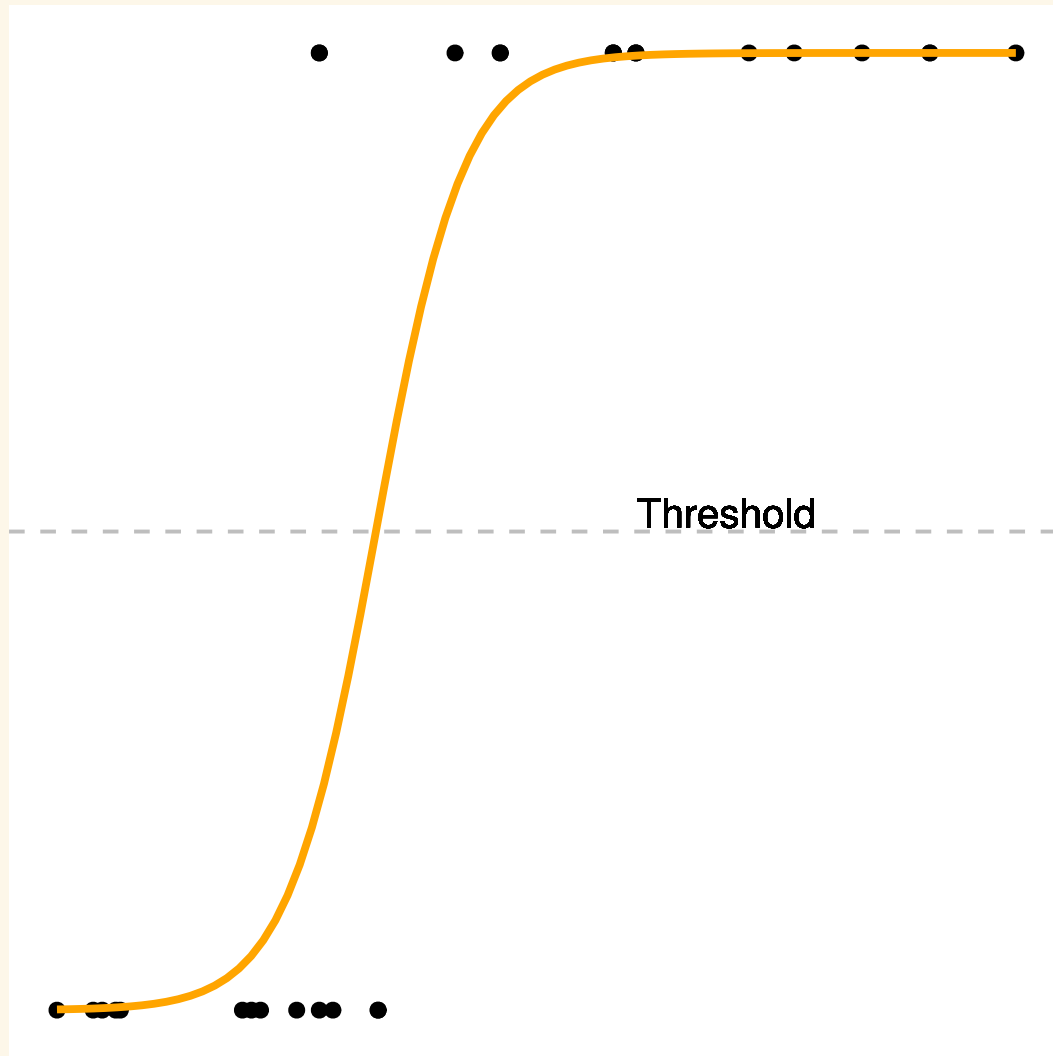


Logistic regression for binary responses: Diagnostics

Stat 230

May 18 2022

Overview



Today:

Binomial responses

Logistic model for binomial responses

EDA and inference

Binomial responses

- **Binomial** counts are defined as
 - Y_i = number of successes in m_i independent Bernoulli (success/failure) trials for case (unit) i
- Each case has predictor values:
 - $X_i = (x_{1,i}, \dots, x_{p,i})$ be the predictors for this case
- We want to use a logistic model for:
 - $\pi(X_i)$ is the probability of success for each of the m_i trials

Case study 21.1: Krunnit Island

In 1949, scientists recorded the number of "at risk" species on each island. Ten years later, they recorded how many of these species were extinct. We want to model the extinction rate for each island as a function of the island size.

- `Island` gives us an identifier of each island (case)
- `Area` is our predictor x_i for each island
- `AtRisk` is m_i , the number of animals available for extinction
- `Extinct` is y_i , the number (out of m_i) that went extinct

```
island <- case2101
glimpse(island)
Rows: 18
Columns: 4
$ Island   <fct> Ulkokrunni, Maakrunni, Ris
$ Area     <dbl> 185.80, 105.80, 30.70, 8.5
$ AtRisk   <int> 75, 67, 66, 51, 28, 20, 43
$ Extinct  <int> 5, 3, 10, 6, 3, 4, 8, 3, 5
```

Binomial responses

Our Binomial counts are modeled by a **Binomial** distribution:

$$Y_i \mid X_i \stackrel{\text{indep.}}{\sim} \text{Binom}(m_i, \pi(X_i))$$

The mean response for each case is

$$E(Y_i \mid X_i) = \mu_{y|x} = m_i \pi(X_i)$$

and the standard deviation is

$$SD(Y_i \mid X_i) = \sigma_{y|x} = \sqrt{m_i \pi(X_i) (1 - \pi(X_i))}$$

Logistic model for Binomial responses

- Logistic function for probabilities:

$$\pi(X_i) = \frac{e^{\eta_i}}{1 + e^{\eta_i}} = \frac{e^{\beta_0 + \beta_1 x_{1,i} + \dots + \beta_p x_{p,i}}}{1 + e^{\beta_0 + \beta_1 x_{1,i} + \dots + \beta_p x_{p,i}}}$$

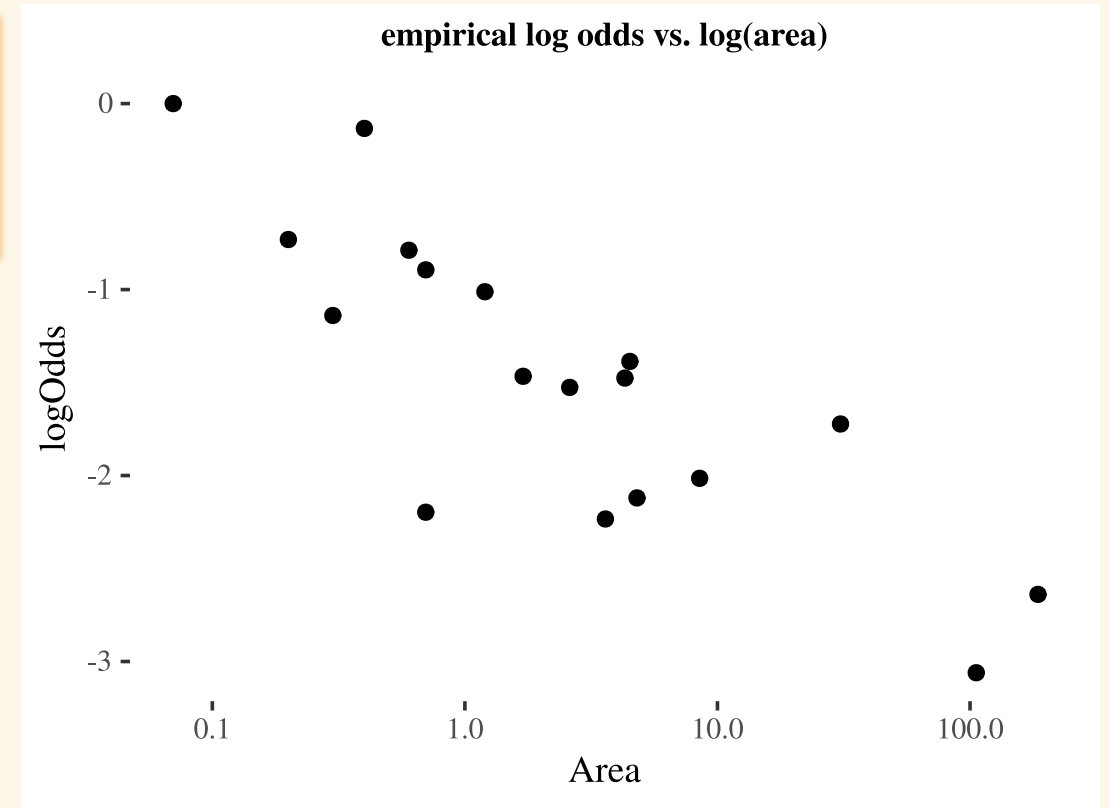
- logit (log odds) function for predictors:

$$\eta_i = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_p x_{p,i} = \ln \left(\frac{\pi(X_i)}{1 - \pi(X_i)} \right)$$

- Interpretation of a logistic model for a binomial response is **exactly the same** as a binary response model.

Case study 21.1: Krunnit Island

```
ggplot(island, aes(x = Area, y = logOdds)) +  
  geom_point() +  
  scale_x_log10() +  
  labs(title="empirical log odds vs. log(area)") +  
  theme(plot.title = element_text(hjust=0.5, size=9,  
                                   face='bold'))
```



Case study 21.1: Krunnit Island

- **Logistic model:** given an island size area_i , Y_i , the number of extinctions on island i , is a binomial variable

$$Y_i \mid \text{area}_i \sim \text{Binom}(m_i, \pi(\text{area}_i))$$

where $\pi(\text{area}_i)$ is the probability of extinction for each of the m_i at risk species on island i .

- Based on the EDA, we will fit the logit model:

$$\log\left(\frac{\pi(\text{area}_i)}{1 - \pi(\text{area}_i)}\right) = \beta_0 + \beta_1 \ln(\text{area}_i)$$

Inference for Binomial response models

Similarities between binomial and binary response logistic models:

- Interpretation of β in the two models is the same.
- Inference for models is the same
- "Wald" z-tests and confidence intervals for β parameters are the same.
- Drop-in-deviance model comparison tests are the same.
- R functions of fitted, predict, augment are the same.

Inference for Binomial response models

Differences between binomial and binary response logistic models:

- The formula for deviance G^2 is different because our probability model is (slightly) different.
- In binary models, Wald inference relies on "large n ". In binomial models, we either need "large n " and/or large m_i values. E.g. In the binomial model, $n = 5$ is fine if all $m_i = 1000$.
- The R function `glm` wants a response equal to the empirical proportion of successes, Y_i/m_i , along with the number of binomial trials m_i in our `glm` specification:

`glm(y/m ~ x1 + x2, family = binomial, weights = m, data =)`

Case study 21.1: Krunit Island

- **response:** the ratio of y (Extinct) to m (AtRisk)
- **weights:** m (AtRisk):

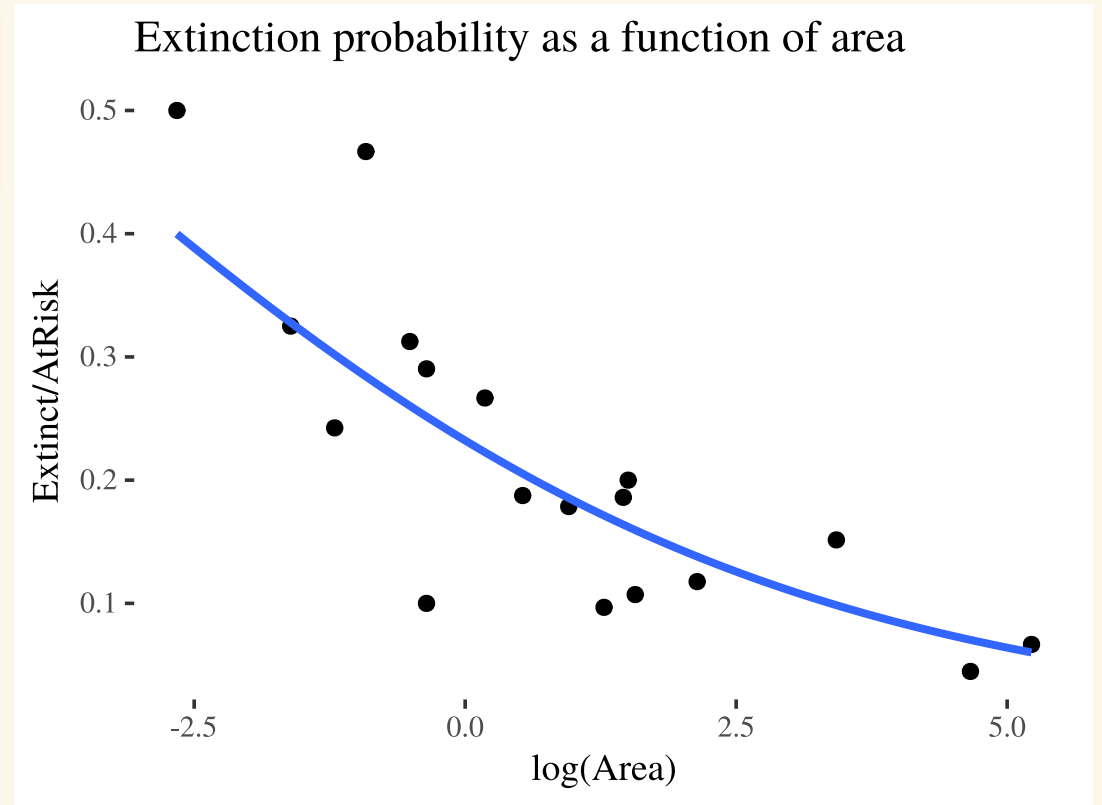
```
krunit.glm <- glm(Extinct/AtRisk ~ log(Area), family="binomial", weights=AtRisk, data=island)
tidy(krunit.glm, conf.int=TRUE)
# A tibble: 2 × 7
  term          estimate std.error statistic  p.value conf.low conf.high
<chr>         <dbl>     <dbl>     <dbl>    <dbl>   <dbl>   <dbl>
1 (Intercept)  -1.20      0.118     -10.1 5.58e-24  -1.43   -0.968
2 log(Area)    -0.297    0.0549     -5.42 6.08e- 8  -0.408  -0.192
```

- The estimated odds of extinction is

$$\frac{\hat{\pi}}{1 - \hat{\pi}} = e^{-1.1962 - 0.29710 \ln(\text{Area})} = e^{-1.1962} \text{Area}^{-0.29710}$$

Case study 21.1: Krunnit Island

```
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")
```



Case study 21.1: Krunit Island

```
tidy(krunnit.glm, conf.int=TRUE)
# A tibble: 2 × 7
  term          estimate std.error statistic  p.value conf.low conf.high
<chr>         <dbl>     <dbl>     <dbl>   <dbl>   <dbl>   <dbl>
1 (Intercept)   -1.20      0.118    -10.1  5.58e-24 -1.43   -0.968
2 log(Area)     -0.297    0.0549    -5.42  6.08e- 8 -0.408  -0.192
```

- The estimated odds of extinction for a 5 km² island is

$$\log\left(\frac{\hat{\pi}(x=5)}{1-\hat{\pi}(x=5)}\right) = -1.1962 - 0.29710 \log(5) \\ = -1.674364$$

```
predict(krunnit.glm, newdata=list(Area=5))
      1
-1.674365
```

- The estimated probability of extinction for at risk species on a 5 km² island is

$$\hat{\pi}(x=5) = \frac{e^{-1.1962-0.29710 \log(5)}}{1 + e^{-1.1962-0.29710 \log(5)}} \\ = \frac{e^{-1.674364}}{1 + e^{-1.674364}} = 0.1578432$$

```
predict(krunnit.glm, newdata=list(Area=5), type="resp
      1
0.157843
```

Case study 21.1: Krunnit Island

```
tidy(krunnit.glm, conf.int=TRUE)
# A tibble: 2 × 7
  term          estimate std.error statistic  p.value conf.low conf.high
<chr>         <dbl>     <dbl>     <dbl>   <dbl>   <dbl>   <dbl>
1 (Intercept)  -1.20      0.118    -10.1  5.58e-24 -1.43   -0.968
2 log(Area)    -0.297     0.0549    -5.42  6.08e- 8 -0.408  -0.192
```

- The effect of area on extinction rates is statistically significant (Wald $z = -5.416$, $p < 0.0001$).
- Effect of doubling island size?

$$OR = \frac{\hat{od}ds(2 \times \text{Area})}{\hat{od}ds(\text{Area})} = 2^{-0.29710} = 0.814, \quad 100\%(0.814 - 1) = -18.6$$

- Doubling an island area is associated with an 18.6% reduction in the odds of extinction of at risk species.
- We are 95% confident that doubling the area of an island is associated with anywhere from a 12.3% to 24.5% decrease in extinction rates.

Case study 21.1: Krunit Island

```
tidy(krunnit.glm, conf.int=TRUE)
# A tibble: 2 × 7
  term          estimate std.error statistic  p.value conf.low conf.high
<chr>         <dbl>     <dbl>     <dbl>   <dbl>   <dbl>   <dbl>
1 (Intercept)  -1.20      0.118    -10.1  5.58e-24 -1.43   -0.968
2 log(Area)    -0.297    0.0549    -5.42  6.08e- 8 -0.408  -0.192
```

- Effect of a 10% increase in island size?

$$OR = 1.1^{-0.29710} = 0.972, \quad 1.1^{-0.408} = 0.962, \quad 1.1^{-0.192} = 0.982$$

- We are 95% confident that a 10% increase in the area of an island is associated with a 1.8% to 3.8% decrease in the odds of extinction for at risk species.

Example: NES revisited

- Suppose that the NES data was collected via a **stratified** random sample
 - in 1980: separate random samples of people were taken within each region
 - in 2000: separate random samples of people were taken within each region
- In any given year/region combo $i = 1, \dots, 8$
 - m_i = number of people surveyed in that year/region (fixed sample size)
 - Y_i = number of Democrat leaning people surveyed in that year/region

Example: NES revisited

```
nes_binom_data <- nes %>%  
  group_by(region, year) %>% # sample size by year/region  
  summarize(m = n(), Dem_count = sum(dem)) # number Dem  
nes_binom_data
```

```
# A tibble: 8 × 4  
# Groups:   region [4]  
  region year      m Dem_count  
  <chr>  <chr>  <int>    <int>  
1 NC    year1980  262     128  
2 NC    year2000  301     162  
3 NE    year1980  226     112  
4 NE    year2000  201     122  
5 S     year1980  380     228  
6 S     year2000  426     210  
7 W     year1980  186      96  
8 W     year2000  250     136
```

- m : the number of people per year/region ($n()$ counts the number of rows for each group)
- **Dem_count**: the number of Democrats per year/region
- Binomial count data with region and year predictors

Example: NES revisited

$$Y_i \mid \text{year, region} \sim \text{Binom}(m_i, \pi_i)$$

where π_i is the probability of Dem in that year/region combo i .

- We can model π as a function of year and region
- we get the same results for both the binomial and binary versions of this data!

Example: NES revisited

Binary version: using person-level responses and our `dem` indicator of Democrat

```
nes_glm_binary <- glm(dem ~ region*year,  
                      data = nes,  
                      family = binomial)
```

Binomial version: using aggregated year/region level counts and our `Dem_count/m` proportion Democrats for each year/region combo

```
nes_glm_binom <- glm(Dem_count/m ~ region*year,  
                    weights = m,  
                    data = nes_binom_data,  
                    family = binomial)
```

Example: NES revisited

```
summary(nes_glm_binary)

Call:
glm(formula = dem ~ region * year, family = binomial, data = nes)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3666  -1.2049   0.9993   1.1131   1.1969

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.04581    0.12359  -0.371   0.71090
regionNE         0.02811    0.18159   0.155   0.87698
regionS          0.45127    0.16199   2.786   0.00534 **
regionW          0.11035    0.19184   0.575   0.56515
yearyear2000     0.19893    0.16924   1.175   0.23982
regionNE:yearyear2000 0.25334    0.25923   0.977   0.32842
regionS:yearyear2000 -0.63257    0.22136  -2.858   0.00427 **
regionW:yearyear2000 -0.08701    0.25748  -0.338   0.73540
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3083.3  on 2231  degrees of freedom
Residual deviance: 3065.5  on 2224  degrees of freedom
AIC: 3081.5

Number of Fisher Scoring iterations: 4
```

Example: NES revisited

```
summary(nes_glm_binom)
```

Call:

```
glm(formula = Dem_count/m ~ region * year, family = binomial,  
    data = nes_binom_data, weights = m)
```

Deviance Residuals:

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

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-0.04581	0.12359	-0.371	0.71090	
regionNE	0.02811	0.18159	0.155	0.87698	
regionS	0.45127	0.16199	2.786	0.00534	**
regionW	0.11035	0.19184	0.575	0.56515	
yearyear2000	0.19893	0.16924	1.175	0.23982	
regionNE:yearyear2000	0.25334	0.25923	0.977	0.32842	
regionS:yearyear2000	-0.63257	0.22136	-2.858	0.00427	**
regionW:yearyear2000	-0.08701	0.25748	-0.338	0.73540	

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1.7792e+01 on 7 degrees of freedom

Residual deviance: 1.7764e-14 on 0 degrees of freedom

AIC: 64.266

Number of Fisher Scoring iterations: 2

Recall: Inference for Binomial response models

Differences between binomial and binary response logistic models:

- The formula for deviance G^2 is different because our probability model is (slightly) different.
- Binomial version assumes that the year/region sample size counts m_i are fixed
- Binomial version we have a sample size of $n = 8$ year/region observations
- In Binomial version we have no degrees of freedom left with 8 parameters in the interaction model!
- The predicted probability $\hat{\pi}_i$ is just the sample proportion of Dem for each year/region

Example: NES revisited

- Which model (binary vs binomial) to use?
- Assuming m_i are fixed sample sizes (e.g. a stratified design), doesn't much matter*
- unless you'd like to incorporate additional individual level predictors of a person's probability of voting Democratic → use binary

```
head(nes) # more individual level data!
```

	year	age	gender	race	region	income	union	dem	educ
1	year1980	70	male	black	S	lower 1/3	no	1	HS or less
2	year1980	67	male	white	NC	middle 1/3	yes	1	HS or less
3	year1980	47	female	black	S	lower 1/3	no	1	HS or less
4	year1980	52	female	white	W	upper 1/3	yes	0	College
5	year1980	30	female	white	NC	upper 1/3	no	1	HS or less
6	year1980	37	male	black	NC	upper 1/3	no	1	College

Your Turn 1

05:00



- Go over to the in class activity file
- Go over the class activity in your group