

- These plots are quite useful for thinking about which explanatory variables might be useful, and/or for considering possible transformations.
- Note: We need to add a small amount like 0.5 to X if any counts are 0.

Let's take a look in R.

Takeaway: the logit appears to be linearly related to log(Area)

Extinctions of wading birds vs. island size in the Krunnit Island archipelago



Preliminary Results for Krunnit Island

Let's fit a binomial logistic regression model for extinction based on the Krunnit Islands data with log(Area) as our explanatory variable, i.e. our model is

$$\text{logit}(\pi) = \beta_0 + \beta_1 \log(Area)$$

Things to note:

- We need to **define a two-column response**: $(Y_i, m_i - Y_i)$
- In R: `resp <- cbind(Extinct, AtRisk - Extinct )`
- We still use `glm(..., family = binomial)`

To R!

Preliminary Results for Krunnit Island

The logistic regression model output is (in part):

```
Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.19620    0.11845 -10.099  < 2e-16 ***
larea       -0.29710    0.05485  -5.416 6.08e-08 ***
---

(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
```

Model Checking

Before we interpret this model, though, we should check the residuals to check for outliers.

- In binary logistic regression, residuals aren't all that useful because the responses are either zeroes or one, but in *Binomial* logistic regression, they can tell use something.

Another concern is **over-dispersion** (also called extra-Binomial variation) (more on that later).

Residuals in Binomial Logistic Regression

There are two types of residuals to consider in Binomial Logistic regression, **deviance residuals** and **Pearson residuals**:

- Deviance residuals are given by:

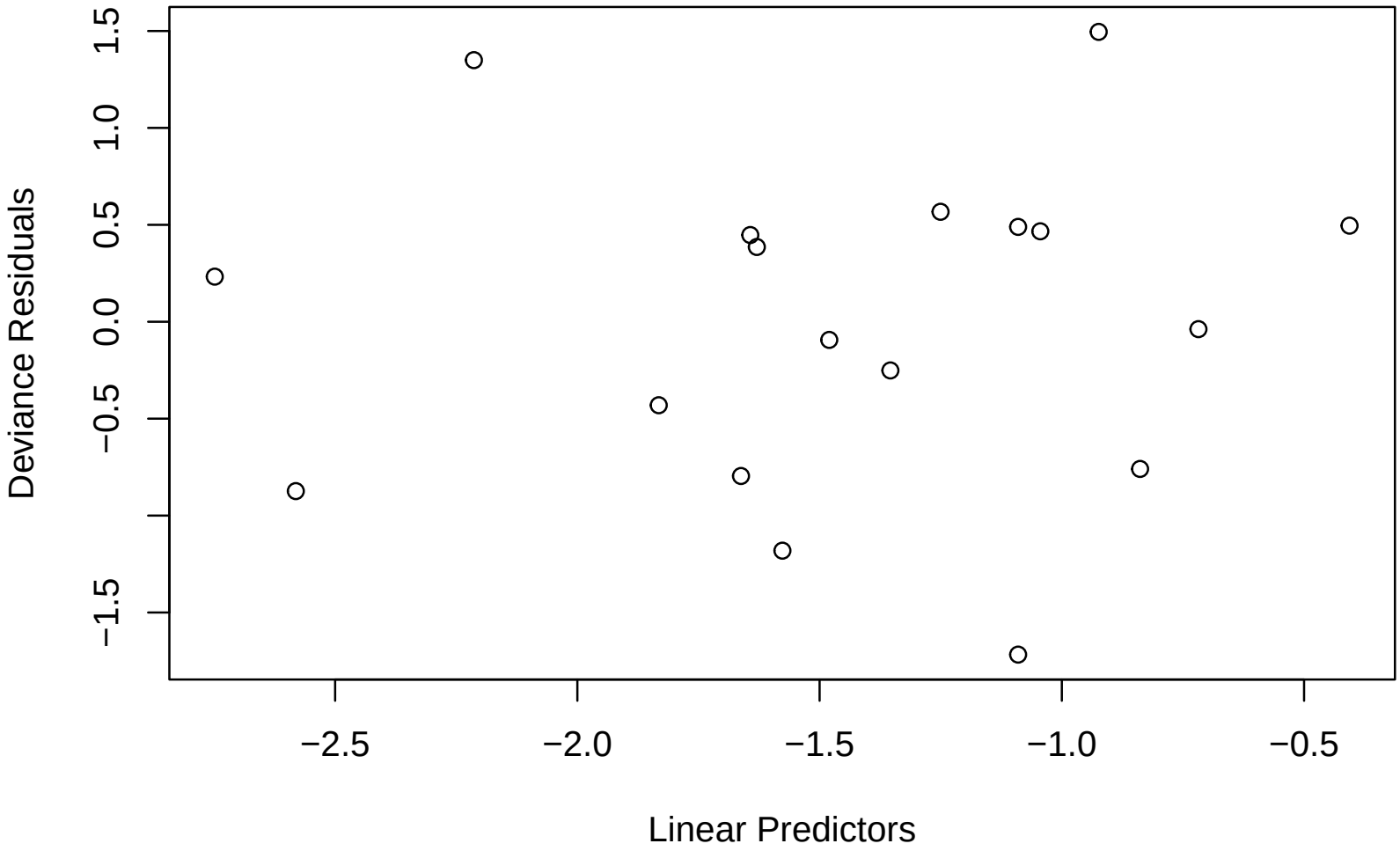
$$D_i = \text{sign}(Y_i - m_i\hat{\pi}_i)\sqrt{2\log\left(\frac{Y_i}{m_i\hat{\pi}_i}\right) + (m_i - Y_i)\log\left(\frac{m_i - Y_i}{m_i - m_i\hat{\pi}_i}\right)}$$

- Pearson residuals are given by:

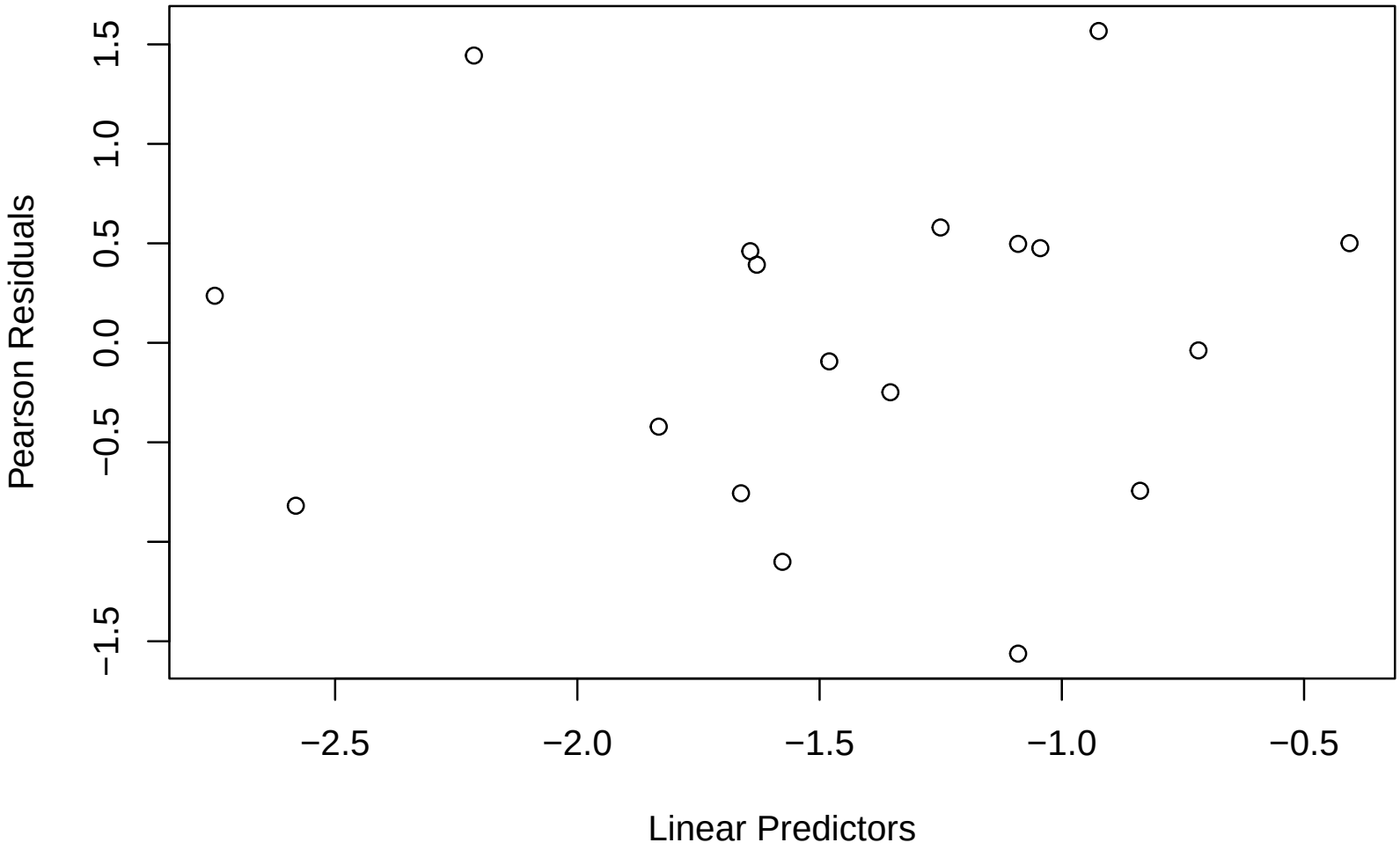
$$P_i = \frac{Y_i - m_i\hat{\pi}_i}{\sqrt{m_i\hat{\pi}_i(1-\hat{\pi}_i)}}.$$

For large m_i and if the model is correct, *both residuals tend to look like they come from standard normal distributions*.

Deviance Residuals



Pearson Residuals



Residuals in Binomial Logistic Regression

Although the Pearson residuals have a more appealing form than the deviance residuals, we more often use the deviance residuals.

- In fact, the **Residual Deviance** that R (and other software packages) reports is the sum of the squared deviance residuals.
- Remembering that residuals give us a sense for "what's left over" after we fit a model, it seems reasonable that small residual deviance is best.

Model Interpretation

Model interpretation is the same as in binary logistic regression.

Given the model:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1$$

we can say

- $\text{logit}(\pi)$ increases by β_1 when x_1 increases by 1
- The multiplicative change in the odds of success is $\exp(\beta_1)$.

Interpreting a logged explanatory variable

In the Krunnit Island example, one of our predictors is $\log(Area)$.

- Recall that in multiple linear regression **there's a specific interpretation for the coefficient of a logged explanatory variable**. The same is true for logistic regression.
- If $\text{logit}(\pi) = \beta_0 + \beta_1 \log(X_1)$, then a k-fold change in X_1 corresponds to a multiplicative change in the odds of k^{β_1}. (Let's verify this on the board.)

Krunnit Island Model Interpretation

In the Krunnit Island example, the estimated model is

$$\log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = -1.196 - 0.297 \times \log(Area)$$

- If $\log(Area)$ increases by 1, then $\text{logit}(\hat{\pi})$ increases by...
- If $Area$ doubles from A to $2A$, then $\text{logit}(\hat{\pi})$ increases by...
- How do the *odds* of extinction change when $Area$ doubles?

Krunnit Archipelago Example

We found that a 2-fold increase (i.e., doubling) of island area corresponds with an estimated multiplicative change in the odds of extinction of $2^{-0.297} = 0.813$.

- That is, on an island of area $2A$ the odds of extinction are 81.3% what they are on an island of area A.
- We might actually want to instead report the change in odds for a **decrease** in island size, i.e.:

  The odds of extinction for an island with area $A/2$ are estimated to be $0.5^{-0.297} = 1.23$ times the odds of extinction for an island with area A.

Recap

Material covered: Ch 21 of *The Sleuth*

- Binomial logistic regression
- Model checking with deviance and Pearson residuals
- Interpreting a logged explanatory variable