Lecture 1 Handout

Elizabeth Colantuoni

3/22/2021

I. Objectives

Upon completion of this session, you will be able to do the following:

- Define the class of generalized linear models; show that the linear model and logistic model are included in this class
- Understand and explain the Bernoulli distribution
- Understand and explain the logistic regression model
- Interpret the regression coefficients from the logistic regression model
- Connect the results of simple logistic regression to estimates from standard analyses of 2x2 tables

II. Generalized linear models

A. Review of linear model

We have a vector of n outcomes y assumed to be realizations of independent random variables Y with the mean μ (a vector of length n) given as:

$$\mu_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi} + \epsilon_i$$

where $i = 1, 2, ..., n$ and $X_{1i}, \ldots, X_{pi}$ are covariates (exposure/confounder/mediators/moderators) representing the mean of Y. Defining the mean using this linear function of Xs defines the *systematic* part of the model.

The *random* part of the model assumes that ϵ_i are independent normally distributed random variables with mean 0 and variance σ^2.

In matrix notation we can specify the model as:

$$Y = X\beta + \epsilon, \epsilon \sim N(0, \Sigma), \Sigma = \sigma^2 I_{nxn}$$

An alternative way to write the model is that the components of Y are independent normal variables with constant variance, σ^2 and

$$E(Y) = \mu, \mu = X\beta$$

B. Generalization of the linear model

Generalized linear models are a class of models that extend the ideas from the linear model setting described above to additional types of outcomes/distributions.

The specification of a generalized linear model requires three components:

- 1. The *random* component: This specifies the distribution of the independent random variables comprising Y; e.g. Y consists of independent normally distributed variables, with $E(Y) = \mu$ and variance σ^2.
- 2. The *systematic* component: The covariates $X_1, ..., Xp$ produce a linear predictor given by:

$$g(\mu_i) = X_i^{\prime}\beta = \beta_0 + \beta_1 X_{1i} + ..., + \beta_p X_{pi}$$

where X_i is a column vector containing a 1 for the intercept and values of the covariates for observation i, i.e. $(1, X_{1i}, ..., X_{pi})$ is a $p + 1 \times 1$ vector.

- 3. The *link* function, linking the *random* and *systematic* components:

$$\mu = g^{-1}(\mu)$$

where in case of the linear model, $g(\mu) = \mu$ known as the *identity* link.

C. Types of data, distributions and link functions

Below is a table providing 4 data types with corresponding *random* components and *link* functions.

Type of data	Distribution defining *random* component	*Link* function	Regression Model name
Continuous	Gaussian/Normal	$g(\mu) = \mu$ (identity)	Linear
Positive continuous	Gamma	$g(\mu) = 1/\mu$ (inverse)	Gamma
Binary	Bernoulli	$g(\mu) = log(\mu/(1-\mu))$ (logit)	Logistic
Count (0, 1, 2, ...)	Poisson	$g(\mu) = log(\mu)$ (log)	Poisson or log-linear

D. Exponential family of distributions

The Gaussian/Normal, Gamma, Bernoulli and Poisson distributions are distributions within a special class of distributions called the "exponential family".

Generalized linear models define a class of regression models that can apply to any distribution within the "exponential family".

A distribution function in the "exponential family" can be expressed as:

$$f_Y(y|\theta, \phi) = exp\{[y\theta - b(\theta)]/a(\phi) + c(y, \phi)\}$$

where y is the data, θ represents the parameter of interest and ϕ is the dispersion parameter.

1. Example: Gaussian/Normal:

$$\begin{aligned} f_Y(y|\theta,\phi) &= \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-(y-\mu)^2/2\sigma^2\right\} \\ &= exp\left\{(y\mu - \mu^2/2)/\sigma^2 - 1/2(y^2/\sigma^2 + log(2\pi\sigma^2))\right\} \end{aligned}$$

where

- $\theta = \mu$
- $\phi = \sigma^2$
- $a(\phi) = \phi$
- $b(\theta) = \theta^2/2$
- $c(y,\phi) = 1/2(y^2/\sigma^2 + log(2\pi\sigma^2))$

2. Example: Bernoulli

$$\begin{aligned} f_Y(y|\theta,\phi) &= p^y(1-p)^{(1-y)} \\ &= exp\left\{ylog(p) + (1-y)log(1-p)\right\} \\ &= exp\left\{y[log(p) - log(1-p)] + log(1-p)\right\} \end{aligned}$$

where

- $\theta = log(p) - log(1-p) = log(\frac{p}{1-p}) \rightarrow p = \frac{exp(\theta)}{1+exp(\theta)}$
- $\phi = 1$
- $a(\phi) = 1$, $c(y,\phi) = 0$
- $b(\theta) = -log(1-p) = -log(\frac{1}{1+exp(\theta)})$

3. Properties of exponential families:

It turns out that $E(Y) = b^{'}(\theta)$ and $Var(Y) = b^{''}(\theta)a(\phi)$.

Biostat students should confirm the $E(Y)$ and $Var(Y)$ for both the Gaussian and Bernoulli distributions.

III. Logistic Regression Background

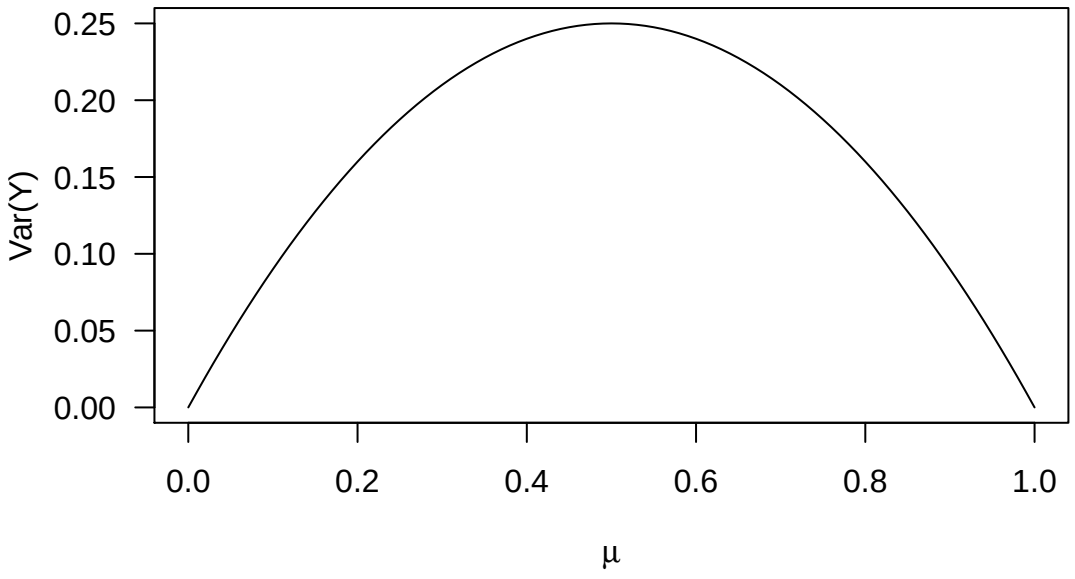
A. Background

You will need to familiarize yourself with the Bernoulli distribution. Assume $Y \sim Bernoulli(\mu)$. Then the $E(Y)$ and $Var(Y)$ are:

$$\begin{aligned} E(Y) &= \Sigma_{y=0}^{1} y Pr(Y=y) \\ &= 0 \times Pr(Y=0) + 1 \times Pr(Y=1) \\ &= 0 \times (1-\mu) + 1 \times \mu \\ &= \mu \end{aligned}$$

$$\begin{aligned} Var(Y) &= \Sigma_{y=0}^{1} (y-\mu)^2 Pr(Y=y) \\ &= (0-\mu)^2 \times Pr(Y=0) + (1-\mu)^2 \times Pr(Y=1) \\ &= \mu^2(1-\mu) + (1-\mu)^2\mu \\ &= \mu(1-\mu)[\mu + (1-\mu)] \\ &= \mu(1-\mu) \end{aligned}$$

NOTE: Unlike the Gaussian/Normal distribution, the $Var(Y) = f(\mu)$.



B. Inference for μ, no covariates

Suppose $Y_1, Y_2, ..., Y_n$ are independent Bernoulli(μ) and $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.

Then, $E(\bar{Y}) = \mu$ and $Var(\bar{Y}) = \frac{\mu(1-\mu)}{n}$ with $\hat{Var}(\bar{Y}) = \frac{\hat{\mu}(1-\hat{\mu})}{n}$.

A 95% confidence interval for μ is given by:

$$\hat{\mu} \pm 1.96\sqrt{\frac{\hat{\mu}(1-\hat{\mu})}{n}}$$

B. Adding covariates and motivation for logit link

Suppose now that we want to associate covariates with μ assuming $Y \sim Bernoulli(\mu)$.

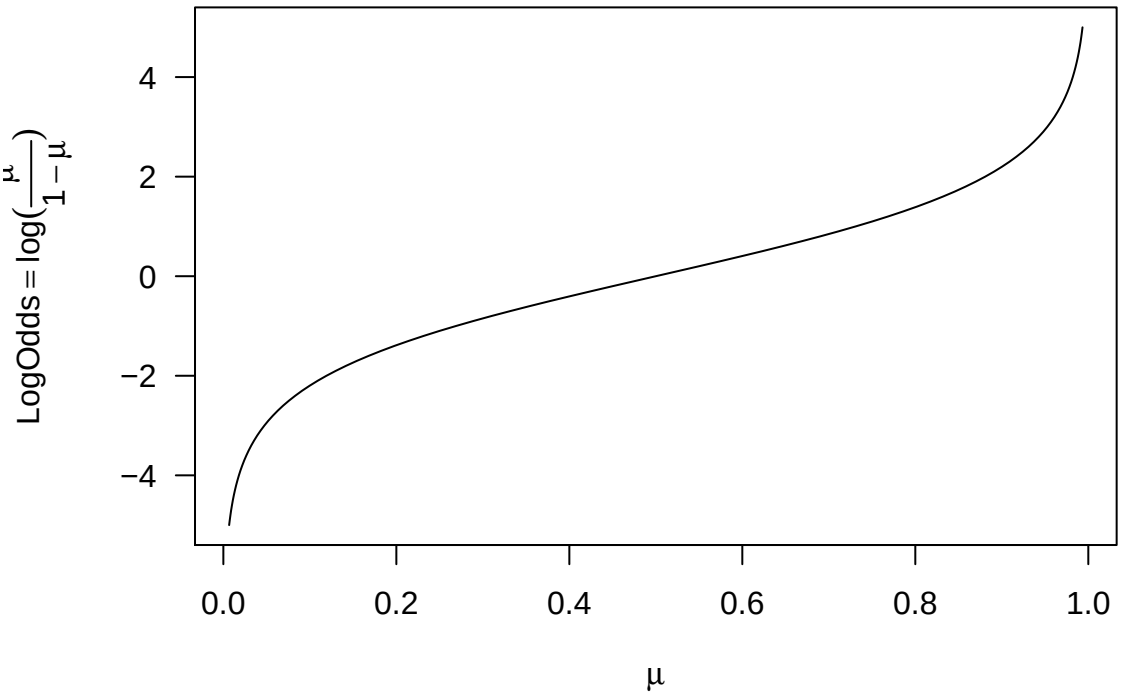
Then we run into a challenge applying the standard linear regression thinking; i.e. $\mu = \beta_0 + \beta_1 X$ because μ is bounded between 0 and 1.

The *link* function in a generalized linear model allows us to transform a mean, which may be bounded, to an unbounded scale. We construct the regression model on the unbounded scale and can then transform back.

In logistic regression we go from:

Probability	ODDS	Log ODDS
μ	$\frac{\mu}{1-\mu}$	$log(\frac{\mu}{1-\mu})$
1	∞	∞
0.95	$\frac{0.95}{0.05} = 19$	$log(19) = 2.94$
0.75	$\frac{0.75}{0.25} = 3$	$log(3) = 1.10$
0.5	$\frac{0.5}{0.5} = 1$	$log(1) = 0$
0.25	$\frac{0.25}{0.75} = 0.33$	$log(0.33) = -1.10$
0.05	$\frac{0.05}{0.95} = 0.05$	$log(0.05) = -2.94$

NOTE: The Log ODDS are unbounded; Log ODDS can take values from $-\infty$ to ∞.



You should get comfortable with going from μ to the log odds of μ and vice-versa!

Probability: μ	ODDS: o	Log ODDS: lo
$\mu = \frac{o}{1+o}$	$o = \frac{\mu}{1-\mu}$	$lo = log(o)$
$\mu = \frac{exp(lo)}{1+exp(lo)}$	$o = exp(lo)$	$lo = log(\frac{\mu}{1-\mu})$

You practice:

1. $logodds = 0$ $odds =$ $\mu =$
2. $logodds = 0.01$ $odds =$ $\mu =$
3. $logodds = 0.10$ $odds =$ $\mu =$

IV. Simple and multiple logistic regression

Assume we have n independent values of $Y_i = 0$ or $1 \sim Bernoulli(\mu_i)$, then we want to build a regression model for $E(Y_i|X_i) = Pr(Y_i = 1|X_i) = \mu_i$.

Logistic regression:

$$log\left\{\frac{\mu_i}{1-\mu_i}\right\} = X_i^{\prime}\beta = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi}$$

- If $p = 1$, then we have a "simple" logistic regression model
- If $p > 1$, then we have a "multiple" logistic regression model

B. Interpretation of model parameters via analysis of 2x2 tables

To motivate logistic regression analysis and work on the interpretion of parameters from the model, we will use data from the NMES survey (used in 140.653)

Let Y_i be the indicator for whether person i spends $\geq$ \$1000 on medical services; i.e. 1 if $totalexp \geq 1000$, 0 otherwise.

Define two covariates:

- $MSCD_i$ is the indicator for whether person i has ($MSCD_i = 1$) or doesn't have ($MSCD_i = 0$) a major smoking related disease (including lung cancer, COPD, heart disease, stroke, esophageal cancer, oropharyngeal cancer).
- $Older_i$ is an indicator for whether person i is greater than 65 years of age.

From the NMES data we have the following:

	MSCD		
	1	0	Total
Y = 1	986	3349	4335
Y = 0	333	7016	7349

B. Intercept only logistic regression model

From the table above, we can estimate the

- Probability of a big expenditure: $\hat{\mu}_i = \frac{4335}{4335+7349} = 0.37$
- Odds of a big expenditure: $\frac{\hat{\mu}_i}{1-\hat{\mu}_i} = \frac{0.37}{0.63} = 0.59$
- Log odds of a big expenditure: $log\left\{\frac{\hat{\mu}_i}{1-\hat{\mu}_i}\right\} = log\left\{\frac{0.37}{0.63}\right\} = log(0.59) = -0.53$

Now, consider obtaining similar information from a logistic regression model:

$$log\left\{\frac{\mu_i}{1-\mu_i}\right\} = \beta_0$$

where β_0 is the log odds of a big expenditure.

From the data provided, the estimated value of

- the log odds of a big expenditure:

$$\hat{\beta}_0 = log\left\{\frac{4335}{7349}\right\} = log(0.59) = -0.53$$

- the standard error for the log odds of a big expenditure:

$$\sqrt{\frac{1}{4335} + \frac{1}{7349}} = 0.019$$

- the odds of a big expendture:

$$exp(\hat{\beta}_0) = exp(-0.53) = 0.59$$

- the probability of a big expenditure:

$$\hat{\mu}_i = \frac{exp(\hat{\beta}_0)}{1 + exp(\hat{\beta}_0)} = 0.37$$

C. Simple Logistic Regression model

From the 2x2 table, we can estimate the odds ratio of a big expenditure comparing persons with a MSCD to those without a MSCD.

$$OR = \frac{Pr(Y_i = 1|MSCD_i = 1)}{Pr(Y_i = 0|MSCD_i = 1)} \div \frac{Pr(Y_i = 1|MSCD_i = 0)}{Pr(Y_i = 0|MSCD_i = 0)}$$

- Estimated odds ratio, $\hat{OR}$:

$$\frac{\frac{986}{333}}{\frac{3349}{7016}} = 6.203$$

- Estimated log odds ratio, $log(\hat{OR})$:

$$log(6.203) = 1.825$$

- Estimated standard error of the $log(\hat{OR})$:

$$\sqrt{\frac{1}{986} + \frac{1}{333} + \frac{1}{3349} + \frac{1}{7016}} = 0.067$$

Next, let's consider the simple logistic regression model:

$$log\left\{\frac{\mu_i}{1 - \mu_i}\right\} = \beta_0 + \beta_1 MSCD_i$$

When $MSCD_i = 0$ then:

$$log\left\{\frac{\mu_i}{1-\mu_i}\right\} = \beta_0$$

When $MSCD_i = 1$ then:

$$log\left\{\frac{\mu_i}{1-\mu_i}\right\} = \beta_0 + \beta_1$$

So that

- β_0 is the log odds of a big expenditure for persons without a MSCD
- $\beta_0 + \beta_1$ is the log odds of a big expenditure for persons with a MSCD
- $\beta_1 = (\beta_0 + \beta_1) - \beta_0$ is the difference in the log odds of a big expenditure comparing persons with a MSCD to those without.
- NOTE: $log(a) - log(b) = log(\frac{a}{b})$
- β_1 is the log odds ratio comparing the odds of a big expenditure among persons with and without a MSCD.

Practice: You estimate β_0 and β_1 and standard errors for the estimated coefficients!

D. Where to next?

In the lab session, you will extend this model to one that includes the interaction of $MSCD$ with $Older$, link the coeffients from the interaction model to analysis of stratified 2x2 tables AND implement these models within *glm* in R.