# STA 303/1002-Methods of Data Analysis II

## Sections L0101& L0201, Winter 2018

**Shivon Sue-Chee**

February 6-8, 2018

# STA 303/1002: Class 11- Binomial Logistic Regression

- Case Study IV: Island size and bird extinction
  - R syntax
  - Data visualization
  - Interpreting coefficients
  - Wald procedures

- Principle of the week: *K-Keep, I-It, S-Simple, S-Stupid*(US Navy, 1960)

# Plot

Q: How would the plot of estimated probabilities change if we modelled probability of death rather than survival?

## Over 50yrs

Q: Should one be reluctant to draw conclusions about the ratio of male and female odds of survival for the Donner Party members over 50?

# Other Model Fit Statistics

- Two popular fit statistics: AIC and BIC; combines log-likelihood with a penalty.
- Useful for comparing models with same response and same data
- Extends from normal regression to GLMs
  1. Akaike's Information Criterion (AIC)

$$AIC = -2\log\mathcal{L} + 2(p+1)$$

  2. Schwarz's (Bayesian Information) Criterion (BIC)

$$BIC = -2\log\mathcal{L} + (p+1)\log N$$

  where
  - $p$-number of explanatory variables, and
  - $N$=sample size

# Model Fit Statistics: AIC and BIC

- Smaller is better!
- BIC applies stronger penalty for model complexity than AIC

- AIC Rule of Thumb:
  - One model fits **better** than another if difference in AIC's $> 10$
  - One model model is essentially **equivalent** to another if the difference in AIC's $< 2$

## Using AIC: Case Study III Example

- Fitted models are based on same response and data.
- Based on AIC, choose a 'best' model.

| Model | Variables | AIC | BIC |
|---|---|---|---|
| 1 | {age,sex} | 57.256 | 62.676 |
| 2 | {age,sex,age*sex,$age^2$,$age^2$*sex} | 57.361 | 68.201 |
| 3 | {age,sex,age*sex,$age^2$ } | 55.830 | 64.863 |
| 4 | {age,sex,age*sex} | 55.346 | 62.573 |

Results:

- Difference in AIC between 1 and 3 is within 2
- There is some indication that 2 is worse than 3 and 4.
- Choose Model 1 (the simplest)

## Related R packages and functions

- Packages:
  - `aod`: analysis of over-dispersed data
  - `ggplot2`: graphics
  - `Sleuth3`: data sets for Ramsey and Schafer's text
  - `effects`: effects displays for GLM and other models
- Functions:
  - `confint()`
  - `coef()`
  - `vcov()`
  - `wald.test()`
  - `AIC()`
  - `BIC()`

# Binomial Logistic Regression

## Suppose $Y \sim \text{Binomial}(m, \pi)$

- $Y$-binomial count of the number of "successes"

$$P(Y = y) = \binom{m}{y} \pi^y (1-\pi)^{m-y}, \quad y = 0, 1, \ldots, m$$

  - Link to Bernoulli:
    $Y = \sum_{i=1}^{m} X_i$ if $X_i$'s are independent Bernoulli($\pi$) r.v.s.
    *Assume that $\pi$ is the same for each Bernoulli trial.*

- Mean: $E(Y) = m\pi$
- Variance: $\text{Var}(Y) = m\pi(1-\pi)$

## Suppose $Y \sim \text{Binomial}(m, \pi)$

- Consider modelling

$$\frac{Y}{m}$$

  - the proportion of "successes" out of $m$ independent Bernoulli trials.
- where,
  - $\text{E}\left(\frac{Y}{m}\right) = \pi$
  - $\text{Var}\left(\frac{Y}{m}\right) = \frac{\pi(1-\pi)}{m}$

# Case Study IV Data Example

- Data: counts of bird species for 18 Krunnit Islands off Finland.

\# of successes

| i | $x_i$ area | $m_i$ nspecies | $y_i$ nextinct | $\pi_i$ observed proportion |
|---|---|---|---|---|
| ISLAND | AREA | ATRISK | EXTINCT | |
| Ulkokrunni | 185.8 | 75 | 5 | 5/75 |
| Maakrunni | 105.8 | 67 | 3 | 3/67 |
| Ristikari | 30.7 | 66 | 10 | ⋮ |
| Isonkivenletto | 8.5 | 51 | 6 | |
| ... | | | | |
| Tiirakari | 0.2 | 40 | 13 | 13/40 |
| Ristikarenletto | 0.07 | 6 | 3 | 3/6 = 0.5 |

- AREA- area of island in $km^2$, $x_i$
- ATRISK- number of species on each island in 1949, $m_i$
- EXTINCT- number of species no longer found on each island in 1959, $y_i$

## Case Study IV: Model

$m_1$ iid species

$x_1, \dots, x_{m_1}$

$y_1 = \sum_{i=1}^{m_1} x_i$

Island 1

$m_1$

- $\pi_i$- probability of 'extinction' for each island.
- ① *Assume that this is the same for each species of bird on a particular island.*

  $\pi_i$
- ② *Assume species survival is independent.* Then

$0 \le \pi_i \le 1$

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

$0 < \pi_i < 1$

- Unlike Case III- Donner party binary logistic example, we can estimate $\pi_i$ from the data.

Bernoulli $\rightarrow$ Binomial counts.

$\pi = \begin{cases} 0 \\ 1 \end{cases}$

Proportions vs Percentages

(0, 1) (0, 100).

cts.

# Case Study IV: Model

Data

- Observed response proportion:

observed counts.

$$\bar{\pi}_i = \frac{y_i}{m_i}$$

total

$\rightarrow \bar{\pi}_i$

- Observed or Empirical logits: (S- "saturated")

$$\log\left(\frac{\bar{\pi}_{S,i}}{1-\bar{\pi}_{S,i}}\right) = \log\left(\frac{y_i}{m_i - y_i}\right)$$

Estimates

- Proposed Model: $\boxed{\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 Area_i,}$ $i = 1, \ldots, 18$

$\rightarrow \hat{\pi}_i$

- AIM:
  - Learn how to create nature preserves that help endangered species.
  - Are large or small preserves better?

## Case Study IV: Initial assessment of data

Visuals

- Plot observed logits versus area to see if a linear relationship seems appropriate.
- From that plot, we decide to look at log(Area) instead.
- The relationship between empirical logits and log(Area) seems linear.
- Hence, we fit

$$\boxed{\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 \log(Area_i),} \quad i = 1, \ldots, 18$$

## Case Study IV: R syntax

- In R, the model formula has the form:

$$\texttt{cbind}(\texttt{y}_\texttt{i}, \texttt{m}_\texttt{i} - \texttt{y}_\texttt{i}) \sim \texttt{log(Area)}$$

x

Need to specify both:

- $y_i$ - number of successes and
- $(m_i - y_i)$ - number of failures

## Case Study IV: Model Summary

- Number of observations: 18
- Number of coefficients: 2
- Fitted model:

$$\boxed{\text{logit}\,(\hat{\pi}) = \overset{\hat{\beta}_0}{-1.196} - \overset{\hat{\beta}_1}{0.297}\log(Area)}$$

## Case Study IV: Wald procedures

(Similar test as in binary logistic regression)

- Hypotheses:

$$H_0 : \beta_1 = 0 \text{ vs } H_a : \beta_1 \neq 0$$

- Test statistic:

$$z = \frac{-0.2971}{0.0549} = -5.42 \sim N(0,1) \text{ or } z^2 = 29.3 \sim \chi_1^2$$

$= (-5.42)^2$

- P-value $< 0.0001$
- Conclusion: Strong evidence that coefficient of log(Area) is not zero. Evidence that extinction probabilities are associated with island area.

- 95% CI for $\beta_1$:

$$-0.2971 \pm 1.96(0.0549) = (-0.40, -0.19)$$

does not include 0

$\hat{\beta}_1 \pm 1.96\,(se(\hat{\beta}_1))$

## Case Study IV: Interpretation of $\beta_1$

$\log a = b$

$\Rightarrow a = e^b$

$\log_{10} 10 = 1 \Rightarrow 10^1$

$\log_{10} 100 = 2 \Rightarrow 10^2 = 10(10)$

- **Model**:

$$\mathrm{logit}(\pi) = \beta_0 + \beta_1 \underline{\log(x)}$$
$$\Longrightarrow \frac{\pi}{1-\pi} = e^{\beta_0} e^{\beta_1 \log(x)} = e^{\beta_0} x^{\beta_1}$$

- **Interpretation**: Hence, changing $x$ by a factor of $h$, changes the odds by a multiplicative factor of $\underline{h^{\beta_1}}$.

$$\frac{e^{\beta_0} x^{\beta_1}}{e^{\beta_0}(1)}$$

$h = \frac{1}{2}$

$x \rightarrow xh$

$h = 2$

## Case Study IV: Interpretation of $\beta_1$

$\frac{\pi}{1-\pi}$

- Example 1: Halving island area changes odds by a factor of $0.5^{-0.2971} = 1.23$.
  $\frac{1}{2}$
  Therefore, the odds of extinction on a smaller island are 123% of the odds of extinction on an island double its size.
  In other words, halving of area is associated with an increase in the odds of extinction by an estimated 23%.
  An approximate 95% confidence interval for the percentage change in odds is 14% to 32%.

- Example 2: Doubling island area changes odds by a factor of $2^{-0.2971} = 0.81$.
  $\frac{2}{1}$
  Therefore, the odds of extinction for an at-risk species on a larger island are only 81% of the odds of extinction for such a species on an island half its size.

## Case Study IV: Estimating probability of extinction

- Q: Estimate the probability of extinction for a species on the Ulkokrunni island.
- Fitted Model (M):

$$\text{logit}(\hat{\pi}_{M,i}) = -1.196 - 0.297 \log(Area_i)$$

- For Ulkokrunni island, $i = 1$ and Area=185.5 $km^2$, then $\text{logit}(\hat{\pi}_{M,1}) =$ $-1.196 - 0.297 \log(185.5) = *$

Est. prob. $\hat{\pi}_{M,1} = \dfrac{e^{*}}{1+e^{*}}$

- Compared to the response proportion, $\bar{\pi}_{S,1} = \frac{5}{75} = 0.067$.

Obs. prob.

# STA303/1004 - Class 11 R Markdown

February 8, 2018

# Case Study IV: The Data

Get the data (from R library):

```
#load Sleuth3 R data library; see case2101
library(Sleuth3); krunnit = case2101
str(krunnit)
```

```
## 'data.frame':   18 obs. of  4 variables:
##  $ Island : Factor w/ 18 levels "Hietakraasukka",..: 16 6 11 2 1 3 4 7 15 12
##  $ Area   : num  185.8 105.8 30.7 8.5 4.8 ...
##  $ AtRisk : int  75 67 66 51 28 20 43 31 28 32 ...
##  $ Extinct: int  5 3 10 6 3 4 8 3 5 6 ...
```

$x_i$

$m_i$

$y_i$

# Case Study IV: New variables

Get the data (from R library):

```
attach(krunnit); head(krunnit)
```

$m_i$ $y_i$

```
##              Island  Area AtRisk Extinct
## 1        Ulkokrunni 185.8     75       5
## 2         Maakrunni 105.8     67       3
## 3         Ristikari  30.7     66      10
## 4 Isonkivenletto   8.5     51       6
## 5 Hietakraasukka   4.8     28       3
## 6          Kraasukka   4.5     20       4
```

NExtinct $= m_i - y_i$

$75-5=70$

$67-3=64$

$\vdots$

$20-4=16$

$\bar{\pi}_i$ $\to$ $\hat{\pi}_i$

$5/75$

$3/67$

$\log(\bar{\pi}_i / 1-\bar{\pi}_i)$

$(5/75) / 1-(5/75)$

$\vdots$

Empirical logits

```
logitpi<-log(Extinct/AtRisk/(1-(Extinct/AtRisk))) #observed logits
logarea<-log(Area) # log transformed Area
NExtinct<-AtRisk-Extinct
pis<-Extinct/AtRisk
```

$\log\left(\frac{\bar{\pi}_i}{1-\bar{\pi}_i}\right)$

log(Area)

$m_i - y_i$

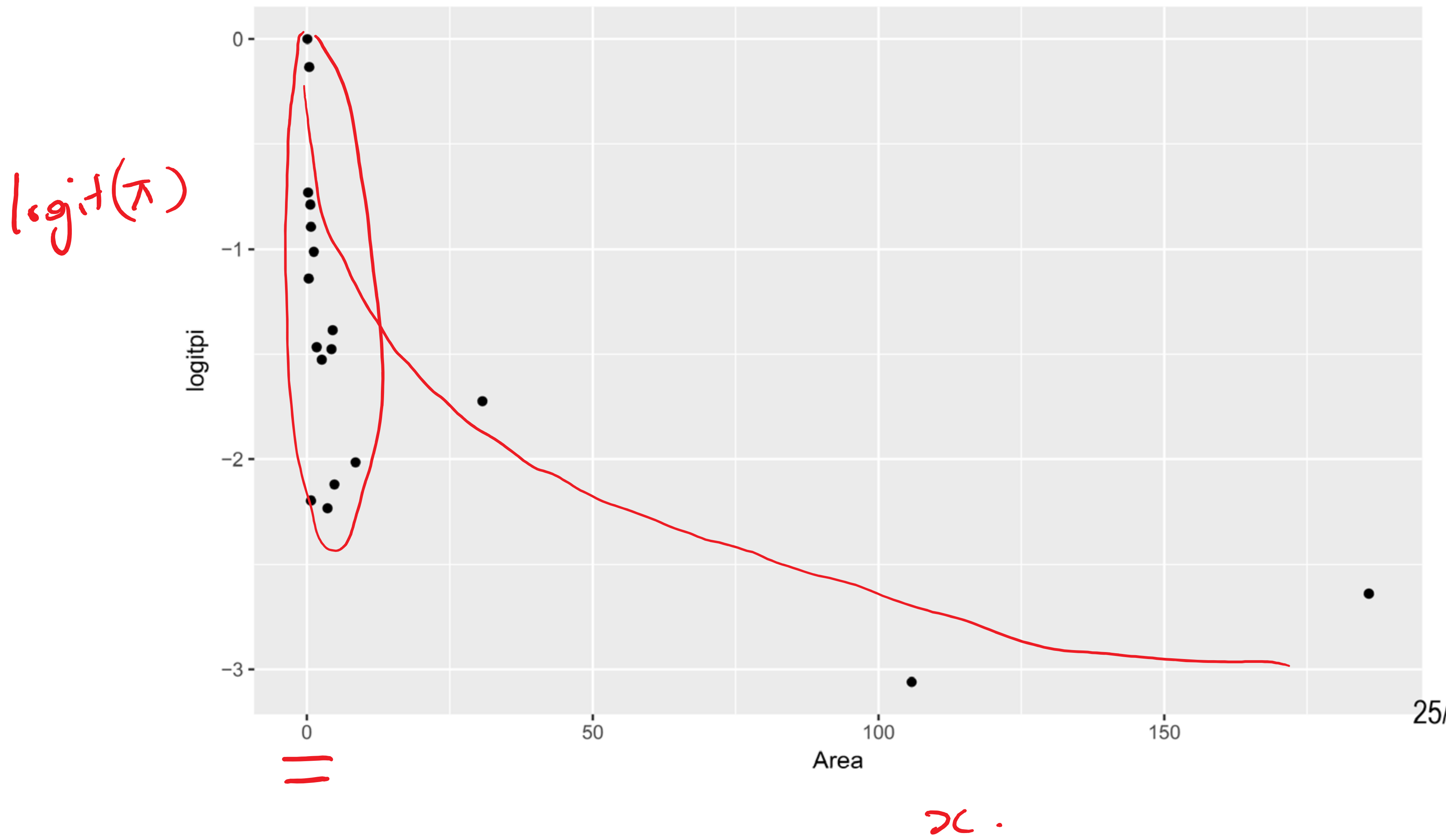
Eg, $\log\left(\frac{5/75}{1-5/75}\right)$

We can compare: $\bar{\pi}_i$ with $\hat{\pi}_i$

: $\log \frac{\bar{\pi}_i}{1-\bar{\pi}_i}$ with $\log \frac{\hat{\pi}_i}{1-\hat{\pi}_i}$

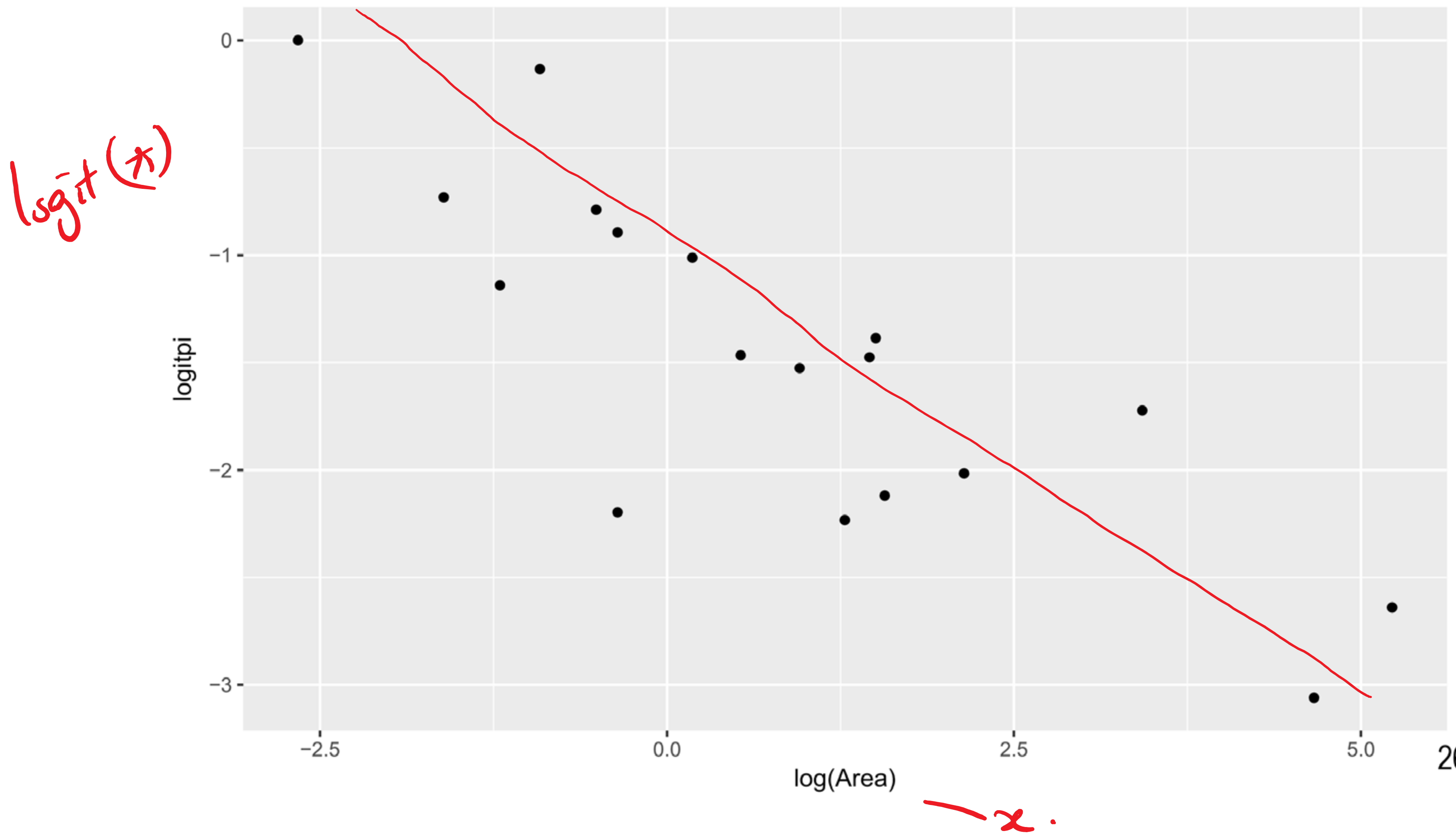
## Case Study IV: Visualizing the data

```
library(ggplot2)
ggplot(krunnit, aes(x=Area, y=logitpi))+geom_point()
```



## Case Study IV: Visualizing the data

```
ggplot(krunnit, aes(x=log(Area), y=logitpi))+geom_point()
```



$y \sim x$

# Case Study IV: Logisitc Model with logged explanatory variable

```
fitbl<-glm(cbind(Extinct,NExtinct)~log(Area), family=binomial, data=krunnit)
summary(fitbl)
```

# of successes $y_i$ # of failures $m_i - y_i$ $\rightarrow m_i$

```
##
## Call:
## glm(formula = cbind(Extinct, NExtinct) ~ log(Area), family = binomial,
##     data = krunnit)
##
## 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
```

$\hat{\beta}_j$ $se(\hat{\beta}_j)$

$z^2 \sim \chi^2_1$

$(0.05485)^2 = 0.003$

## Case IV: Deviance test and Estimated Var-Cov of $\beta$

```
anova(fitbl, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: cbind(Extinct, NExtinct)
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                         17     45.338
## log(Area)  1   33.277        16     12.062 7.994e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

→ Used for Global LRT.

```
print(vcov(fitbl))
```

```
##              (Intercept)    log(Area)
## (Intercept)  0.014029452 -0.002602237
## log(Area)   -0.002602237  0.003008830
```

$var(\hat{\beta}_1) = (se(\hat{\beta}_1))^2$

# Case IV: Wald tests in R

```
library(aod) # Analysis of Overdispersed Data
wald.test(Sigma=vcov(fitbl), b=coef(fitbl), Terms=2)
```

$\mathrm{var}(\hat{\beta}_j)$ $\hat{\beta}_j$

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 29.3, df = 1, P(> X2) = 6.1e-08
```

$(-5.42)^2 = 29.3$

## Case IV: Confidence Intervals for $\beta$'s

```
CL=cbind(bhat=coef(fitbl), confint.default(fitbl))  # 95% CI for betas
CL
```

```
##                   bhat      2.5 %     97.5 %
## (Intercept) -1.1961955 -1.4283454 -0.9640456
## log(Area)   -0.2971037 -0.4046132 -0.1895942
```

$\hat{\beta}_1 \pm 1.96\ se(\hat{\beta}_1)$

```
2^(CL) # doubling Area
```

```
##                  bhat     2.5 %    97.5 %
## (Intercept) 0.4364247 0.3715568 0.5126174
## log(Area)   0.8138847 0.7554388 0.8768524
```

```
.5^(CL) # halving Area
```

```
##                 bhat    2.5 %   97.5 %
## (Intercept) 2.291346 2.691379 1.950773
## log(Area)   1.228675 1.323734 1.140443
```

## Case IV: Estimated probabilities of extinction per island

```
phats<-predict.glm(fitbl, type="response") # estimated probability of extinctio
options(digits=4)
rbind(Extinct, NExtinct, pis,phats)
```

```
##                  1        2        3       4       5      6       7
## Extinct    5.00000  3.00000 10.00000  6.0000  3.0000  4.000  8.0000
## NExtinct  70.00000 64.00000 56.00000 45.0000 25.0000 16.000 35.0000
## pis        0.06667  0.04478  0.15152  0.1176  0.1071  0.200  0.1860
## phats      0.06017  0.07036  0.09854  0.1380  0.1595  0.162  0.1639
##                  8       9      10      11      12      13      14     15
## Extinct    3.00000  5.0000  6.0000  8.0000  2.0000  9.0000  5.0000 7.0000
## NExtinct  28.00000 23.0000 26.0000 22.0000 18.0000 22.0000 11.0000 8.0000
## pis        0.09677  0.1786  0.1875  0.2667  0.1000  0.2903  0.3125 0.4667
## phats      0.17125  0.1854  0.2052  0.2226  0.2516  0.2516  0.2603 0.2842
##                16      17     18
## Extinct    8.0000 13.0000 3.0000
## NExtinct  25.0000 27.0000 3.0000
## pis        0.2424  0.3250 0.5000
## phats      0.3019  0.3278 0.3998
```

$y_i$

$m_i - y_i$

$\bar{\pi}_i$

$\hat{\pi}_i$

Observed

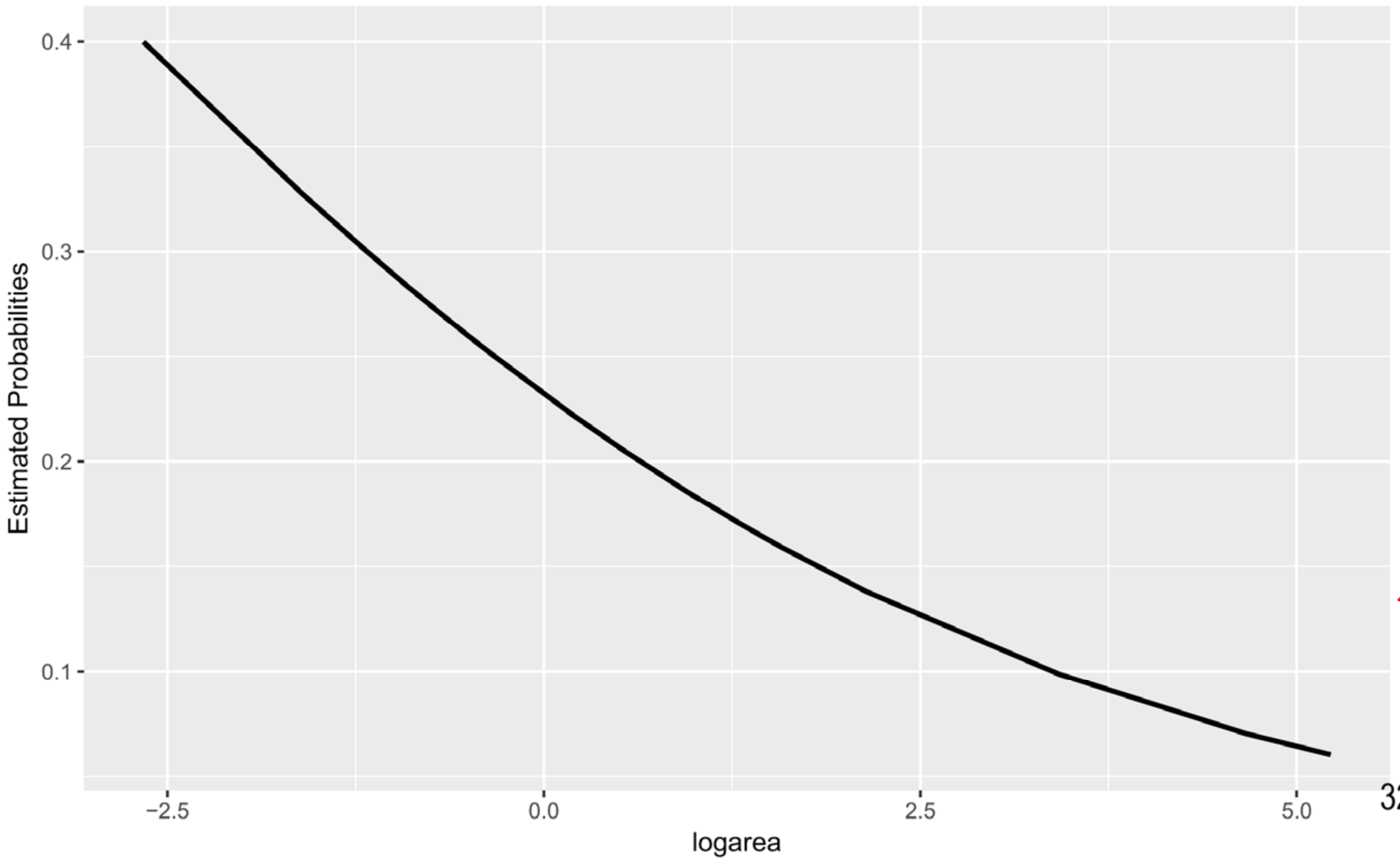
$p_i s = \dfrac{\text{Extinct}}{\text{Extinct}+\text{NExtinct}} = y_i / m_i = \bar{\pi}_i$

Estimated

phats. $\rightarrow \hat{\pi}_i$

# Case IV Effect Plot

```
ggplot(krunnit,aes(x=logarea, y=phats))+ylab("Estimated Probabilities")+
  geom_line(size=1)
```




→ larger area associated to lower odds of extinction