* Tues 3/23
* Thurs 3/25
Tues 3/30
Break day
* Thurs 4/1

JOHNS HOPKINS
BLOOMBERG SCHOOL of PUBLIC HEALTH

Lecture 1

Course Introduction

Introduction to Generalized Linear Models and Logistic Regression

Course Description

- 140.653 Linear regression for continuous outcomes
- 140.654 Regression for discrete outcomes plus some survival analysis
- 140.654 Introduction to machinic learning approaches (classification/regression trees and random forests)
- You will be learning the underlying theory behind how linear regression works with an emphasis on developing, fitting, interpreting and evaluating models to address specific scientific questions.

Course Objectives:

1. Formulate a scientific question about the relationship of a response variable Y and predictor variables X in terms of the appropriate logistic, log-linear or survival regression model

2. Interpret the meaning of regression coefficients in scientific terms as if for a substantive journal. For binary responses collected in clusters, distinguish between marginal and cluster-specific regression coefficients estimated by ordinary and conditional logistic regression

3. Develop graphical and/or tabular displays of the data to show the evidence relevant to describing the relationship of Y with X. For survival data, produce Kaplan-Meier and complimentary log, log plots of survival functions with standard errors

4. Estimate the model using a modern statistical package such as R and interpret the results for substantive colleagues. Derive the estimating equations for the maximum likelihood estimates for the class of generalized linear models and state the asymptotic distributions of the regression coefficients and linear combinations thereof

Course Objectives:

5. Give a heuristic derivation of the Cox proportional hazards estimating function in terms of Poisson regression for grouped survival data
6. Check the major assumptions of the model including independence and model form (mean, variance, proportional hazards) and make changes to the model or method of estimation and inference to appropriately handle violations. For example, use robust variance estimates for violations of independence or variance model
7. Use regression diagnostics to determine whether a small fraction of observations is having undue influence on the results
8. Correctly interpret the regression results to answer the specific substantive questions posed in terms that can be understood by substantive experts
9. Write a methods and results section for a substantive journal, correctly describing the regression model in scientific terms and the method used to specify and estimate the model
10. Critique the methods and results from the perspective of the statistical methods chosen and alternative approaches that might have been used

Key Dates

- Problem Set 1: Friday April 9th
- Quiz 1: Monday April 12th
- Problem Set 2: Thursday April 29th
- Quiz 2: Monday May 3rd
- Problem Set 3: Friday May 14th
- Quiz 3: Sunday May 16th
- Problem Set 4: Friday May 21st

Course Communication

- Direct email from course faculty and announcement via Courseplus in the event of major changes to due dates or important messages

- Slack workspace:
  - Please subscribe to this forum
  - We have set up topic categories to help keep things organized
  - No question is too big/small
  - Join #problemset1_654, #problemset2_654, #problemset3_654, #problemset4_654

- Questions about grading:
  - Please send questions about grading to Elizabeth only.

- General guidelines on personal emails to course faculty
  - Please use the Slack workspace for questions relating to course content and problem sets. There will be other students in the course who have similar questions as you. So by posting questions in a public forum, we gain efficiency
  - Please send personal communications (e.g. problem set due date extension request) to Elizabeth

Review: Linear model

- Model specification: y_i is a realization of Y_i $\quad \varepsilon_i \sim N(0, \sigma^2)$

  $i = 1, 2, \dots, n$ $\quad Y_i = \underbrace{\beta_0 + \beta_1 X_{1i} + \dots + \beta_p X_{pi}}_{\mu_i} + \varepsilon_i$ $\quad Y_i \sim N(\mu_i, \sigma^2)$

- Vector notation: $Y = X\beta + \varepsilon$, $\varepsilon \sim MVN(0, \sigma^2 I_{n\times n})$

  $Y_i = X_i'\beta + \varepsilon_i$ $\quad X_i$ = $(p+1)\times 1$ vector of 1s (intercept), $X_{1i}, X_{2i}, \dots, X_{pi}$

- Model can be decomposed into:
  - Systematic component $\mu_i = X_i'\beta \qquad \mu = X\beta$
  - Random component

$$Var(\varepsilon) = Var(Y) = \sigma^2 I_{n\times n}$$

Generalized Linear Models

GLMs

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

- GLMs have three components:
  - Random component: Distribution of Y_i $\Rightarrow$ mean $\mu_i = E(Y_i)$
    $Var(Y_i)$
  - Systematic component: $X_{1i}, X_{2i}, \ldots, X_{pi}$ $\Leftrightarrow$ $E(Y_i)$
    $g(\mu_i) = \underbrace{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi}}_{LP}$
  - Link function:
    mapping of random component to the systematic component.
    $g^{-1}(LP) = \mu_i$

Linear model: link function: identity link
$g(\mu_i) = \mu_i$

Generalized Linear Models

- Examples of data types / distributions / link functions

- Continuous

  Distribution: $Y_i \sim N(\mu_i, \sigma^2)$

  Identity link $= g(\mu_i) = \mu_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi}$

- Positive continuous

  Distribution: Gamma

  Link function: inverse link $g(\mu_i) = 1/\mu_i$ Gamma regression

- Binary $\rightarrow 0, 1$

  Distribution: Bernoulli

  Link function: Logit link $g(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right)$ Logistic regression

- Count, $0, 1, 2, 3, \ldots$

  Distribution: Poisson

  Link function: Log link $g(\mu_i) = \log(\mu_i)$ Poisson regression

  log-linear models

Generalized Linear Models

- Defines a regression model for outcome that is distributed according to a member of the exponential family of distributions.

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

Normal
Bernoulli
Gamma
Poisson

Generalized Linear Models: Exponential family

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

Normal distn

$$f_Y(y|\theta,\phi) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\{-(y-\mu)^2/2\sigma^2\}$$

$$= exp\{(y\underbrace{\mu}_{\theta} - \underbrace{\mu^2/2}_{b(\theta)})/\sigma^2 - 1/2(y^2/\sigma^2 + log(2\pi\sigma^2))\}$$

where

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

Generalized Linear Models: Exponential family

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

$p = Pr(Y=1)$

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

Bernoulli

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

where

- $\theta = log(p) + log(1-p) = \boxed{log(\frac{p}{1-p})} \rightarrow p = \frac{exp(\theta)}{1+exp(\theta)}$

  $p/1-p$ odds

  Canonical link function
- $\phi = 1$
- $a(\phi) = 1$
- $b(\theta) = -log(1-p) = -log(\frac{1}{1+exp(\theta)})$

Background for Bernoulli Distribution

- Properties of exponential family distributions: $E(Y) = b'(\theta)$, $\mathrm{Var(Y)} = -\mathrm{b}''(\theta)a(\varphi)$

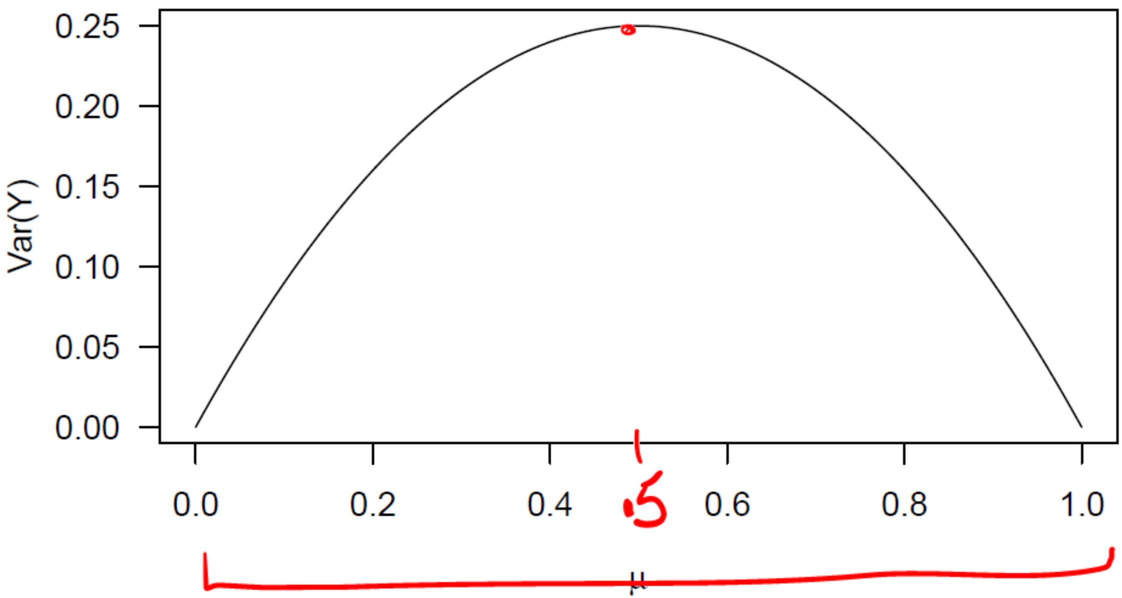
$$\begin{aligned} E(Y) &= \Sigma_{y=0}^{1} y Pr(Y=y) \\ &= 0 \times Pr(Y=0) + 1 \times Pr(Y=1) \\ &= 0 \times (1-\mu) + 1 \times \mu \\ &= \mu \end{aligned}$$

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

μ, σ^2

NOTE: Unlink the normal distribution, the variance of the Bernoulli is a function of the mean

Background for Bernoulli Distribution: Mean/Var relationship



Background for Bernoulli Distribution: Inference for mean

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 $\hat{Var}(\bar{Y}) = \frac{\hat{\mu}(1-\hat{\mu})}{n}$.

sample proportion of 1s

A 95% confidence interval for μ is given by:

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

$Pr(Y=1) = \mu$

Background for Bernoulli Distribution: Motivate link function

- Now we want to correlate the mean of the Bernoulli distribution with covariates!
- But we run into a challenge because the mean is bounded between 0 and 1.
- What if we can transform the mean to an unbounded space, perform the regression and then transform back?!?
  - Generalized linear models!

Probability	ODDS	Log ODDS
μ	$\frac{\mu}{1-\mu}$	$\boxed{log(\frac{\mu}{1-\mu})}$

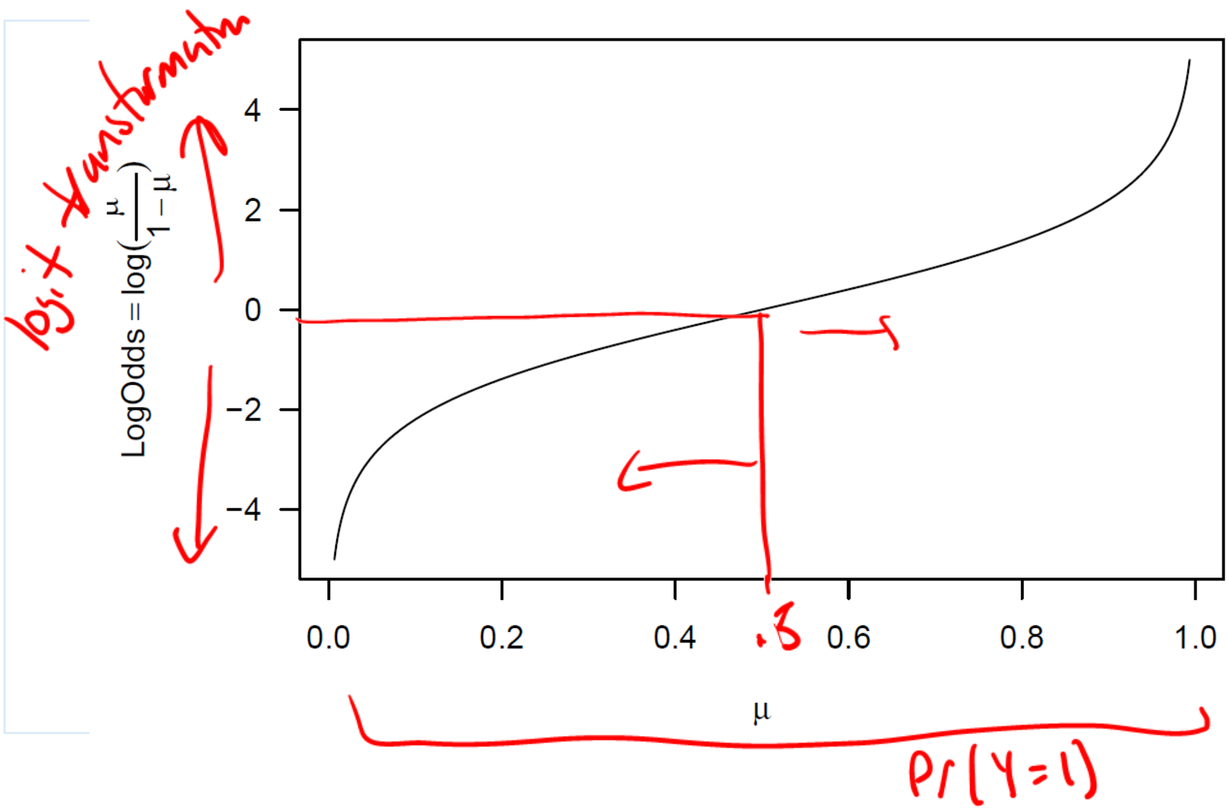
$(0,1)$ $[0,\infty)$ $(-\infty,0)+(0,\infty)$

$[0,1],[1,\infty)$ $(-\infty,\infty)$

Background for Bernoulli Distribution: Logit link function

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	$\frac{0.05}{0.95} = 0.05$	$log(0.05) = -2.99$

Background for Bernoulli Distribution: Logit link function



Some practice for you!

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

You practice:

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

Logistic regression model

- Simple logistic regression model: $Y_i \sim \text{Bernoulli}(\mu_i)$ $\Rightarrow$ random component

link function + systematic component

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 X_{1i}$$

* $\text{Var}(Y_i) = \mu_i(1-\mu_i)$

- Multiple logistic regression model:

$Y_i \sim \text{Bernoulli}(\mu_i)$

link function + systematic component

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 X_{1i} + \dots + \beta_p X_{pi}$$

Logistic regression models: Motivated by analysis of 2x2 tables

Lecture 16 → 653

- Consider the 1987 NMES
  - Outcome: big expenditure
  - Covariates:
    - MSCD : $\begin{cases} 1 & \text{LC, COPD, CVD, stroke + cancers} \\ 0 & \text{o/w} \end{cases}$
    - Older : $\begin{cases} 1 & >65 \text{ yrs} \\ 0 & \text{o/w} \end{cases}$

$$Y_i = \begin{cases} 0 & \text{if totalexp} < \$1000 \\ 1 & \text{if totalexp} \geq \$1000 \end{cases}$$

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

$$Pr(Y=1) = \frac{4335}{4335+7349}$$

$$\text{odds}(Y=1) = \frac{4335}{7349}$$

$$Pr(Y=1 \mid MSCD=1) = \frac{986}{986+333}$$

$$Pr(Y=1 \mid MSCD=0) = \frac{3349}{3349+7016}$$

relative rate or odds ratio

Intercept only logistic regression model

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

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

β_0 = log odds of a big expenditure

$$\hat{\beta}_0 = \log\left[4335/7349\right] = \underline{-.53}$$

$$se(\hat{\beta}_0) = \sqrt{\frac{1}{4335} + \frac{1}{7349}} \underline{= .019}$$

odds of a big expenditure $= \exp(-.53) = .59$

$$Pr(\text{Big expenditure}) = \frac{\exp(-.53)}{1+\exp(-.53)} = .37$$

$Pr(Y=1)$

$$\frac{4335}{4335+7349}$$

Simple logistic regression model

risk difference : $Pr(Y=1|MSCD=1) - Pr(Y=1|MSCD=0)$

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

risk ratio : $\dfrac{Pr(Y=1|MSCD=1)}{Pr(Y=1|MSCD=0)}$

Odds ratio : $\dfrac{Pr(Y=1|MSCD=1)}{Pr(Y=0|MSCD=1)} \Big/ \dfrac{Pr(Y=1|MSCD=0)}{Pr(Y=0|MSCD=0)}$

$= \dfrac{Pr(Y=1|MSCD=1)\,Pr(Y=0|MSCD=0)}{Pr(Y=1|MSCD=0)\,Pr(Y=0|MSCD=1)}$

$se(\log OR) = \sqrt{\frac{1}{986}+\frac{1}{333}+\frac{1}{3349}+\frac{1}{7016}} = .067$

$986/333 \Big/ \dfrac{3349}{7016} = 6.2$ $\quad \log(OR) = 1.83$

Simple logistic regression model

$$\log\left[\frac{\mu_i}{1-\mu_i}\right] = \beta_0 + \beta_1 \text{MSCD}_i$$

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

MSCD = 0 $\quad \log\left[\frac{\mu_i}{1-\mu_i}\right] = \beta_0$

log odds of Big expenditure among persons w/o a MSCD

$$\hat\beta_0 = \log\left[3349/7016\right] \qquad se(\hat\beta_0) = \sqrt{\frac{1}{3349} + \frac{1}{7016}}$$

MSCD = 1 $\quad \log\left[\frac{\mu_i}{1-\mu_i}\right] = \beta_0 + \beta_1$ log odds of Big \$ among persons w MSCD

$$\hat\beta_0 + \hat\beta_1 = \log\left[986/333\right] \qquad se(\hat\beta_0 + \hat\beta_1) = \sqrt{\frac{1}{986} + \frac{1}{333}}$$

β_1 = difference in log odds of Big exp comparing MSCD=1 to MSCD=0

$$\log(a) - \log(b) = \log(a/b)$$

= log of ratio of odds = log(OR)

$$= \log\left(986/333 \,\big/\, 3349/7016\right) \qquad se(\hat\beta_1) = \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$$

Where to next?

- In Lab 1, you will continue with this example but consider analyzing stratified 2x2 tables
  - Equates to a logistic regression for two binary covariates, including interaction
- Lecture 2:
  - Multiple logistic regression models cont.
  - Adjustment for binary confounder
  - Adjustment for continuous confounder
  - Exploring functional forms for continuous covariates
  - Assessing for confounding in generalized linear models