STA 303/1002-Methods of Data Analysis II

Sections L0101& L0201, Winter 2019

Dr. Shivon Sue-Chee

UNIVERSITY OF TORONTO

Week 8: March 4-8, 2019

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)

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(π) r.v.s.
    Assume that π 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,
  - $\mathrm{E}\left(\frac{Y}{m}\right) = \pi$
  - $\mathrm{Var}\left(\frac{Y}{m}\right) = \frac{\pi(1-\pi)}{m}$

Case Study IV Data Example

@1949

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

i =1,...,18	x_i	m_i	y_i @1959
	area	nspecies	nextinct
ISLAND	AREA	ATRISK	EXTINCT
Ulkokrunni	185.8	75	5
Maakrunni	105.8	67	3
Ristikari	30.7	66	10
Isonkivenletto	8.5	51	6
...			
Tiirakari	0.2	40	13
Ristikarenletto	0.07	6	3

- 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

π_1 — 1st island
π_2
⋮
π_{18}

- π_i - probability of 'extinction' for each island. *Assume that this is the same for each species of bird on a particular island.*
- *Assume species survival is independent.* Then

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

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

Case Study IV: Model

- Observed response proportion:

$$\bar{\pi}_{i,s} = \frac{y_i}{m_i}$$

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

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

- Proposed Model:

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

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

Case Study IV: Initial assessment of data

logit (π) Area

- 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:

$$\mathtt{cbind(y_i, m_i - y_i) \sim log(Area)}$$

Need to specify both:

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

Case Study IV: Model Summary

summary(fittedmodel)

- Number of observations: 18 islands
- Number of coefficients: 2 $p=1$, $p+1=2$
- Fitted model:

$$\boxed{\text{logit}\,(\hat{\pi}) = -1.196 - 0.297 \log(\mathit{Area})}$$

Case Study IV: Wald procedures

(Similar test as in binary logistic regression)

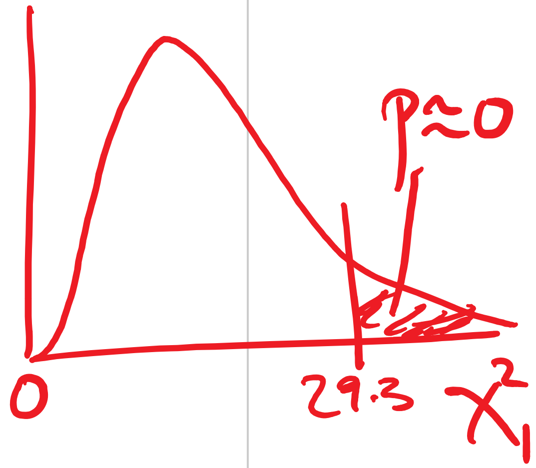
- Hypotheses:

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

- Test statistic:

$P(\chi^2_1 \geq 29.3)$

$$z = \frac{-0.2971}{0.0549} = -5.42 \sim N(0,1) \text{ or } z^2 = 29.3 \sim \chi^2_1$$



- 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 β_1:

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

not incl. 0

Case Study IV: Interpretation of β_1

- Model:

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

- Interpretation: Hence, changing x by a factor of h, changes the odds by a multiplicative factor of h^{β_1}.

$$e^{\beta_0} (x)^{\beta_1}$$
$$e^{\beta_0} (x h)^{\beta_1}$$
$$= e^{\beta_0} x^{\beta_1} h^{\beta_1}$$

Case Study IV: Interpretation of β_1

- Example 1: Halving island area changes odds by a factor of $0.5^{-0.2971} = 1.23$.
  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 $2^{-0.2971} = 0.81$.
  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.

$x_1 =$

$x_2 = \frac{1}{2} x_1$

$\left(\frac{1}{2}\right)^{\beta_1}$

h^{β_1}

$2^{\hat{\beta}_1}$

Case Study IV: Estimating probability of extinction

$\hat{\pi}_i$

- 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
  $\text{logit}(\hat{\pi}_{M,1}) = -1.196 - 0.297 \log(185.5) = *$

  $\hat{\pi}_{M,1} = 0.06 = \frac{e^*}{1+e^*}$
- Compared to the response proportion, $\bar{\pi}_{S,1} = \frac{5}{75} = 0.067$.

Checking Model Assumptions

Model Assumptions for Binomial Logistic Regression

1. Underlying probability model for response is Binomial.
   - Variance is not constant; is a function of the mean.

2. Observations are independent.

3. The form of the model is correct
   - Linear relationship between logits and explanatory variables
   - All relevant variables are included; irrelevant ones excluded

4. Sample size is large enough for valid inference-tests and CIs. (Recall large-sample properties of MLEs.)
   - Check for outliers.

What is the SATURATED Model?

- Observed response proportion:

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

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

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

- Fits the model exactly with the data
- Most general model possible for the data.

Which Models are often compared?

Consider one explanatory variable, X with n unique levels for the outcome, $Y \sim (Bin(m, \pi))$

- Saturated (FULL) Model: as many parameter coefficients as n

$$logit(\widehat{\pi}) = \widehat{\alpha}_0 + \widehat{\alpha}_1 \mathbb{1}_1 + \cdots + \widehat{\alpha}_{n-1} \mathbb{1}_{n-1}$$

- Fitted (REDUCED) Model: nested within a FULL model; has $(p+1)$ parameters

$$logit(\widehat{\pi}) = \widehat{\beta}_0 + \widehat{\beta}_1 X$$

- NULL Model: Intercept only model

$$logit(\widehat{\pi}) = \widehat{\gamma}_0$$

$n-(p+1)$

$p+1-1=p$

Checking model adequacy: Form of the model

Deviance Goodness -Of -Fit (G-O-F) Test

- To check model adequacy in binomial logistic regression, we can use the Deviance Goodness -Of -Fit (G-O-F) Test.
- Analogous to GOF test for comparing 2 models in Linear Regression.

- Form of hypotheses: H_0: REDUCED model, H_a: FULL model
- The DEVIANCE GOF test compares the fitted model (M) to the saturated model (S).

$$H_0 : (Fitted) logit(\widehat{\pi}) = \widehat{\beta}_0 + \widehat{\beta}_1 X \qquad p+1$$

$$H_a : (Saturated) logit(\widehat{\pi}) = \widehat{\alpha}_0 + \widehat{\alpha}_1 \mathbb{1}_1 + \cdots + \widehat{\alpha}_{n-1} \mathbb{1}_{n-1} \qquad n$$

Compared to Saturated model: Deviance G-O-F test

- Uses LRT
- Sometimes called "Drop-in-Deviance" test
- as extra-sum-of-squares tests; based on the deviance residual
- Hypotheses:

  H_0: $logit(\pi) = \alpha_0 + \alpha_1 X$

  (Fitted model fits data as well as Saturated model)

  H_a: $logit(\pi) = \beta_0 + \beta_1 \mathbb{1}_1 + \cdots + \beta_{n-1} \mathbb{1}_{n-1}$

  (Saturated model is better)

- Test Statistic:

$$Deviance = -2 \log \left(\frac{\mathcal{L}_R}{\mathcal{L}_F} \right) = -2 \log \left(\frac{\mathcal{L}_M}{\mathcal{L}_S} \right)$$

- Under H_0, $Deviance \sim$ Chi-square distribution with $\boxed{n - (p+1)}$ df.
- Warning: This is an asymptotic approximation, so it works better if each $m_i > 5$.)

Calculating the Deviance test statistic

Recall underlying model of Y: $Y_i \sim Binomial(m_i, \pi_i)$

$$P(Y_i = y_i) = \binom{m_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i}, \quad y_i = 0, 1, \ldots, m_i$$

Hence the likelihood is:

$$\mathcal{L} = \Pi_{i=1}^{n} \binom{m_i}{y_i} \pi_i^{y_i} (1 - \pi_i)^{m_i - y_i}$$

where

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{ip})} = \frac{e^{\mu}}{1+e^{\mu}}$$

$$\hat{\pi}_i = \frac{e^{\hat{\mu}}}{1+e^{\hat{\mu}}}$$

Calculating the Deviance test statistic

Then the log-likelihood is:

$$\log \mathcal{L} = \sum_{i=1}^{n} [y_i \log(\pi_i) + (m_i - y_i)\log(1-\pi_i) + \log \binom{m_i}{y_i}]$$

The deviance test statistic is based on a ratio of likelihoods.

$$\begin{aligned} Deviance &= -2\log\frac{\mathcal{L}_M}{\mathcal{L}_S} \\ &= -2(\log \mathcal{L}_M - \log \mathcal{L}_S) \\ &= 2(\log \mathcal{L}_S - \log \mathcal{L}_M) \end{aligned}$$

M = S

- Q: A Saturated Model has *Deviance* = 0 $= 2(\log L_S - \log L_S)$

Calculating the Deviance test statistic

$$\frac{y_i}{m_i} = \pi_i \qquad \frac{\hat{y}_i}{m_i} = \hat{\pi}_i$$

$$
\begin{aligned}
Deviance &= 2(\log \mathcal{L}_S - \log \mathcal{L}_M) \\
&= 2\sum_{i=1}^{n} \Bigg[\underbrace{y_i \log\left(\frac{y_i}{m_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i}\right) + \cancel{\log\binom{m_i}{y_i}}}_{\text{Sat.}} \\
&\qquad \underbrace{- y_i \log\left(\frac{\hat{y}_i}{m_i}\right) - (m_i - y_i)\log\left(\frac{m_i - \hat{y}_i}{m_i}\right) - \cancel{\log\binom{m_i}{y_i}}}_{\text{Fitted M.}}\Bigg] \\
&= 2\sum_{i=1}^{n} \Big[\underbrace{y_i \log(y_i) + (m_i - y_i)\log(m_i - y_i)}_{S} \\
&\qquad \underbrace{- y_i \log(\hat{y}_i) - (m_i - y_i)\log(m_i - \hat{y}_i)}_{M.}\Big] \\
G^2 &= 2\sum_{i=1}^{n} \left[y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - \hat{y}_i}\right)\right]
\end{aligned}
$$

Case Study IV Exercise: Using Deviance

0.2

Using R output,

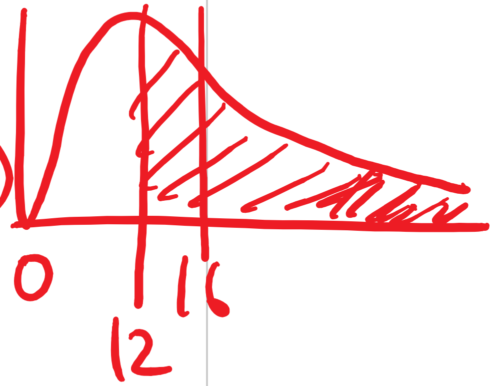
Q: Determine whether a saturated model is an improvement over the simpler model with linear function of log(*Area*).

(In R, we get deviance of a model by using `deviance('fittedmodel')`)

- Hypotheses: H_0: Fitted Model : $\text{logit}(\pi) \sim \log(\text{Area})$ $\quad p+1$

  H_a: Saturated $\quad n$
- Test Statistic: Deviance=12.062 $\quad$ In R: Residual deviance
- Distribution of TS: χ^2 (with $n-(p+1) = 18-(1+1) = 16$) df
- P-value: $P(\chi^2_{16} \geq 12.062) = 0.74$

  In R: $1 - \text{pchisq}(12.062, 16)$



- Conclusion: The data are consistent with H_0; the simpler model with linear function of log(*Area*) is adequate (fits as well as the saturated model).

Binomial Logistic Regression: Interpreting Deviance

- Smaller deviance leads to larger p-value and vice versa.
- Large p-values means:
  - Fitted model is adequate, OR
  - Test is not powerful enough to detect inadequacies
- Small p-values means:
  - Fitted model is not adequate; consider a more complex model with more explanatory variables or higher order terms and so on, OR
  - Response distribution is not adequately modelled by the Binomial distribution, OR
  - There are severe outliers.

Can we do a Deviance GOF test in Binary case?

In Binary logistic regression case, $m_i = 1$ for all i, and $y_i = \begin{cases} 0 \\ 1 \end{cases}$

Then deviance becomes:

$$\begin{aligned} Deviance &= 2 \sum_{i=1}^{n} \big[y_i \log(y_i) + (1 - y_i) \log(1 - y_i) \\ &\qquad - y_i \log(\widehat{y}_i) - (1 - y_i) \log(1 - \widehat{y}_i) \big] \\ &= 2 \sum_{i=1}^{n} \big[- y_i \log(\widehat{y}_i) - (1 - y_i) \log(1 - \widehat{y}_i) \big]. \end{aligned}$$

Notice that the terms that came from the saturated model, $\log \mathcal{L}_S$ are gone, so deviance is no longer useful to compare $\mathcal{L}_M$ with $\mathcal{L}_S$.

Model assessment in Binomial Logistic Regression

- Is linear relationship appropriate?
  - Plot observed logit versus quantitative explanatory variable
- Is the form of the model correct?
  - Use Wald or LRT tests
- Is saturated model better than fitted model?
  - Deviance GOF test
- Are there outliers?
  - Examine standardized residuals: Pearson and Deviance Residuals

$\bar{\pi}_{s,i} - \hat{\pi}_{M,i}$

- Consider other model fit statistics: AIC, BIC
- Other issues/concerns in model fitting

Residuals: Pearson and Deviance

(0,1) (0,1)

- Response (raw) residuals: (*observed* − *fitted*) proportion

(−1,1)

$$\widehat{\pi}_{S,i} - \widehat{\pi}_{M,i} = \frac{y_i}{m_i} - \widehat{\pi}_{M,i}$$

$y_i - m_i\hat{\pi}_{M,i}$

- Standardized residuals:
  - (1) Pearson Residuals: uses estimate of s.d. of Y (in denominator)

(−∞,∞)

$$P_{res,i} = \frac{y_i - m_i\widehat{\pi}_{M,i}}{\sqrt{m_i\widehat{\pi}_{M,i}(1-\widehat{\pi}_{M,i})}}$$

  - (2) Deviance Residuals: defined so that the sum of the squares of the residuals is the deviance

(−∞,∞)

−, +

$$D_{res,i} = \text{sign}(y_i - m_i\widehat{\pi}_{M,i}) \times \sqrt{2\left\{y_i \log\left(\frac{y_i}{m_i\widehat{\pi}_{M,i}}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - m_i\widehat{\pi}_{M,i}}\right)\right\}}$$

$G^2 = \sum_{i=1}^{n} D_i^2$

Response, Pearson and Deviance Residuals in R

- Response residuals

Model object

```
residuals(fitbl, type="response")
```

- Pearson residuals

```
residuals(fitbl, type="pearson")
```

- Deviance residuals

```
residuals(fitbl, type="deviance")
```

Case Study IV Example: Were there outliers in the data?

	Pearson, $P_{res,i}$	Deviance, $D_{res,i}$
Asymptotic Dist.	$\boxed{N(0,1)}$	$N(0,1)$
R code	`pearson`	`deviance`
Possible outlier if	$\lvert P_{res,i}\rvert > 2$	$\lvert D_{res,i}\rvert > 2$
Outlier if	$\lvert P_{res,i}\rvert > 3$	$\lvert D_{res,i}\rvert > 3$
Under small n	D_{res} closer to $N(0,1)$ than P_{res}	
$\hat{\pi}$ close to 0 or 1	P_{res} are unstable; related to instability of Wald	

Regularity conditions for MLE

- Results: Both are $< \lvert 2 \rvert$, so no outliers

Case IV

Other Model Fit Statistics

- Useful for comparing models with same response and same data
- Two popular fit statistics: AIC and BIC; combines log-likelihood with a penalty
  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
    - $N = \sum_{i=1}^{n} m_i$.
- Example: see AIC, BIC for Case IV model

In R: AIC( ), BIC( )

Problems and Solutions

Problems and Complications common to Linear and Logistic Regression

- *Extrapolation*- don't make inferences/predictions outside range of observed data; model may no longer be appropriate.
- *Multicollinearity*- highly correlated explanatory variables; difficult to assess individual effects on response. Consequences include:
  - Unstable fitted equation
  - Coefficient that should be statistically significant is not
  - Coefficient may have the wrong sign
  - Sometimes, large s.e. of $\widehat{\beta}$
  - Sometimes numerical procedure to find MLEs does not converge

Problems and Complications common to Linear and Logistic Regression

- *Extrapolation*- don't make inferences/predictions outside range of observed data; model may no longer be appropriate.
- *Multicollinearity*- highly correlated explanatory variables; difficult to assess individual effects on response. Consequences include:
  - Unstable fitted equation
  - Coefficient that should be statistically significant is not
  - Coefficient may have the wrong sign
  - Sometimes, large s.e. of $\widehat{\beta}$
  - Sometimes numerical procedure to find MLEs does not converge

Problems and Complications common to Linear and Logistic Regression

- *Influential points*- an observation is influential if its removal substantially changes estimated coefficients (such as, fitted $\widehat{\beta}$'s, deviance)
- *Model Building*- choosing explanatory variables and their forms (eg. polynomial terms, interaction and transformations) tend to overfit the data; should build model on training data and test on test data (cross validation).

Problems and Complications common to Linear and Logistic Regression

- *Influential points*- an observation is influential if its removal substantially changes estimated coefficients (such as, fitted $\widehat{\beta}$'s, deviance)
- *Model Building*- choosing explanatory variables and their forms (eg. polynomial terms, interaction and transformations) tend to overfit the data; should build model on training data and test on test data (cross validation).

Two problems specific to Logistic Regression

1. Extra-binomial variation

- variance of Y_i greater than $m_i\pi_i(1-\pi_i)$ $\psi = 1$
- also called "over dispersion"
- does not bias $\widehat{\beta}$'s but s.e. of $\widehat{\beta}$'s will be too small (too small p-values, too narrow CIs)

SOLUTION: add one more parameter to the model, ψ- dispersion parameter. Then $\text{Var}(Y_i) = \psi m_i\pi_i(1-\pi_i)$.

Two problems specific to logistic regression

2. Complete and Quasi-complete separation

- *Complete separation*:
  - one or a linear combination of explanatory variables perfectly predict whether $Y = 1$ or $Y = 0$
  - In Binary response, when $y_i = 1$, $\hat{y}_i = 1$, then $\sum_{i=1}^{n} \{y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)\} = 0$.
  - MLE's cannot be computed
- *Quasi-complete separation*:
  - explanatory variables predict $Y = 1$ or $Y = 0$ almost perfectly (just a few points wrong)
  - MLE's are numerically unstable

SOLUTION: simplify the model. Other options- penalized maximum likelihood, exact logistic regression, bayesian methods