Lecture 2 Handout

Elizabeth Colantuoni

3/23/2021

I. Objectives

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

- Connect logistic regression to the analysis of 2x2 tables
- Describe how covariate adjustment works in logistic regression analysis
- Create visual displays of data to motivate assumptions you are making for continuous covariates in logistic regression models

II. Lecture 1 Review

A. Generalized Linear Models

Generalized linear models are a class of regression models that can be formulated for any outcome whose distribution is in the exponential family of distributions.

To define a generalized linear model, we need to specify:

- 1. The *random* component: Distribution of Y defining the $E(Y)$ and $Var(Y)$ with Y independent.
- 2. The *systematic* component: Defines the linear model for a function g of $E(Y) = \mu$, i.e. for subject i $g(\mu_i) = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi}$.
- 3. The *link* function: The mapping g taking the $E(Y) = \mu$ and linking it to the systematic component/linear predictor.

$$g(\mu) = X\beta$$

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

B. General Linear Model as a Generalized Linear Model

For the general linear model, we can specifiy:

- 1. The *random* component: $Y_i \sim N(\mu_i, \sigma^2)$, with $i = 1, ..., n$ independent observations.
- 2. The *systematic* component: $g(\mu_i) = X_i^{'}\beta$ where $X_i^{'}\beta$ is determined from the scientific question of interest.
- 3. The *link* function: $g(\mu_i) = \mu_i$, the Identity link!

C. Logistic Regression Model as a Generalized Linear Model

For the logistic regression model, we can specify:

- 1. The *random* component: $Y_i \sim Bernoulli(\mu_i)$, with $i = 1, ..., n$ independent observations. This implies $E(Y_i) = Pr(Y_i = 1) = \mu_i$ and $Var(Y_i) = \mu_i(1 - \mu_i)$.
- 2. The *systematic* component: $g(\mu_i) = X_i^{'}\beta$ where $X_i^{'}\beta$ is determined from the scientific question of interest.
- 3. The *link* function: $g(\mu_i) = log(\frac{\mu_i}{1-\mu_i})$, the logit link. To translate back to the mean, $\mu_i = g^{-1}(X_i^{'}\beta)$ use the inverse-logit function given by:

$$g^{-1}(a) = \frac{exp(a)}{1 + exp(a)}$$

D. Simple Logistic Regression

Define a simple logistic regression model with

- The outcome Y_i is 1 or 0.
- The primary exposure X_i is 1 or 0.
- Define $\mu_i = Pr(Y_i = 1|X_i)$
- Define $odds(Pr(Y_i = 1|X_i)) = \frac{Pr(Y_i=1|X_i)}{1-Pr(Y_i=1|X_i)}$
- Define $logit(Pr(Y_i = 1|X_i)) = log(\frac{Pr(Y_i=1|X_i)}{Pr(Y_i=0|X_i)})$
- Define $OR(Y_i, X_i) = \frac{Pr(Y_i=1|X_i=1)}{Pr(Y_i=0|X_i=1)} / \frac{Pr(Y_i=1|X_i=0)}{Pr(Y_i=0|X_i=0)}$
- The simple logistic regression is:

$$logit[\mu_i] = logit[Pr(Y_i = 1|X_i)] = \beta_0 + \beta_1 X_i$$

- The interpretation of the intercept is:

$$\beta_0 = logit[Pr(Y_i = 1|X_i = 0)] = log\left\{\frac{Pr(Y_i = 1|X_i = 0)}{Pr(Y_i = 0|X_i = 0)}\right\}$$

- The value of the intercept can be back-transformed to the probability scale:

$$\frac{exp(\beta_0)}{1 + exp(\beta_0)} = Pr(Y_i = 1|X_i = 0)$$

- The intepretation of the slope for X is:

$$\beta_1 = log\left\{\frac{odds(Y_i = 1|X_i = 1)}{odds(Y_i = 1|X_i = 0)}\right\} = \text{ log odds ratio}$$

III. Logistic regression models motivated through analysis of 2x2 tables

In Lecture 1 and Lab 1, we motivated the interpretation of logistic regression model parameters via analysis of 2x2 tables using the NMES data.

We will continue this discussion here by briefly reviewing the models you fit and considering additional models.

Back to our NMES example,

- The outcome is "big expenditure": $Y_i = 1$ if $totalexp_i > 1000, 0$ otherwise
- The primary exposure is "major smoking caused disease": $MSCD_i$, 1 if yes, 0 if no
- We will also explore "older": $older_i = 1$ if $age_i > 65, 0$ otherwise, as an effect modifier and confounder.

We will consider 4 models:

1. Intercept only model, Model A: $logit[Pr(Y_i = 1)] = \beta_0$
2. Main term of $MSCD$, Model B: $logit[Pr(Y_i = 1|MSCD_i)] = \beta_0 + \beta_1$
3. Interaction model, Model C:

$$logit[Pr(Y_i = 1|MSCD_i, Older_i)] = \beta_0 + \beta_1 MSCD_i + \beta_2 Older_i + \beta_3 MSCD_i \times Older_i$$

4. Main effects of $MSCD$ and $Older$, Model D:

$$logit[Pr(Y_i = 1|MSCD_i, Older_i)] = \beta_0 + \beta_1 MSCD_i + \beta_2 Older_i$$

A. Model B fit and interpretation

```
load('./nmes.rdata')
data = nmes
data[data=='.'] = NA

## Create the necessary variables:
data$posexp=ifelse(data$totalexp>0,1,0)
data$mscd=ifelse(data$lc5+data$chd5>0,1,0)
data1=data[!is.na(data$eversmk),]
data1$older=ifelse(data1$lastage<65,0,1)
data1$bigexp=ifelse(data1$totalexp>1000,1,0)

## Model B
modelB = glm(bigexp~mscd,data=data1,family="binomial")
lincom(modelB,c("(Intercept)","mscd"))
```

```
##             Estimate   2.5 %      97.5 %     Chisq    Pr(>Chisq)
## (Intercept) -0.7395315 -0.7806967 -0.6983663 1239.792 1.372226e-271
## mscd        1.825045   1.694177   1.955913   747.095  1.718138e-164
```

```
lincom(modelB,c("(Intercept)","mscd"),eform=TRUE)
```

```
##             Estimate  2.5 %     97.5 %    Chisq    Pr(>Chisq)
## (Intercept) 0.4773375 0.4580868 0.4973973 1239.792 1.372226e-271
## mscd        6.203076  5.442166  7.070374  747.095  1.718138e-164
```

From the output, we observe:

- The estimated log odds (95% CI) and odds of a big expenditure among persons without a MSCD are: $\hat{\beta}_0 = -0.74(-0.78, -0.70), exp(\hat{\beta}_0) = 0.48(0.46, 0.50)$, respectively.
- The estimates log odds ratio (95% CI) and odds ratio (95% CI) comparing the odds of a big expenditure among persons with and without a MSCD are: $\hat{\beta}_1 = 1.83(1.70, 1.96), exp(\hat{\beta}_1) = 6.20(5.44, 7.08)$, respectively.
- The odds of a big expenditure for persons with a MSCD are 6.20 times the odds of a big expenditure for persons without a MSCD.
- The odds of a big expenditure for persons with a MSCD are 520% greater, i.e. 100*(6.20-1), than the odds of a big expenditure for persons without a MSCD.

B. Model C fit and interpretation

```
## Model C
modelC = glm(bigexp~mscd+older+mscd:older,data=data1,family="binomial")
lincom(modelC,c("mscd","mscd+mscd:older","mscd:older"))
```

```
##                 Estimate   2.5 %      97.5 %    Chisq    Pr(>Chisq)
## mscd            1.969895   1.735287   2.204503  270.8301 7.481555e-61
## mscd+mscd:older 1.491115   1.329415   1.652815  326.6619 5.126712e-73
## mscd:older      -0.4787796 -0.7637143 -0.193845 10.84618 0.0009899951
```

```
lincom(modelC,c("mscd","mscd+mscd:older","mscd:older"),eform=TRUE)
```

```
##                 Estimate 2.5 %     97.5 %    Chisq    Pr(>Chisq)
## mscd            7.169921 5.670554  9.065741  270.8301 7.481555e-61
## mscd+mscd:older 4.442046 3.778832  5.221658  326.6619 5.126712e-73
## mscd:older      0.619539 0.4659326 0.8237856 10.84618 0.0009899951
```

From the output we observe:

- The estimated log odds ratio ($\hat{\beta}_1$) and odds ratio ($exp(\hat{\beta}_1)$) comparing the odds of a big expenditure among persons 65 years of age or younger with and without a MSCD is 1.97 (1.74, 2.20) and 7.17 (5.67, 9.06), respectively.
- The estimated log odds ratio ($\hat{\beta}_1 + \hat{\beta}_3$) and odds ratio ($exp(\hat{\beta}_1 + \hat{\beta}_3)$) comparing the odds of a big expenditure among persons over the age of 65 with and without a MSCD is 1.49 (1.33, 1.65) and 4.44 (3.78, 5.22), respectively.
- NOTE: The estimate of the coefficient for the interaction between $MSCD$ and $Older$ (β_3) is -0.48. Recall this is the **difference in the log odds ratios** for a big expenditure comparing persons with and without a MSCD among older vs. younger persons (1.49 - 1.97 = -0.48).
- NOTE: The estimate of the exponentiated coefficient for the interaction between $MSCD$ and $Older$ is 0.62. Recall this is the **ratio of the odds ratios* comparing the odds ratio of a big expenditure and MSCD for older vs. younger persons, i.e.

$$\frac{4.44}{7.17} = 0.62$$

C. Model D fit and interpretation

$$logit[Pr(Y_i = 1|MSCD_i, Older_i)] = \beta_0 + \beta_1 MSCD_i + \beta_2 Older_i$$

What assumption is Model D making? Think "adjustment" in linear regression!

- This model assumes $OR(Y_i, MSCD_i|Older_i = 1) = OR(Y_i, MSCD_i|Older_i = 0)$
- I.e. The relative odds of a big expenditure comparing persons with and without a $MSCD$ are the same for younger and older people.

How would we go about estimating β_1?

- Think inverse variance weighting; same as we did in linear regression!

Age group	$log\hat{OR}$	$se(log\hat{OR})$	$var(log\hat{OR})$	$\frac{1}{var(log\hat{OR})}$	$w = \frac{\frac{1}{var(log\hat{OR})}}{\Sigma(\frac{1}{var(log\hat{OR})})}$
Younger	1.97	0.12	0.0144	69.4	0.32
Older	1.49	0.083	0.0069	144.9	0.68

$$\hat{\beta}_1 = 1.97 \times 0.32 + 1.49 \times 0.68 = 1.64$$

$$se(\hat{\beta}_1) = \frac{1}{\sqrt{\sum(\frac{1}{var(log\hat{OR})})}} = \frac{1}{\sqrt{214.3}} = 0.068$$

```
modelD = glm(bigexp~mscd+older,data=data1,family="binomial")
summary(modelD)$coeff
```

```
##               Estimate Std. Error   z value      Pr(>|z|)
## (Intercept) -0.9577826 0.02700779 -35.46321 1.815505e-275
## mscd         1.6549130 0.06803662  24.32386 1.096494e-130
## older        0.5638298 0.04104938  13.73540  6.230701e-43
```

```
lincom(modelD,c("mscd","older"),eform=TRUE)
```

```
##       Estimate 2.5 %    97.5 %   Chisq    Pr(>Chisq)
## mscd  5.232625 4.57938  5.979054 591.65   1.096494e-130
## older 1.75739  1.621537 1.904625 188.6613 6.230701e-43
```

You practice: Use the output above, interpret $exp(\hat{\beta}_2)$.

IV. Adjustment for continuous variable

Instead of making age_i binary, suppose we want to treat age as a continuous variable for the adjustment!

Then we can modify Model D as follows:

$$logit[Pr(Y_i = 1|MSCD_i, age_i)] = \beta_0 + \beta_1 MSCD_i + \beta_2 age_i$$

- Can you draw a picture of this model?

```r
modelDagecont = glm(bigexp~mscd+lastage,data=data1,family="binomial")
summary(modelDagecont)$coeff
```

```
##                Estimate  Std. Error   z value      Pr(>|z|)
## (Intercept) -2.27990966 0.099135981 -22.99780 4.903428e-117
## mscd         1.60502065 0.068269770  23.50998 3.224831e-122
## lastage      0.02574057 0.001599682  16.09105  2.947835e-58
```

```r
lincom(modelDagecont,c("mscd","lastage"),eform=TRUE)
```

```
##         Estimate 2.5 %    97.5 %   Chisq   Pr(>Chisq)
## mscd    4.977962 4.35452  5.690664 552.719 3.224831e-122
## lastage 1.026075 1.022863 1.029297 258.922 2.947835e-58
```

From the fit of the model we estimate:

- For persons of the same age, the odds of a big expenditure among those with a MSCD are roughly 5 times the odds among those without a MSCD (estimated odds ratio: 4.98, 95% CI: 4.35 to 5.69).
- Among persons with the same disease status (i.e. have a MSCD or not), the odds of a big expenditure increase by 2.6 percent, i.e. $100(1.026 - 1) = 2$, per additional year of age (estimated odds ratio: 1.026, 95% CI: 1.023 to 1.029).

A. How do we know if the log odds change linearly with age?

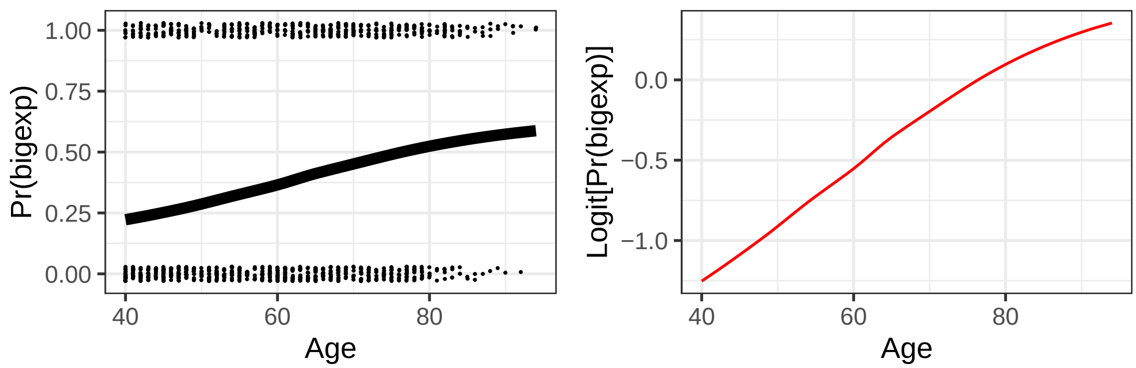
We can explore this visually!

```r
# Make a plot exploring Pr(bigexp) ~ age
fit <- loess(bigexp ~ lastage, alpha=0.2, data=data1)
data1$smoothProb = fit$fitted
data1$smoothLogit = log(fit$fitted/(1-fit$fitted))

# For plotting, take a sample of the large dataset
data2 = data1[sample(seq(1,nrow(data1)),1000),]
plot1 = ggplot(data2, aes(lastage, bigexp)) + theme_bw() +
         geom_line(aes(lastage,smoothProb), size=2) +
  geom_point(position=position_jitter(height=0.03, width=0.03),size=0.05) +
  xlab("Age") + ylab("Pr(bigexp)")

# Make a plot exploring logit[Pr(bigexp)] ~ age
plot2 = ggplot(data2,aes(lastage, bigexp)) + xlab("Age") + ylab("Logit[Pr(bigexp)]") + theme_bw() +
  geom_line(aes(lastage, smoothLogit), color="red")

suppressWarnings(grid.arrange(plot1,plot2,ncol=2))
```



Perhaps modeling age as a linear spline with knot at 65? 70? 80?

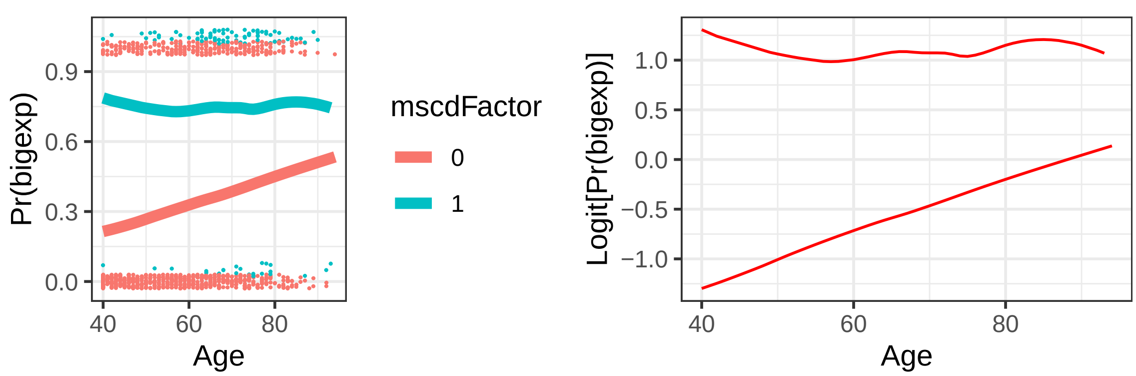
Now do the same descriptive analysis stratified by $MSCD$.

```
# Make a plot exploring Pr(bigexp) ~ age for each mscd group
data1$mscdFactor = as.factor(data1$mscd)
data1$smoothProb2 = 0
data1$smoothLogit2 = 0
fit <- loess(bigexp ~ lastage, alpha=0.2, data=data1[data1$mscd==1,])
data1$smoothProb2[data1$mscd==1] = fit$fitted
data1$smoothLogit2[data1$mscd==1] = log(fit$fitted/(1-fit$fitted))
fit <- loess(bigexp ~ lastage, alpha=0.2, data=data1[data1$mscd==0,])
data1$smoothProb2[data1$mscd==0] = fit$fitted
data1$smoothLogit2[data1$mscd==0] = log(fit$fitted/(1-fit$fitted))
# Create a new bigexp variable so we can separately see the data by mscd
data1$newy = data1$bigexp + data1$mscd*0.05

data2 = data1[sample(seq(1,nrow(data1)),1000),]

plot1 = ggplot(data2, aes(lastage, newy, color=mscdFactor)) +
         geom_line(aes(lastage,smoothProb2), size=2) +
  geom_point(position=position_jitter(height=0.03, width=0.03),size=0.05) + theme_bw() +
  xlab("Age") + ylab("Pr(bigexp)")

# Make a plot exploring logit[Pr(bigexp)] ~ age
plot2 = ggplot(data2,aes(lastage, bigexp,color=mscdFactor)) +
  xlab("Age") + ylab("Logit[Pr(bigexp)]") +
  geom_line(aes(lastage, smoothLogit2,group=mscdFactor), color="red") + theme_bw()
suppressWarnings(grid.arrange(plot1,plot2,ncol=2))
```



In this analysis where we have stratified by $MSCD$ status, we see a very strong interaction between age and $MSCD$. Our earlier analysis supported this!

Also, it looks like it would not be so horrible to assume the log odds of a big expenditure are linear with age.

For fun: Fit the interaction model again, but this time with age as a continuous variable. Interpret all the coefficients in the model!

```r
# Helpful to center lastage in an interaction model like this
# so that the main effect of mscd has a useful interpretation
data1$age_c = data1$lastage - 60

modelCcont = glm(bigexp~mscd+age_c+mscd:age_c,data=data1,family="binomial")
lincom(modelCcont,c("mscd","mscd+mscd:age_c","mscd+20*mscd:age_c","mscd:age_c"))
```

```
##                     Estimate   2.5 %       97.5 %      Chisq    Pr(>Chisq)
## mscd                1.792367   1.625218    1.959516    441.7169 4.579093e-98
## mscd+mscd:age_c     1.768144   1.607967    1.928321    468.0905 8.342527e-104
## mscd+20*mscd:age_c 1.307903   1.113342    1.502464    173.595  1.213445e-39
## mscd:age_c          -0.0242232 -0.03631431 -0.01213209 15.41797 8.616514e-05
```

```r
lincom(modelCcont,c("mscd","mscd+mscd:age_c","mscd+20*mscd:age_c","mscd:age_c"),eform=TRUE)
```

```
##                     Estimate  2.5 %     97.5 %    Chisq    Pr(>Chisq)
## mscd                6.003646  5.079528  7.09589   441.7169 4.579093e-98
## mscd+mscd:age_c     5.859966  4.992649  6.877953  468.0905 8.342527e-104
## mscd+20*mscd:age_c 3.69841   3.044517  4.492744  173.595  1.213445e-39
## mscd:age_c          0.9760678 0.9643371 0.9879412 15.41797 8.616514e-05
```

V. Identifying confounding in logistic models

We will discuss this next time!

To get a head start on this discussion, please see Handout linked with today's class written by Dr. Scott Zeger "Note on confounding and effect modification 2019".

You may also be interested in reading:

- Rothman and Greenland, Modern Epidemiology, pp 52-55.
- Janes, Holly, Francesca Dominici, and Scott Zeger. "On quantifying the magnitude of confounding." Biostatistics 11.3 (2010): 572-582.

I will be highlighting some of the results from this handout and the Janes et al paper in Lecture 3.