Chapter 13. Logistic regression

1. The models have been investigated so far is of the form:

   $$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

   with assumption that ε_i follows iid $N(0, \sigma^2)$.

   This means response variable y must be a continuous and normal random variable.

2. In reality, many response variables are not continuous variable and do not follow normal distribution. Examples? What do we do if we still want to model these variables by standard linear regression models?

3. Generalized linear models (GLM)

   $$\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x}\boldsymbol{\beta}) + \boldsymbol{\varepsilon}$$

   In linear regression, the error $\boldsymbol{\varepsilon}$ follows a normal distribution and $\boldsymbol{g}$ is an identity function, i.e., $\boldsymbol{g}(\boldsymbol{x}\boldsymbol{\beta}) = \boldsymbol{x}\boldsymbol{\beta}$.

   In GLMs, $\boldsymbol{g}$ basically can be any function and the error $\boldsymbol{\varepsilon}$ will not normally follow the normal distribution.

4. Logistic regression is a special type of the GLMs. It deals with binary (dichotomous) variable. The log-linear regression introduced in the next chapter is another type of GLM which deals with count data.

5. For the logistic regression, the value of the binary variable y given x is

$$y = \pi(x) + \varepsilon$$

where the error ε assumes one of the two possible values:

$$\varepsilon = \begin{cases} 1-\pi(x) & \text{with } p = \pi(x) \quad \text{if } y = 1 \\ -\pi(x) & \text{with } q = 1-\pi(x) \quad \text{if } y = 0 \end{cases}$$

Therefore, $\varepsilon \sim [\pi(x)]^{\varepsilon+\pi(x)}[1-\pi(x)]^{1-\pi(x)-\varepsilon}$

6. All GLMs consist of three components:

(1). Random component – Identify the response variable *y* and assumes a probability distribution for it.

(2). Systematic component – Specify the explanatory variables used as regressors in the model.

(3). Link function – Describe the functional relationship between the systematic component and the expected value $\mu = E(y)$ of the random component.

7. Exponential family of distributions

$$f(y;\theta) = s(y)t(\theta)e^{a(y)b(\theta)}$$

or

$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)]$$

If $a(y) = y$, the pdf is said to be in the *canonical form*, and $b(\theta)$ is called the *natural parameter*. Natural parameter is typically used as a link function.

If there are other parameters in addition to θ (the parameter of interest), they are regarded as *nuisance parameters*.

8. Many pdfs belong to the exponential family: Normal, Poisson, Binomial, Geometric, NBD, Gamma dist. etc.

Example 1 – Normal distribution:

$$f(y;\mu) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

$$f(y;\mu) = \exp\left(\frac{y}{\sigma^2}\mu - \frac{\mu^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma) - \frac{y^2}{2\sigma^2}\right)$$

σ^2 is a nuisance parameter which is assumed to be known.

Example 2 – Poisson distribution:

$$f(y;\mu) = \frac{\lambda^y e^{-\lambda}}{y!}$$

$$f(y;\mu) = \exp\left(y\log\lambda - \lambda - \log y!\right)$$

Example 3 – Binomial distribution:

$$f(y;\mu) = \binom{n}{y}\pi^y(1-\pi)^{n-y}$$

$$f(y;\mu) = \exp\left(y\log\frac{\pi}{1-\pi} + n\log(1-\pi) + \log\binom{n}{y}\right)$$

9. As an example, let's model the distribution of species *Ocotea whitei* (ocotwh) in BCI using explanatory variable *slope*:

(1). Data – The 50 ha plot is divided into a 20×20 m grid system

(2). Random component – Identify the probability model for the occurrence (y) of the species in a particular cell. Obviously, y is a Bernoulli trial:

$$f(y;\pi) = \pi^y(1-\pi)^{1-y}$$

$$f(y;\pi) = \exp\left(y\log\frac{\pi}{1-\pi} + \log(1-\pi)\right)$$

(3). Systematic component – Specify the slope x (explanatory variable) as

$$\beta_0 + \beta_1 x$$

(4). Define a link function – In GLM, it is most typical to use the natural parameter as a link function

$$b(\theta) = \beta_0 + \beta_1 x$$

In our case, the natural parameter is (called *logit*):

$$b(\theta) = g(x) = \log \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x$$

or expressed as

$$\pi(y = 1 \mid x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

(Note: This can be viewed as the conditional probability that y equals 1 given an x. It follows that $1 - \pi(y = 0 \mid x)$ is the conditional probability that y equals 0 given an x.)

10. Maximized likelihood estimators (MLE)

For normal data, MLEs and LSEs are the same. However, for GLMs maximum likelihood method is the best method for parameter estimation.

Given a set of observations:

i	***y***	***x***
1	0	x_1
2	1	x_2
3	1	x_3
…	…	…
n	0	x_n

Assume y follows distribution:

$$f(y;\pi) = \pi^{y}(1-\pi)^{1-y}$$

The likelihood function is the probability that a set of data is observed, defined as

$$L(\pi; y) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1-\pi(x_i)]^{1-y_i}$$

or

$$L(\beta_0, \beta_1; y) = \prod_{i=1}^{n}\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right)^{y_i}\left(\frac{1}{1+e^{\beta_0+\beta_1 x_i}}\right)^{1-y_i}$$

(Note: Likelihood function is a function of parameters, not data. Data here are considered as random variables.)

11. Maximum likelihood principle: *The best explanation of a set of data is provided by the values of* (β_0, β_1) *that maximize the likelihood function.*

12. Log-likelihood function

$$l(\beta_0, \beta_1; y) = \sum\left(y_i \log\left(\frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}\right) + (1-y_i)\log\left(\frac{1}{1+e^{\beta_0+\beta_1 x_i}}\right)\right)$$

$$l(\beta_0, \beta_1; y) = \sum\left[y_i(\beta_0+\beta_1 x_i) - \log(1+e^{\beta_0+\beta_1 x_i})\right]$$

13. R-code:

ocotwh.glm=glm(y~slope,family=binomial(link=logit),data=ocotwh.dat)

> **summary(ocotwh.glm)** #view the outputs

Call:
glm(formula = y ~ slope, family = binomial(link = logit), data = ocotwh.dat)

Deviance Residuals:

```
    Min      1Q  Median      3Q     Max
-2.1134  -0.5213  -0.4005  -0.3117   2.4726

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.33010    0.16935  -19.66   <2e-16 ***
slope        0.30488    0.02174   14.03   <2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1196.49  on 1249  degrees of freedom
Residual deviance:  948.64  on 1248  degrees of freedom
AIC: 952.64

Number of Fisher Scoring iterations: 4
```

Note: 1. The test on *Intercept* and *slope* are called Wald statistic (see 16).

2. *Null deviance* is the deviance for the model including intercept term only, whereas the *Residual deviance* is for the model including intercept and slope terms (see 14). Large residual deviance suggests the model is a poor model.

14. Deviance (likelihood ratio statistic):

$$D = -2\log\left[\frac{\text{likelihood of the current model}}{\text{likelihood of the saturated model}}\right]$$

Note:

(1). The current model is the model of interest.

(2). The saturated model is the full model that considers observed data as parameters, thus there are as many parameters as data points (the full model gives a perfect

fit to the data). Under this model, the maximum of the likelihood is achieved as much as we can.

(3). If the current model is a good model, the ratio in the bracket will be close to 1. Otherwise, the ratio will be small.

(4). Therefore, large *D* suggests the current model is a poor description of the data.

(5). The deviance for logistic regression plays the same role as the residual sum of squares in linear regression.

15. Test for the significance of coef. H_0: $\beta_1 = 0$

Scaled deviance (similar to the extra-sum-of squares):

G = D(for the model without β_1) – D(for the model with β_1)

$$G = -2\log\left[\frac{L(\beta_0)}{L(\beta_0, \beta_1)}\right]$$

$G \sim \chi_1^2$, large G suggests rejecting H_0: $\beta_1 = 0$. p-value = $P(\chi_1^2 > G)$.

Note: G measures the contribution of the explanatory variable x to the model given β_0 already in the model. This scaled deviance (G) can be read from the outputs of **anova(object.glm)**.

> **anova(ocotwh.glm)** #view outputs, if test is desired, use option **test = "Chisq"**

Analysis of Deviance Table

Model: binomial, link: logit

Response: y

Terms added sequentially (first to last)

	Df	Deviance	Resid. Df	Resid. Dev
NULL			1249	1196.49
slope	1	247.85	1248	948.64

16. Wald statistic

This is another way to test for the significance of coefs. H_0: $\beta_1 = 0$

$$W = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}$$

This test statistic follows the standard normal distribution $N(0, 1)$.

p-value = $P(|z| > W)$, large W suggests rejection of H_0: $\beta_1 = 0$.

Note: Different from deviance test (χ^2 test in 15), the Wald statistic suggests a direction of change.

17. Analysis of residuals – assessing the goodness-of-fit

(1) Like the linear regression, **R** also performs all sort of residual analysis for GLMs. Try **plot(glm.object)**. However, interpretation of any such analysis must be cautious because GLM residuals are usually not normally distributed, particularly for small samples. As a result, the curvature in Q-Q plot cannot be taken seriously.

18. Two residuals

(1). The Pearson residuals:

$$r_{p_i} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

where $\hat{\mu}_i$ is the fitted value for i^{th} observation.

$\chi^2 = \sum r_{p_i}^2$ is a generalized Pearson χ^2 statistic

Problem with r_p is that its distribution is often skewed for non-normal distribution, therefore, Q-Q plot is not reliable. Large $|r_{pi}|$ suggests outliers.

(2). Deviance residuals

$$r_{D_i} = sign(y_i - \hat{\mu}_i)\sqrt{d_i}$$

where d_i is the contribution of each observation to the total deviance D, i.e., $D = \sum d_i$.

19. Model interpretations (Hosemer & Lemeshow 1989)

(1). The key question is: *What do the estimated coefs in the model tell us about the research questions that motivated the study?*

(2). The estimated coefs for the independent variables represent the slope or rate of change of the dependent variable given per unit of change in the independent variable. Therefore, the interpretation involves two steps: Determining the functional relationship between the dependent variable and the independent variable, and appropriately defining the unit of change for the independent variable.

(3). Proper interpretation of the coefs in a logistic regression model depends on being able to place meaning on the difference between two logits.

$$\pi(y = 1 \mid x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

logit:

$$g(x) = \log \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x$$

The difference of two logits:

$$\beta_1 = g(x+1) - g(x)$$

20. Odds ratio

Assume x is dichotomous independent variable, from the logistic model in (19) we have

		Independent variable x	
		$x = 1$	$x = 0$
Dependent variable y	$y = 1$	$\pi(1\mid 1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$	$\pi(1\mid 0) = \frac{e^{\beta_0}}{1+e^{\beta_0}}$
	$y = 0$	$\pi(0\mid 1) = \frac{1}{1+e^{\beta_0+\beta_1}}$	$\pi(0\mid 0) = \frac{1}{1+e^{\beta_0}}$

Odds of outcome being present among individuals with $x = 1$ is defined as:

$$\text{Odds:} \qquad \frac{\pi(1\mid 1)}{1-\pi(1\mid 1)} = \frac{\pi(1\mid 1)}{\pi(0\mid 1)} = e^{\beta_0+\beta_1}$$

Odds of outcome being present among individuals with $x = 0$ is:

$$\text{Odds:} \qquad \frac{\pi(1\mid 0)}{1-\pi(1\mid 0)} = \frac{\pi(1\mid 0)}{\pi(0\mid 0)} = e^{\beta_0}$$

Odds ratio:

$$\psi = \frac{\dfrac{\pi(1|1)}{1-\pi(1|1)}}{\dfrac{\pi(1|0)}{1-\pi(1|0)}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Odds ratio is a measure of association which has wide applications. It approximates how much more likely (or unlikely) it is for the outcome to be present among those with $x = 1$ than among those with $x = 0$.

For example, if y denotes the presence or absence of lung cancer and if x denotes whether or not the person is a smoker, then $\hat{\psi} = 2$ indicates that lung cancer occurs twice as often among smokers than among nonsmokers in the study population.

(Note: The interpretation for the odds ratio is based on the fact that in many instances it approximates a quantity called the *relative risk*, defined as $\dfrac{\pi(1|1)}{\pi(1|0)}$)

21. *log-odds ratio*:

Take a log on the odds ratio, it is easy to show

$$\log(\psi) = \log \frac{\dfrac{\pi(1)}{1-\pi(1)}}{\dfrac{\pi(0)}{1-\pi(0)}} = g(1) - g(0) = \beta_1$$

Thus, log-odds ratio is actually the difference in two logits which is the coef.

22. *probit* model:

$$\Phi^{-1}(\pi) = \beta_0 + \beta_1 x$$

Or

$$\pi = \Phi(\beta_0 + \beta_1 x)$$

ocotwh.probit=glm(y~slope,family=binomial(link=probit),data=ocotwh.dat)

23. *log-log* model:

$$\log[-\log(1-\pi)] = \beta_0 + \beta_1 x$$

Or

$$\pi = 1 - \exp[-\exp(\beta_0 + \beta_1 x)]$$

ocotwh.loglog=glm(y~slope,family=binomial(link=cloglog),data=ocotwh.dat)

24. Variable selection:

(1). Start from model with constant term only:

ocotwh.1=glm(y~1,family=binomial,data=ocotwh.dat)

(2). Use **step**

> **ocotwh.step=step(ocotwh.1,~slope+meanelev+convex+habcat)**

```
Start:  AIC= 1198.49
 y ~ 1

          Df Deviance     AIC
+ slope     1   948.64  952.64
+ habcat    6   958.06  972.06
+ convex    1  1161.53 1165.53
+ meanelev  1  1177.46 1181.46
<none>         1196.49 1198.49

Step:  AIC= 952.64
 y ~ slope

          Df Deviance     AIC
+ habcat    6   926.40  942.40
+ convex    1   944.67  950.67
<none>          948.64  952.64
+ meanelev  1   947.19  953.19
- slope     1  1196.49 1198.49

Step:  AIC= 942.4
 y ~ slope + habcat

          Df Deviance    AIC
```

```
+ meanelev  1   919.18 937.18
+ convex    1   921.90 939.90
<none>          926.40 942.40
- habcat    6   948.64 952.64
- slope     1   958.06 972.06

Step:  AIC= 937.18
 y ~ slope + habcat + meanelev

         Df Deviance    AIC
+ convex    1   904.45 924.45
<none>          919.18 937.18
- meanelev  1   926.40 942.40
- habcat    6   947.19 953.19
- slope     1   954.05 970.05

Step:  AIC= 924.45
 y ~ slope + habcat + meanelev + convex

         Df Deviance    AIC
<none>          904.45 924.45
- convex    1   919.18 937.18
- meanelev  1   921.90 939.90
- habcat    6   939.42 947.42
- slope     1   936.66 954.66
```

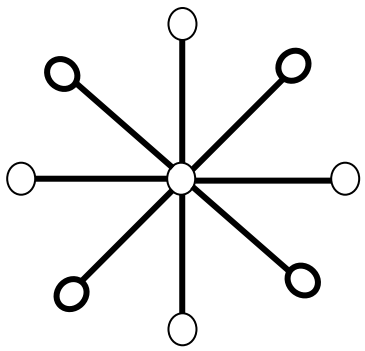
Note: Choose a model with smallest AIC. But care must be taken to accept such a model as the step may include trivial explanatory variables in some circumstances. In such situation, use **anova** is useful to screen out those trivial variables.

(3). >**summary(ocotwh.step)** #view the outputs for the best model

(4). >**anova(ocotwh.step)** #deviance analysis for the best model

25. Autologistic regression model:

$$\pi(y=1\mid x)=\frac{e^{\beta_0+\beta_1 x+\gamma_1 y_1{}^*+\gamma_2 y_2{}^*}}{1+e^{\beta_0+\beta_1 x+\gamma_1 y_1{}^*+\gamma_2 y_2{}^*}}$$



where

1^{st} order neighborhoods: $y_1 * = \sum (\text{vertical and horizontal neighbors})$

2^{nd} order neighborhoods: $y_2 * = \sum (\text{diagonal neighbors})$

26. Nominal logistic regression

If the response variable is categorical, with more than two categories, then there are two options for generalized linear models. One is the generalization of the logistic regression, extending the binary model to multicategorical models. The other option is the log-linear model – to model the frequencies or counts of the covariate patterns as the response variables with Poisson distribution. We will focus on the multicategorical logistic model here and defer the log-linear model to the next chapter.

27. Multicategorical (nominal) logistic regression

Nominal logistic regression models are used when there is no natural order among the response categories. Consider two species are distributed in 1250 cells on the 20×20 m grided BCI plot. Each cell can either be occupied by one of the species or empty. The distribution can be denoted as:

0 – Empty, none of the species is there,

1 – The presence of species A,

2 – The presence of species B.

The explanatory variables on each cell are denoted as x_i, where $i = 1, 2, \ldots, p$ are the number of explanatory variables.

$$\log(\frac{\pi_j}{\pi_1}) = \beta_{0j} + \beta_{1j} x_1 + \beta_{2j} x_2 + \ldots + \beta_{pj} x_p = \boldsymbol{\beta_j x}$$

where j is the j^{th} category (species).

Because $\pi_1 + \pi_2 + ... + \pi_J = 1$, we have

$$\pi_1 = \frac{1}{1 + \exp(\boldsymbol{\beta_2 x}) + \exp(\boldsymbol{\beta_3 x})}$$

$$\pi_2 = \frac{\exp(\boldsymbol{\beta_2 x})}{1 + \exp(\boldsymbol{\beta_2 x}) + \exp(\boldsymbol{\beta_3 x})}$$

$$\pi_3 = \frac{\exp(\boldsymbol{\beta_3 x})}{1 + \exp(\boldsymbol{\beta_2 x}) + \exp(\boldsymbol{\beta_3 x})}$$

The parameters of the nominal logistic model can be solved using mle for multinomial distribution.

28. Over-dispersion (extra-binomial distribution)

The logistic regression is based on the binomial distribution – modeling the count of yes/no: $f(y;\mu) = \binom{n}{y}\pi^y(1-\pi)^{n-y}$. This model assumes that the occurrences of yes/no are independent events. Its mean and variance are:

Mean: $E(y) = n\pi$

Variance: $V(y) = n\pi(1-\pi)$

The binomial distribution may not be strictly valid if in the following situations:

(1) Binary trials are not independent,

(2) The π's for the binary responses are not the same,

(3) Important explanatory variables are not included in the model of π.

In these situations, the mean for the binomial distribution still roughly holds, but the variance is inflated. In other words, the parameter estimates of the logistic model are still roughly unbiased, but the standard errors will tend to be smaller than they should be. This means that *p*-value will tend to be too small and confidence interval will tend to be two narrow. Which type of statistical error is committed here?

29. Logistic model with overdispersion

$E(y) = \mu[y = 1 \mid x] = n\pi$

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

$\text{var}[y \mid x] = \phi n\pi(1-\pi)$. The standard logistic regression has $\phi = 1$.

When you are not sure if your data are overdispersion, it is safer to suppose that overdispersion is present than to ignore it. The consequences of assuming the presence of overdispersion when there is actually none are minor. Inferences may be less precise, but they will not be misleading.

30. Checking for extra-binomial variation
    (1) Asking whether extra-binomial variation is likely present: are the binary responses independent? Are observations with identical values of the explanatory variables likely to have different π's? May any important explanatory variables be missing?
    (2) Examining the goodness-of-fit test: A large deviance test statistic may indicate that the binomial distribution with specified explanatory variables may not be supported by the data. Large deviance probably indicates overdispersion.
    (3) Examining residuals: Since deviance statistic is the sum of the squared deviance residuals, it is useful to examine the deviance residuals themselves to see whether a few of these are responsible for a large deviance statistic. If so, the analysis should focus on the outlier problem rather than on the overdispersion problem.

31. Solution: Quasi-likelihood approach

    The quasi-likelihood method was developed in such a way that the maximum quasi-likelihood estimates of parameters are identical to the mles, but the standard errors are larger and inferences are adjusted to account for the extra-binomial variation.

    One possible estimate of the dispersion parameter is the deviance statistic divided by its degree of freedom:

    $\hat{\phi} = \frac{\text{Deviance}}{\text{degree of freedom}}$, where df = *N*-*p*. (sample size - # of parameters)

This is amount to the sample variance of the deviance residuals. It should approximately equal 1 if the data are binomial, and larger than 1 for over-dispersion.

32. References:

(1) Dean, C.B. 1994 A modified pseudo-likelihood estimator of the over-dispersion parameter in Poisson mixture models, J. Appl. Stat. 21:523-532.

(2) Dean, C.B. 1992. Testing for overdispersion in Poisson and binomial regression models, J. Amer. Statist. Assoc. 87:451-457.

(3) Dean, C. and Lawless, J.F. 1989. Tests for detecting overdispersion in Poisson regression models, J. Amer. Statist. Asoc. 84:467-472.