STAT 120 C

Introduction to Probability and Statistics III

Dustin Pluta
2019/04/01

Odds, Odds Ratio, and Logistic Regression

- To better understand and interpret the statistical analysis of categorical data, we need to introduce the concept of the *odds* and the *odds ratio*

- **Definition**

The *odds* of an event A is

$$\textbf{Odds} = \frac{P(A)}{P(\text{not } A)} = \frac{P(A)}{1 - P(A)}$$

- **Example**

Suppose we roll a fair, 6-sided die. The odds of rolling a 1 is

$$\textbf{Odds} = \frac{P(\text{roll a } 1)}{P(\text{don't roll a } 1)}$$

$$Odds = \frac{1/6}{5/6} = \frac{1}{5}$$

Odds

- The *odds* is a measure of how likely an event is relative to a non-event
- When we say "3 to 1 odds", we're making a statement about the relative probabilty of events
- Since

$$P(A) = \frac{odds(A)}{odds(A) + 1},$$

we can convert the odds to a probability.

- For instance, "3 to 1" corresponds to an odds of 3, which implies $P(A) = \frac{3}{4}$.

Odds Ratio

- Suppose that X represents the event that an individual is exposed to a harmful factor for a disease, and that D represents the event that the individual develops the disease.
- A key goal of clinical trials and epidemiology is to determine the degree of risk associated with exposure.
- It is often not possible or practical to estimate the odds, especially for rare diseases
- In these cases, it is useful to consider the *odds ratio* instead.

Odds Ratio

Definition

The conditional odds of D given exposure X is

$$odds(D|X) = \frac{P(D|X)}{1 - P(D|X)}.$$

The conditional odds of D given that there is no exposure ($\bar X$) is

$$odds(D|\bar{X}) = \frac{P(D|\bar{X})}{1 - P(D|\bar{X})}.$$

The **odds ratio** is

$$\Delta = \frac{odds(D|X)}{odds(D|\bar{X})}$$

Odds Ratio

Contingency table perspective

	$\bar{D}$	D	
$\bar{X}$	π_{00}	π_{01}	$\pi_{0\cdot}$
X	π_{10}	π_{11}	$\pi_{1\cdot}$
	$\pi_{\cdot 0}$	$\pi_{\cdot 1}$	1

With this notation, the odds ratio can be written

$$\Delta = \frac{\pi_{11}/(\pi_{10}+\pi_{11})}{\pi_{00}/(\pi_{01}+\pi_{01})} = \frac{\pi_{11}\pi_{00}}{\pi_{01}\pi_{10}}$$

Estimating the Odds Ratio

There are three common sampling designs that can be used to estimate the odds ratio.

Method 1: Simple random sample

- If we draw a simple random sample from the population, all of the probabilities in the contingency table can be estimated by $\frac{n_{ij}}{n_{..}}$.
- However, if the disease D is rare, then we will need a very large sample size to accurately estimate $P(D|X)$ and $P(D|\bar{X})$.
- This method is theoretically ideal, but often impractical for rare diseases and/or rare exposures.

Estimating the Odds Ratio

Method 2: Prospective Study

- In a prospective study, a fixed number of exposed and nonexposed individuals are sampled, and the incidence of disease recorded in each group.
- This allows us to make sure that we have a sufficient number of exposed and unexposed individuals
- From this sample, we can compute $P(D|X)$ and $P(D|\bar{X})$, and so can compute the odds ratio.
- However, we could still run into problems if the disease is rare (which would again require a large sample size).
- Note that in this design, we cannot estimate the individual cell probabilities π_{ij} since the number of exposed and unexposed individuals is fixed by the sampling design.

Estimating the Odds Ratio

Method 3: Retrospective Study

- In a retrospective study, the number of diseased and undiseased individuals are fixed by the sample design, and the exposure incidences are counted.
- In this setting, we can estimated $P(X|D)$ and $P(X|\bar{D})$, but cannot estimate $P(D|X)$ or $P(D|\bar{X})$, since the number of diseased and undiseased individuals are fixed.
- This seems problematic at first, but we can actually still recover an estimate of the odds ratio.
- Retrospective studies are generally the easiest means of estimating the odds ratio, and often it is the only practical method for studying rare diseases.

Estimating the Odds Ratio

Method 3: Retrospective Study

- Observe that:

$$P(X|D) = \frac{\pi_{11}}{\pi_{01} + \pi_{11}}$$
$$1 - P(X|D) = \frac{\pi_{01}}{\pi_{01} + \pi_{11}}$$
$$odds(X|D) = \frac{\pi_{11}}{\pi_{01}}.$$

Similarly,

$$odds(X|\bar{D}) = \frac{\pi_{10}}{\pi_{00}}.$$

Thus, the same odds ratio defined above can be expressed as

$$\Delta = \frac{odds(X|D)}{odds(X|\bar{D})}.$$

Estimating the Odds Ratio

Method 3: Retrospective Study

The probabilities in a retrospective study can be estimated as

$$\hat{P}(X|D) = \frac{n_{11}}{n_{\cdot 1}}$$
$$1 - \hat{P}(X|D) = \frac{n_{01}}{n_{\cdot 1}}$$
$$\hat{odds}(X|\bar{D}) = \frac{n_{11}}{n_{01}}$$
$$\hat{odds}(X|\bar{D}) = \frac{n_{10}}{n_{00}}$$

The resulting estimate of the odds ratio is

$$\hat{\Delta} = \frac{n_{00}n_{11}}{n_{01}n_{10}}$$

Estimating the Odds Ratio

- To do inference on $\hat{\Delta}$, we can generate a confidence interval using the asymptotic distribution of the $\log$ odds ratio:

$$\frac{\log(\hat{\Delta}) - \log(\Delta)}{se(\log(\hat{\Delta}))} \dot{\sim} \mathcal{N}(0,1),$$

where $se(\log \hat{\Delta}) = \sqrt{1/n_{00} + 1/n_{10} + 1/n_{01} + 1/n_{11}}$.

The resulting $(1-\alpha)100\%$ confidence interval is

$$\exp\Big\{\log \hat{\Delta} \pm z_{\alpha/2} se(\log \hat{\Delta})\Big\}$$

Estimating the Odds Ratio

Example: Estimating Risk of Alzheimer's Disease by APOE4 Exposure

	AD Yes	AD No	Total
APOE4 Yes	44	11	55
APOE4 No	6	39	45
Total	50	50	100

Table 5: Retrospective sample of Alzheimer's patients and healthy controls.

$$\hat{\Delta} = \frac{n_{00}n_{11}}{n_{01}n_{10}} = \frac{44 \cdot 39}{6 \cdot 11} = 26$$

$$se(\log(\hat{\Delta})) = 0.55$$

- 95% CI for odds ratio:

$$\exp\left\{\log \hat{\Delta} \pm z_{\alpha/2} se(\log \hat{\Delta})\right\} = (8.85, 76.4)$$

- **Interpretation:** the odds of developing AD given the presence of the APOE4 gene is estimated to be 26 times the odds of developing AD given the absence of the APOE4

Estimating the Odds Ratio

Example: Estimating Risk of Alzheimer's Disease by APOE4 Exposure

NIH Public Access
Author Manuscript
Nat Rev Neurol. Author manuscript; available in PMC 2013 July 29.

Published in final edited form as:
Nat Rev Neurol. 2013 February ; 9(2): 106–118. doi:10.1038/nrneurol.2012.263.

Apolipoprotein E and Alzheimer disease: risk, mechanisms, and therapy

Chia-Chen Liu[1], **Takahisa Kanekiyo**[2], **Huaxi Xu**[1], and **Guojun Bu**[1]

[1]Fujian Provincial Key Laboratory of Neurodegenerative Disease and Aging Research, Institute of Neuroscience, College of Medicine, Xiamen University, Xiamen, Fujian 361005, China

[2]Department of Neuroscience, Mayo Clinic, 4500 San Pablo Road, Jacksonville, Florida 32224, USA

APOE genotypes, AD and cognition

APOE ε4 as a strong risk factor for AD

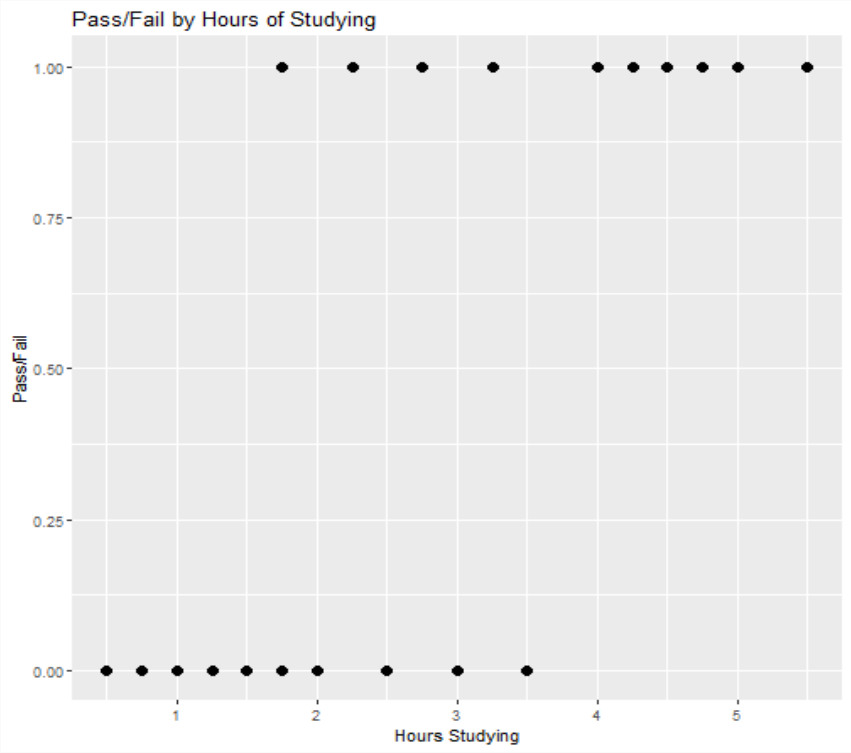
Genome-wide association studies have confirmed that the ε4 allele of *APOE* is the strongest genetic risk factor for AD.[16, 17] The presence of this allele is associated with increased risk for both early-onset AD and LOAD.[18, 19] A meta-analysis of clinical and autopsy-based studies demonstrated that, compared with individuals with an ε3/ε3 genotype, risk of AD was increased in individuals with one copy of the ε4 allele (ε2/ε4, OR 2.6; ε3/ε4, OR 3.2) or two copies (ε4/ε4, OR 14.9) among Caucasian subjects.[10] The ε2 allele of *APOE* has protective effects against AD: the risk of AD in individuals carrying *APOE* ε2/ε2 (OR 0.6) or ε2/ε3 (OR 0.6) are lower than those of ε3/ε3.[10] In population-based studies, the *APOE4*–AD association was weaker among African Americans (ε4/ε4, OR 5.7) and Hispanics (ε4/ε4, OR 2.2) and was stronger in Japanese people (ε4/ε4, OR 33.1) compared with Caucasian cases (ε4/ε4, OR 12.5).[10] *APOE* ε4 is associated with increased prevalence of AD and lower age of onset.[7, 10, 20] The frequency of AD and mean age at clinical onset are 91% and 68 years of age in ε4 homozygotes, 47% and 76 years of age in ε4 heterozygotes, and 20% and 84 years in ε4 noncarriers,[7, 20] indicating that *APOE* ε4 confers dramatically increased risk of development of AD with an earlier age of onset in a gene dose-dependent manner (Figure 1b).

Logistic Regression

- Consider binary response variables Y_i, with $Y_i = 1$ with probability π_i and $Y_i = 0$ with probability $1 - \pi_i$.
- That is, each Y_i is a Bernoulli random variable with probability π_i.
- In Fisher's exact test and Pearson's chi-squared test, we are modeling the association of categorical responses with categorical covariates.
- Logistic regression allows us to extend this idea for the modelling of categorical responses with *continuous* covariates.

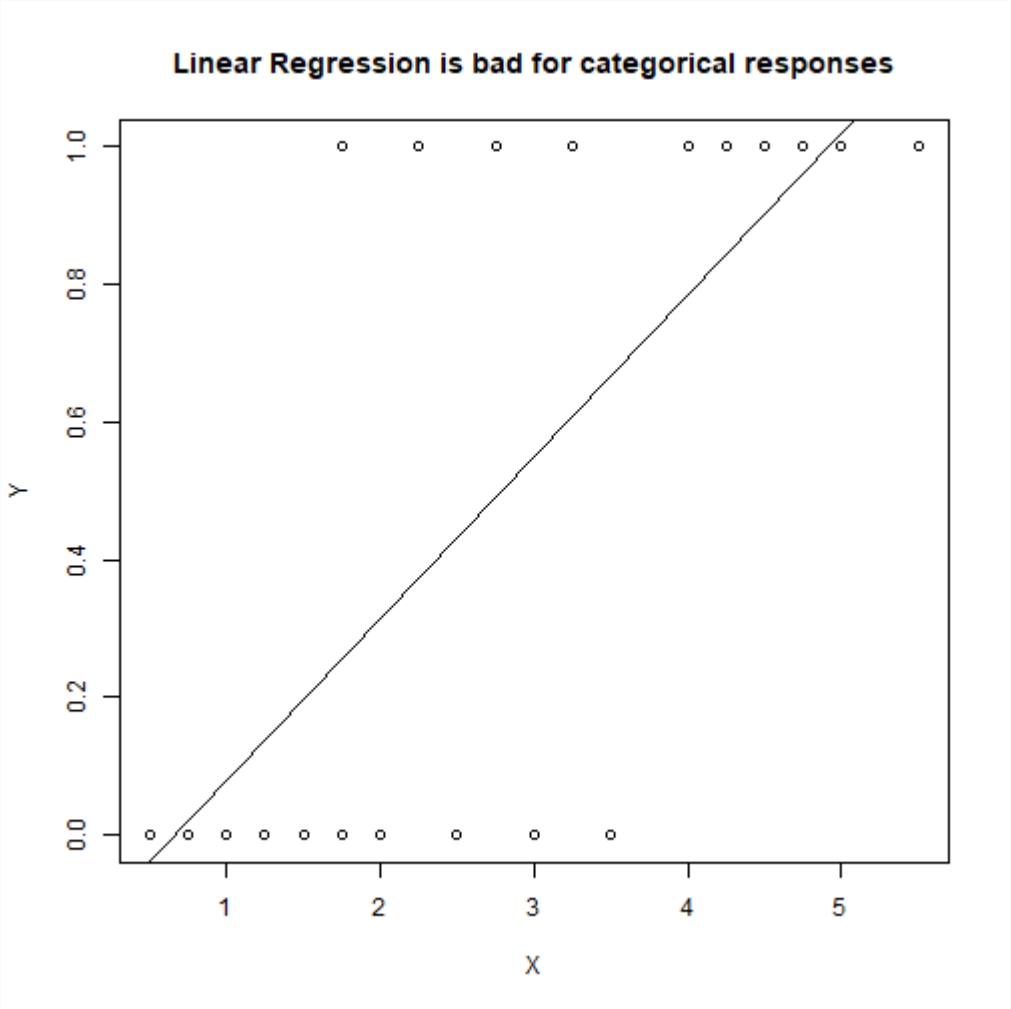
Logistic Regression

- Consider the following example (taken from Wikipedia article on logistic regression).
- Let Y_i be the Pass/Fail (Pass = 1, Fail = 0) status of student i, and let X_i be the number of hours the student studied the weekend before the final exam.
- We wish to determine how the probability of passing is related to the number of hours spent studying.



Logistic Regression

If we pretend that the data are continuous (coded as 0 or 1), and fit a usual linear regression, we may run into problems.



Logistic Regression

- In logistic regression, we model the *log odds* as a linear combination of the predictors.
- If $Y_i \overset{ind}{\sim} Bernoulli(\pi_i)$, the log odds is

$$\log\left(\frac{\pi_i}{1-\pi_i}\right)$$

- For measured covariates X_i, the logistic regression model is

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i.$$

Logistic Regression

Model Properties

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_i.$$

- Note that the odds $\frac{\pi_i}{1-\pi_i} \in [0, \infty)$, and so $\log\left(\frac{\pi_i}{1-\pi_i}\right) \in \mathbb{R}$.
- Thus, it is reasonable to model this quantify as linear in X_i.
- We can solve the model equation for π_i to see that

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}.$$

Logistic Regression

Interpretation of Parameters

- β_0: log-odds when all predictors are 0
- β_1: difference in log-odds when comparing subpopulations that differ in X by 1 unit.
- β_1 can also be interpreted as the *odds ratio* when comparing two subpopulations that differ in X by 1 unit:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1(X+1)$$
$$\log\left(\frac{\pi_j}{1-\pi_j}\right) = \beta_0 + \beta_1 X$$
$$\beta_1 = \log\left(\frac{\pi_i/(1-\pi_i)}{\pi_j/(1-\pi_j)}\right)$$

Logistic Regression

Model Fitting

We can write the likelihood for the logistic regression model:

$$\begin{aligned}\mathcal{L}(\beta_0, \beta_1 | X_i, Y_i) &= \prod_i P(Y_i | X_i, \beta_0, \beta_1) \\ &= \prod_i \pi_i^{Y_i} (1 - \pi_i)^{1 - Y_i}\end{aligned}$$

where $\pi_i = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}$.

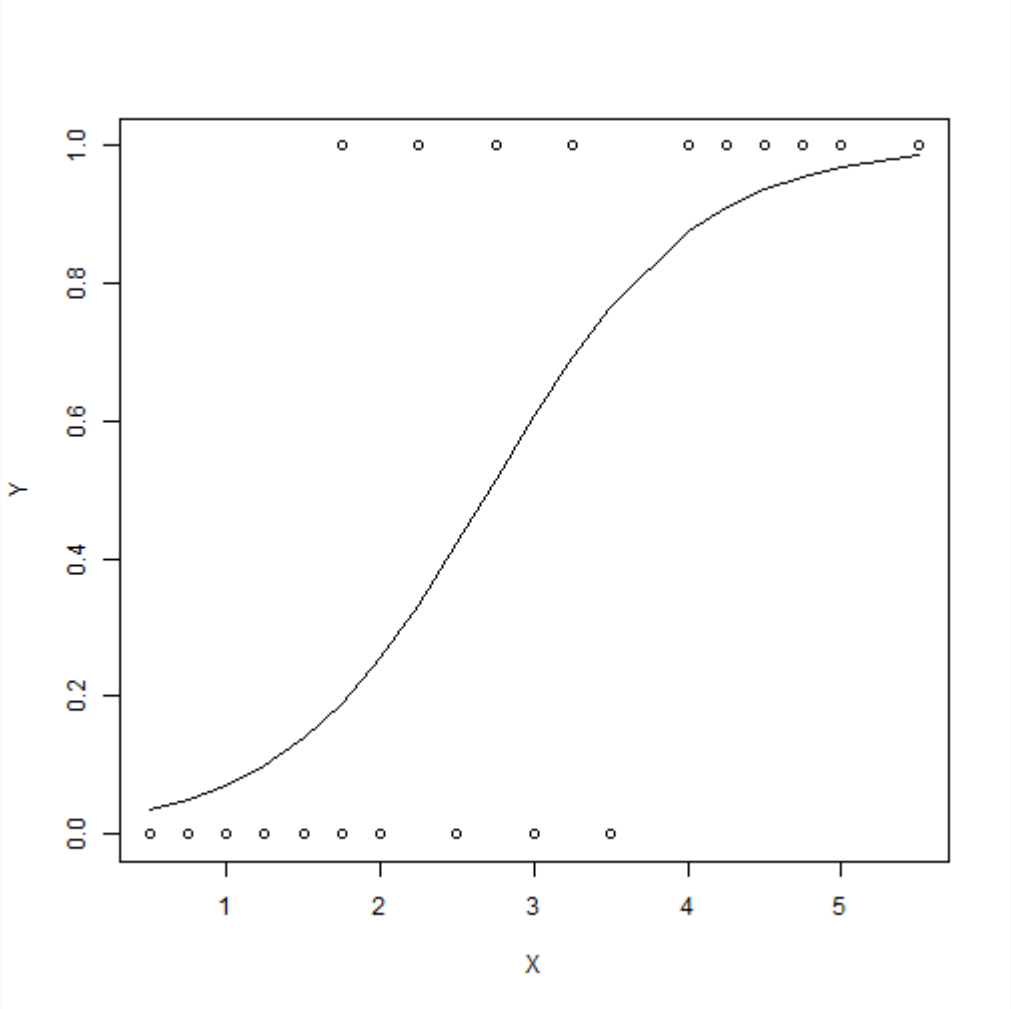
- No closed form expression for β_0, β_1 exists, but the likelihood can be maximized by numerical optimization methods (Newton's method, gradient ascent), or by the *Iteratively Reweighted Least Squares* method.

Logistic Regression

```r
fit ← glm(Y ~ X, family = "binomial")
summary(fit)
```

```
##
## Call:
## glm(formula = Y ~ X, family = "binomial")
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.70557  -0.57357  -0.04654   0.45470   1.82008
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -4.0777     1.7610  -2.316   0.0206 *
## X             1.5046     0.6287   2.393   0.0167 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 27.726  on 19  degrees of freedom
## Residual deviance: 16.060  on 18  degrees of freedom
## AIC: 20.06
##
```

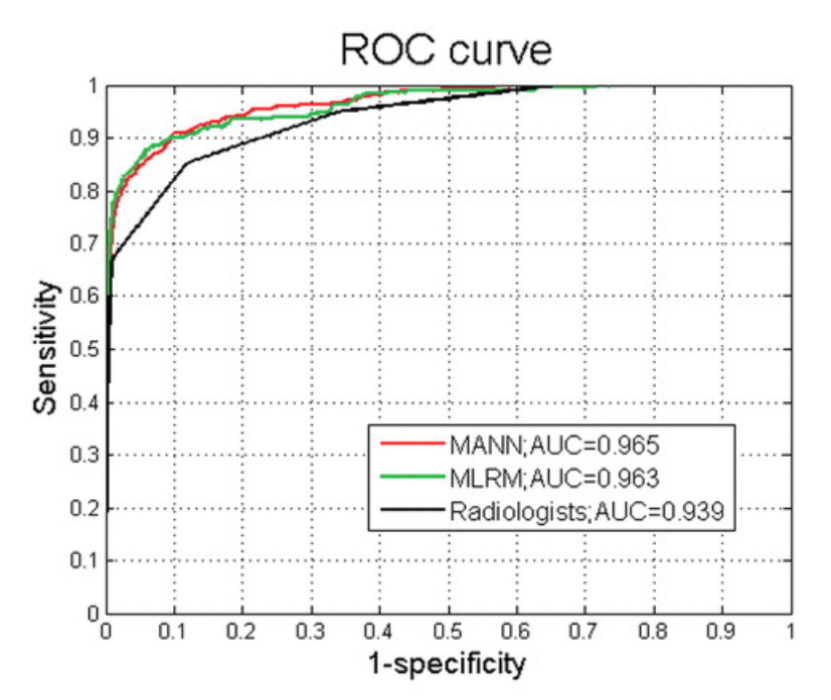
Logistic Regression



Logistic Regression

- Logistic regression is a simple model, but is often competitive with more advanced "state of the art" classifers (like neural networks)

https://pubs.rsna.org/doi/full/10.1148/rg.301095057



Logistic Regression

Much more information on logistic regression and many other statistical methods for machine learning is provided in the classic text *Elements of Statistical Learning*

https://web.stanford.edu/~hastie/ElemStatLearn/printings/ESLII_print12.pdf