Lecture 14: Interpreting logistic regression models

Sandy Eckel
seckel@jhsph.edu

15 May 2008

Logistic regression

- Framework and ideas of logistic regression similar to linear regression
- Still have a systematic and probabilistic part to any model
- Coefficients have a new interpretation, based on log(odds) and log(odds ratios)

Recall from last time: The logit function

- In logistic regression, we are always modelling the outcome log(p/(1-p))
- We define the function:

$$\text{logit}(p) = \log(p/(1-p))$$

- We often use the name **logit** for convenience
- In logistic regression, we have the logit on the left-hand side of the equation

Example: Public health graduate students

- 323 graduate students in introductory biostatistics took a health survey. Current smoking status was assessed, which we will predict with gender
  - Associating demographics with smoking is vital to planning public health programs.
  - Information was also collected on age, exercise, and history of smoking; potential confounders of the association between gender and current smoking.
  - First we will focus only on the association between gender and current smoking status

Coding our two variables for the first example

- Outcome:
  - smoking = 1 for current smokers
    0 for current nonsmokers
- Primary predictor:
  - gender = 1 for men
    0 for women

Recall: an analogous linear regression model

- In linear regression, if we had only one binary X like gender, we would be predicting two means: $E(Y) = \beta_0 + \beta_1(Gender)$
  - β_0 – the mean outcome when X=0
  - $\beta_0 + \beta_1$ – the mean outcome when X=1
  - β_1 – the ***difference*** in mean outcome when X=1 vs. when X=0

Logistic regression model and Results

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(Gender) \quad \Rightarrow \quad \log\left(\frac{p}{1-p}\right) = -3.1 + 1.0(Gender)$$

```
Logit estimates                                   Number of obs   =        323
                                                  LR chi2(1)      =       4.46
                                                  Prob > chi2     =     0.0348
Log likelihood = -75.469757                       Pseudo R2       =     0.0287

------------------------------------------------------------------------------
       smoke |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |    .967966   .4547931     2.13   0.033     .0765879    1.859344
  (Intercept)|  -3.058707   .3235656    -9.45   0.000    -3.692884    -2.42453
------------------------------------------------------------------------------
```

gender = 1 for men
0 for women

Logistic Regression Gender-specific results

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(Gender) \quad \Rightarrow \quad \ln\left(\frac{p}{1-p}\right) = -3.1 + 1.0(Gender)$$

- For women, gender=0: $\ln\left(\frac{p}{1-p}\right) = -3.1 + 1.0(0) = -3.1$
- For men, gender=1: $\ln\left(\frac{p}{1-p}\right) = -3.1 + 1.0(1) = -2.1$
- β_1 is the difference between men and women
- β_1 is the ***change in log odds*** comparing men to women

Logistic Regression
Interpretation 1: log(odds) scale

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(Gender) \quad \Rightarrow \quad \ln\left(\frac{p}{1-p}\right) = -3.1 + 1.0(Gender)$$

gender = 1 for men
0 for women

- β_0: the ***log odds*** of smoking for women
- $\beta_0+\beta_1$: the ***log odds*** of smoking for men
- β_1: the ***difference*** in the log odds of smoking for men compared to women

What if we wanted to get the odds interpretation, not the log odds…

- We can start to "untransform" the equations
  - Recall:

    $$\text{if} \quad \log(a) = b, \quad \text{then} \quad \exp(\log(a)) = a = e^{b}$$

  - For women, X=0: log(odds)= $\beta_0+\beta_1(0) = \beta_0$

    $$\text{odds of smoking for women} = e^{\beta_0} = e^{-3.1} = 0.05$$

  - For men, X=1: log(odds)= $\beta_0+\beta_1(1)$

    $$\text{odds of smoking for men} = e^{\beta_0+\beta_1} = e^{-3.1+1.0} = e^{-2.1} = 0.12$$

Logistic Regression
Interpretation 2: odds scale

- e^{β_0}: the ***odds*** of smoking for women (when X=0)
- $e^{\beta_0+\beta_1}$: the ***odds*** of smoking for men (when X=1)
- In the past, we've compared two sets of odds by dividing to find the odds ratio (OR)

Comparing odds

- If we ***subtract*** the log odds, mathematically that's equivalent to dividing inside the log:
  - log(a) – log(b) = log(a/b)
- So, if
  - $e^{\beta_0+\beta_1} = e^{-3.1+1.0} = e^{-2.1} = 0.12$ is the odds when X=1, and
  - $e^{\beta_0} = e^{-3.1} = 0.05$ is the odds when X=0, then
  - we want to ***divide*** them in order to compare

$$\text{Odds Ratio} = \frac{\text{odds for men}}{\text{odds for women}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = \frac{0.12}{0.05} = 2.4$$

Logistic Regression
Interpretation: the odds ratio

$$\text{Odds Ratio} = \frac{\text{odds for men}}{\text{odds for women}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = \frac{0.12}{0.05} = 2.4$$

- The odds of smoking is about 2 ½ times greater for men than for women.
- Based on this study, perhaps smoking **cessation** programs should be targeted toward men

Useful math – ratios of exponentiated terms

- We can usually simplify an equation like this

$$\text{Odds Ratio} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}}$$
$$= e^{(\beta_0+\beta_1)-(\beta_0)}$$
$$= e^{\beta_1}$$

because $\frac{e^a}{e^b} = e^{a-b}$

Taking a ratio of odds to get the odds ratio

- e^{β_0} : the ***odds*** when X=0
- $e^{\beta_0+\beta_1}$: the ***odds*** when X=1
- $\frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$ the ***odds ratio***

  comparing the odds when X=1 vs. X=0

Two interpretations of logistic regression slopes

- $\beta_0+\beta_1$ = log(odds) (for X=1)
  - β_1 = ***difference*** in log odds
- $e^{\beta_0+\beta_1}$ = odds (for X=1)
  - e^{β_1} = odds ***ratio***
- But we started with P(Y=1)
- Can we find that?

More useful math – how to get the probability from the odds

- $\text{odds} = \dfrac{\text{probability}}{1 - \text{probability}}$
- $\text{probability} = \dfrac{\text{odds}}{1 + \text{odds}}$
- so $\quad P(X=1) = \dfrac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$

Finding the probability from the log odds

Find the log odds:

For X=0: log(odds) = β_0

For X=1: log(odds) = $\beta_0 + \beta_1$

Find odds:

For X=0: odds = e^{β_0}

For X=1: odds = $e^{\beta_0+\beta_1}$

Transform odds into probability:

(next slide…)

Finding the probability from the log odds, cont…

Transform odds into probability:

$$p = \frac{\text{odds}}{1+\text{odds}}$$

$$\text{For } X=0: \quad \text{probability} = \frac{e^{\beta_0}}{1+e^{\beta_0}}$$

$$\text{For } X=1: \quad \text{probability} = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$$

We could even go one step further

- $\text{Relative Risk (RR)} = \dfrac{p_1}{p_2}$
- $\text{For } X=1: \quad P(\text{smoke} \mid \text{male}) = \dfrac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$

  $\text{For } X=0: \quad P(\text{smoke} \mid \text{female}) = \dfrac{e^{\beta_0}}{1+e^{\beta_0}}$
- $\text{Relative Risk for Men vs. Women}: \quad \dfrac{p_1}{p_2} = \dfrac{\left(\dfrac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)}{\left(\dfrac{e^{\beta_0}}{1+e^{\beta_0}}\right)}$
  - no way to simplify

Remember to consider study design

- We always *can* calculate the relative risk
- The relative risk is not appropriate for case-control studies
  - Again, because the investigators decide the number of cases and controls to study
- The odds ratio is appropriate for all study designs

In General

- Logistic regression for a binary outcome
- Left side of equation is log odds
  - Can transform the equation to find
    - odds
    - probability
  - Can compare two groups
    - difference of log odds ≡ log odds ratio
    - odds ratio
    - relative risk
- (Almost) everything we learned before applies

Summary: Useful math for logistic regression

- $\textit{If}\ \log(\mathrm{a}) = b,\ \text{then}\ \exp(\log(\mathrm{a})) = \mathrm{a} = \mathrm{e}^{b}$

  X=1: log(odds)= $\beta_0+\beta_1(1)$ $\quad$ so odds for $(\mathrm{X}=1) = e^{\beta_0+\beta_1}$

- log(a) – log(b) = log(a/b)

  so log(odds|X=1) – log(odds|X=0) = log(OR for X=1 vs. X=0)

- $\dfrac{e^a}{e^b} = e^{a-b}$ $\quad$ so $\quad \dfrac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$ $\qquad$ Also: $e^{a+b} = e^a \times e^b$

  so $e^{2\beta_1} = e^{\beta_1} \times e^{\beta_1} = \left(e^{\beta_1}\right)^2$

- $\text{probability} = \dfrac{\text{odds}}{1+\text{odds}}$

  $$\text{so probability for } (\mathrm{X}=1) = \frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}$$

Another Example

- Regular physical examination is an important preventative public health measure
- We'll study this outcome using the public health graduate student dataset
  - Outcome: No physical exam in the past two years
  - Primary predictor: age (centered)
  - Secondary predictor and potential confounder: regularly taking a multivitamin

Problem with outcome variable:

- The original "physician visit" variable was meant to be continuous, but it was collected categorically
  - time since last physician visit
- Since it is now categorical and we wish to use it as the outcome for a regression model, we will make it binary and use logistic regression

Phys = 1 if over 2 years
0 if 2 years or less

Length of time since last check-up	Freq.	Percent	Cum.
Within the past year	182	54.17	54.17
Within the past 1-2 years	72	21.43	75.60
Within the past 2-5 years	53	15.77	91.37
5 or more years	29	8.63	100.00
Total	336	100.00	

Goals

- **Predict Phys (no physician visit within the past two years=1) with centered Age (continuous)**
- After adjusting for age, is taking a multivitamin (1=yes) a statistically significant predictor for not regularly visiting a physician?
- Is taking a multivitamin a confounder for the age-physician visit relationship?

Results
Model 1: Intercept and Age

Note that agec = age-30 (centered age)

```
Logit estimates                                   Number of obs   =        336
                                                  LR chi2(1)      =       0.00
                                                  Prob > chi2     =     0.9567
Log likelihood = -186.71399                       Pseudo R2       =     0.0000
```

phys_no	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
agec	-.0009585	.0176509	-0.05	0.957	-.0355536	.0336365
(Intercept)	-1.130428	.1270539	-8.90	0.000	-1.379449	-.8814066

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1(Age-30) \quad \Rightarrow \quad \log\left(\frac{p}{1-p}\right)=\text{-}1.13-0.001(Age-30)$$

Model 1:
Interpretation of coefficients on log odds scale

- β_0: the log odds of not visiting a physician for a 30-year-old
- β_1: the difference in the log odds of not visiting a physician for a one year increase in age

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1(Age-30) \quad \Rightarrow \quad \log\left(\frac{p}{1-p}\right)=\text{-}1.13-0.001(Age-30)$$

Model 1: How did we get the difference in log odds interpretation of β_1 ?

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(Age-30) \quad \Rightarrow \quad \log\left(\frac{p}{1-p}\right) = -1.13 - 0.001(Age-30)$$

- For a 30-year-old:

$$\log\left(\frac{p}{1-p}\right) = -1.13 - 0.001(30-30) = -1.13$$

- For a 31-year-old:

$$\log\left(\frac{p}{1-p}\right) = -1.13 - 0.001(31-30) = -1.13 - 0.001 = -1.129$$

- β_1 is the difference in the log odds associated with a 1 year increase in age

Model 1: Interpretation of β_1 (diff log odds = log OR)

- log(a) – log(b) = log(a/b)
  - so log(odds|X=31) – log(odds|X=30) = log(OR for X=31 vs. X=30)
  - **difference of log odds = log odds ratio**
- Alternate interpretation for β_1:
  - The ***log odds ratio*** of not visiting a physician associated with a one year increase in age

Model 1: Interpretation of β_1 (OR = ratio of odds)

$$\text{odds of not visiting a physician} = \frac{p}{1-p} = e^{-1.13-0.001(Age-30)}$$

- For a 31-year-old:

$$\frac{p}{1-p} = e^{-1.13-0.001(31-30)} = e^{-1.13-0.001} = e^{-1.131} = 0.3227$$

- For a 30-year-old:

$$\frac{p}{1-p} = e^{-1.13} = 0.3230$$

- Odds ratio = $\dfrac{0.3227}{0.3230} = 0.999 = \dfrac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$

Model 1: Interpretation of β_1 odds ratio for one year age difference

- e^{β_0} is the odds of not visiting a physician for 30-year-olds
- $e^{\beta_0+\beta_1}$ is the odds of not visiting a physician for 31-year-olds
- e^{β_1} is the odds ratio of not visiting a physician corresponding to a one year increase in age

Model 1: Interpretation of β_1 What is the OR for ***two*** year age difference?

$$\text{odds of not visiting a physician} = \frac{p}{1-p} = e^{-1.13-0.001(Age-30)}$$

- For a 32-year-old:

$$\frac{p}{1-p} = e^{-1.13-0.001(32-30)} = e^{-1.13-0.001\times 2} = e^{-1.132} = 0.3224$$

- For a 30-year-old:

$$\frac{p}{1-p} = e^{-1.13} = 0.3230$$

- Ratio = $\dfrac{0.3224}{0.3230} = 0.998 = \dfrac{e^{\beta_0+2\beta_1}}{e^{\beta_0}} = e^{2\beta_1} = \left(e^{\beta_1}\right)^2$

Model 1: Interpretation of β_1 What is the OR for ***ten*** year age difference?

$$\text{odds of not visiting a physician} = \frac{p}{1-p} = e^{-1.13-0.001(Age-30)}$$

- For a 40-year-old:

$$\frac{p}{1-p} = e^{-1.13-0.001(40-30)} = e^{-1.13-0.01} = e^{-1.14} = 0.3198$$

- For a 30-year-old:

$$\frac{p}{1-p} = e^{-1.13} = 0.3230$$

- Ratio = $\dfrac{0.3198}{0.3230} = 0.990 = \dfrac{e^{\beta_0+10\beta_1}}{e^{\beta_0}} = e^{10\beta_1} = \left(e^{\beta_1}\right)^{10}$

Model 1: Interpretation of β_1 What is the OR for ***any*** age difference?

- e^{β_1} is the ***proportional increase*** of the odds of not visiting a physician corresponding to a one year increase in age

$$(\text{odds for 30-yr-old}) \times \boxed{\frac{(\text{odds for 31-yr-old})}{(\text{odds for 30-yr-old})}} = (\text{odds for 31-yr-old})$$

- $\left(e^{\beta_1}\right)^{10} = e^{10\beta_1}$ is the ***proportional increase*** of the odds of not visiting a physician corresponding to a ten year increase in age

Model 1: How could we get a Relative Risk? (if it was appropriate based on our study design)

$$\text{probability of not visiting a physician} = p = \frac{e^{-1.13-0.001(Age-30)}}{1+e^{-1.13-0.001(Age-30)}}$$

- For a 40-year-old:

$$p = \frac{e^{-1.13-0.001(40-30)}}{1+e^{-1.13-0.001(40-30)}} = \frac{e^{-1.13-0.01}}{1+e^{-1.13-0.01}} \frac{e^{-1.14}}{1+e^{-1.14}} = 0.2423$$

- For a 30-year-old:

$$p = \frac{e^{-1.13-0.001(0)}}{1+e^{-1.13-0.001(0)}} = \frac{e^{-1.13}}{1+e^{-1.13}} = 0.2442$$

- The relative risk (RR) is

$$p_1/p_2 = 0.2423/0.2442 = 0.992$$

Model 1: Probabilities and Relative Risk for 10 year diff

- $\frac{e^{\beta_0}}{1+e^{\beta_0}}$ is the probability of not visiting a physician for 30-year-olds
- $\frac{e^{\beta_0+\beta_1\times10}}{1+e^{\beta_0+\beta_1\times10}}$ is the probability of not visiting a physician for 40-year-olds
- $\frac{e^{\beta_0+\beta_1\times10}}{1+e^{\beta_0+\beta_1\times10}} \Big/ \frac{e^{\beta_0}}{1+e^{\beta_0}}$ is the relative risk of not visiting a physician for 40-year-olds vs. 30-year-olds

Remember those Goals?

- Predict Phys (no physician visit within the past two years=1) with Age (continuous)
- **After adjusting for age, is taking a multivitamin (1=yes) a statistically significant predictor for not regularly visiting a physician?**
- Is taking a multivitamin a confounder for the age-physician visit relationship?

Nested models

- Adding a single new variable to the model
  - Model 1: $\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1(Age-30)$
  - Model 2: $\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1(Age-30)+\beta_2(\text{Multivitamin})$

Logistic regression: Comparing nested models that differ by one variable

- Compare models with p-value or CI
  - What method is this?
    - The Wald test, a test that applies the CLT, like
      - Z test comparing proportions in 2x2 table
      - X^2 test for independence in 2x2 table
      - analogous to the t test for linear regression
  - H_0: the new variable is not needed

    Or, equivalently

    H_0: $\beta_{new}=0$ in the population

Model 2: Results

```
Logit estimates                                   Number of obs   =        317
                                                  LR chi2(2)      =       7.87
                                                  Prob > chi2     =     0.0195
Log likelihood = -171.80997                       Pseudo R2       =     0.0224

------------------------------------------------------------------------------
     phys_no |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        agec |   .0012855   .0192619     0.07   0.947    -.0364671    .0390381
    multivit |  -.7808889   .2871247    -2.72   0.007    -1.343643   -.2181349
 (Intercept) |  -.8571962    .159519    -5.37   0.000    -1.169848   -.5445446
------------------------------------------------------------------------------
```

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(Age-30) + \beta_2(\text{Multivitamin})$$

$$\Rightarrow \quad \log\left(\frac{p}{1-p}\right) = -0.86 + 0.001(Age-30) - 0.78(\text{Multivitamin})$$

Conclusion from the Wald test

- The p-value for multivitamin is 0.007 (<0.05) and the CI for coefficient multivitamin does not include 0 (CI for OR doesn't include 1)
- Reject H_0
- Conclude that the larger model is better: after adjusting for age, multivitamin use is still an important predictor of physician visits in the population

Model 2: Coefficient interpretation on the log odds scale

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(Age-30) + \beta_2(\text{Multivitamin})$$

$$\Rightarrow \quad \log\left(\frac{p}{1-p}\right) = -0.86 + 0.001(Age-30) - 0.78(\text{Multivitamin})$$

- β_0: the **log odds** of not visiting a physician for a 30-year-old person who reports not regularly taking multivitamins
- β_1: the **log odds ratio** of not visiting a physician for a one year increase in age controlling for multivitamin use
- β_2: the **log odds ratio** of not visiting a physician for those who take multivitamins compared with those who do not, adjusting for age

Model 2: Interpretation – odds and odds ratio scale

- $\exp(\beta_0)$: the **odds** of not visiting a physician for a 30-year-old person who reports not regularly taking multivitamins

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(Age-30) + \beta_2(\text{Multivitamin})$$

$$\Rightarrow \quad \log\left(\frac{p}{1-p}\right) = -0.86 + 0.001(Age-30) - 0.78(\text{Multivitamin})$$

Model 2: Interpretation – odds and odds ratio scale

- $\exp(\beta_1)$: after adjusting for multivitamin use, the **odds ratio** of not visiting a physician changes by a ***factor*** of $\exp(\beta_1)=1.001$ ***for each additional year of age***
  - additional age is associated with lower frequency of physician visits in these students, but the association is not statistically significant ($p>0.05$)

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1(Age-30)+\beta_2(\text{Multivitamin})$$

$$\Rightarrow \quad \log\left(\frac{p}{1-p}\right)=\text{-}0.86+0.001(Age-30)-0.78(\text{Multivitamin})$$

Model 2: Interpretation – odds and odds ratio scale

- $\exp(\beta_2)$: the **odds ratio** of not visiting a physician for those who take multivitamins compared with those who do not is $\exp(\beta_2)=0.46$, adjusting for age
  - taking multivitamins is associated with regular physician visits ($p=0.007$)

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1(Age-30)+\beta_2(\text{Multivitamin})$$

$$\Rightarrow \quad \log\left(\frac{p}{1-p}\right)=\text{-}0.86+0.001(Age-30)-0.78(\text{Multivitamin})$$

Goals

- Predict Phys (no physician visit within the past two years=1) with Age (continuous)
- After adjusting for age, is taking a multivitamin (1=yes) a statistically significant predictor for not regularly visiting a physician?
- **Is taking a multivitamin a confounder for the age-physician visit relationship?**

Was multivitamin use a confounder?

- CI for β_1 in model 1: (-0.036, 0.034)
  - Estimate for β_1 in model 2: 0.001
- CI for $\exp(\beta_1)$ in model 1: (exp(-0.036), exp(0.034)) → (0.97, 1.03)
  - Estimate for $\exp\{\beta_1\}$ in model 2: exp(0.001) = 1.001
- Estimate from model 2 is in original CI: **multivitamin use is not a statistically significant confounder**

Interpretation of lack of confounding result

- The factor by which the odds of irregular physician visits changes for each additional year of age does not change appreciably when we adjust for multivitamin use

- The "slope" is roughly the same before and after adjusting for multivitamin use.

Goals: conclusion 1

- Predict Phys (no physician visit within the past two years=1) with Age (continuous)
  - There is no statistically significant effect of age on physician visits in the population

Goals: conclusion 2

- After adjusting for age, is taking a multivitamin (1=yes) a statistically significant predictor for not regularly visiting a physician?
  - After adjusting for age, those who regularly take a multivitamin are also more likely to have visited a physician during the past two years (p=0.007)

Goals: conclusion 3

- Is taking a multivitamin a confounder for the age-physician visit relationship?
  - The effect of age on physician visit is still nonsignificant after adjusting for multivitamin use and multivitamin use is not a confounder

Summary of Lecture 14

- Logistic regression interpretation
  - Intercept – log odds when all X's are 0
  - Slope
    - difference in log odds for a 1 unit increase in X, controlling for other X's
    - log odds ratio associated with a 1 unit increase in X, controlling for other X's
  - Transform log odds/ log odds ratio to odds/odds ratio scale by exponentiating
    - For a continuous X, e^{β} is the factor by which the odds changes (or odds ratio) for each unit change of X
  - Can also transform from log odds to probability
- Nested models in Logistic regression that differ by one variable
  - Use the Wald test (z-test) for the new variable