Lecture 16: Logistic regression diagnostics, splines and interactions

Sandy Eckel
seckel@jhsph.edu

19 May 2007

Logistic Regression Diagnostics
Graphs to check assumptions

- **Recall:** Graphing was used to check the assumptions of linear regression
- Graphing binary outcomes for logistic regression is not as straightforward as graphing a continuous outcome for linear regression
- Several methods have been developed to visualize the logistic regression model for use in checking the assumptions
  - Tables
  - Graphs with lowess curves

Nepali breastfeeding study
Example: data

- Breastfeeding tends to be protective for numerous infant health risks
- A study was conducted in Nepal to evaluate the odds of breastfeeding using a number of possible factors
- **Outcome**: breastfeeding (1=yes, 0=no)
- **Primary predictor**: baby's gender (1=F, 0=M)
- **Secondary predictors**:
  - Child's age (0 to 76 months)
  - Mother's age (17 to 52)
  - Number of children (parity) (1 to 14)

How to look at the data? Binary Y and Binary (or categorical) X

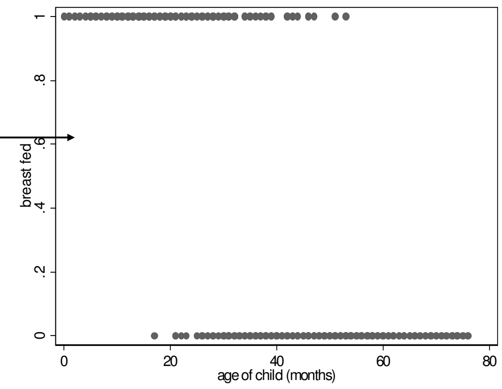
- Breastfeeding vs. baby's gender
  - both binary
  - make a **table**!
  - This method would work for any binary or categorical predictor

How to look at the data? Binary Y and Continuous X

- Breastfeeding vs. child's age
  - Breastfeeding is binary
  - Child's age is continuous
- Could make child's age categorical or binary by
  - breaking it at the quartiles
  - defining groups by years
    e.g. <1 year, 1 year, 2-3 years, 4+ years
  - then use tables
- Or, we could ***graph*** the relationship

How to look at the data? Binary Y and Continuous X
A scatter plot

Actual breastfeeding



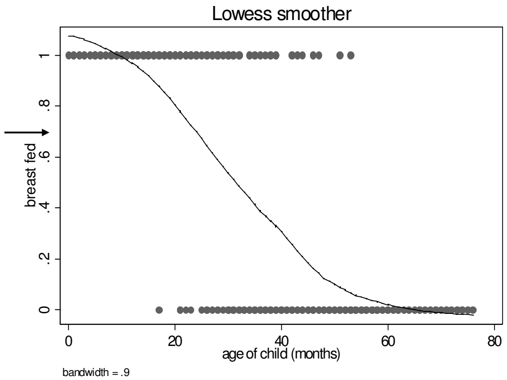
This isn't very informative…how can we fix this?

How to look at the data? Binary Y and Continuous X
Allow a smoothed relationship

- The "lowess" command is a smoothed graph
- It's like a window has been pulled across the graph
  - at each moment, the probability of a 1 within the window is graphed
  - as the window moves, the probability of a 1 is shown as a line
  - changing the width of the window yields different levels of smoothing

How to look at the data? Binary Y and Continuous X
A scatter plot with **Lowess curve**

Probability of breastfeeding



Much more informative! Now we can talk about how the probability of breastfeeding changes with child's age

- We want this to look like a nice 'logistic' curve

Checking form of the model

- Lowess allows us to visualize how the probability of our outcome varies by a certain predictor
- We really want to graph log[p/(1-p)], because that function is assumed to be linear in logistic regression
  - Get the lowess smooth of the probability and then you can transform the smoothed probability to the log odds scale
  - Plot the `smoothed' log odds versus the continuous covariate of interest
  - This relation should look linear
- By looking at lowess plots within key subgroups, we can detect whether the relationship varies across covariates
- Looking at these plots helps us decide if interactions or splines are needed in the model

Assumptions of logistic regression

Two assumptions:

- L – the model fits the data
- I – the observations are all independent

- Independence still cannot be assessed graphically; must know how the data were collected

How can we assess our model ?
L – the model fits the data

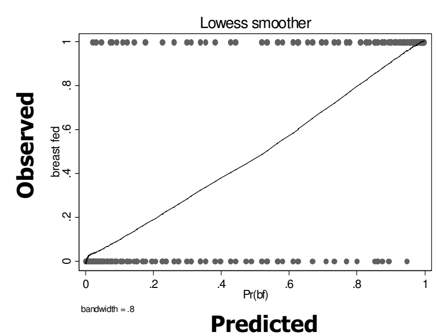
3 methods for assessing model fit

- "Look" at the data
  - Binary or categorical predictors: **tables**
    - Do you see a need for interaction?
  - Continuous predictors: **lowess curves**
    - Do you see a need for interaction or splines?
- Graph observed probability vs. the predicted probability

- Use the X^2 Test of Goodness of Fit to assess the predicted probabilities

Assess model fit : Method 2
Graphing observed vs. predicted probabilities

- Run the model
- Save the predicted probability of breastfeeding for each child
- Plot observed vs predicted probabilities



If the relationship is close to a straight line

- the predicted and observed probabilities are almost the same
- the model fits the data very well

If not, try to add more X's, splines or interactions

Assess model fit : Method 3
X^2 Test of Goodness of Fit

- Run the model

X^2 Test of Goodness of Fit

- Breaks data groups of equal size
- Compares observed and predicted numbers of observations in each group with a X^2 test
  (also called the Hosmer-Lemeshow X^2 Test)

- H_0: the model fits the observed data well
  - We ***want*** $p>0.05$ so we don't reject H_0

Method 3:
X^2 Test of Goodness of Fit

- $p = 0.20 > \alpha = 0.05$

- Fail to reject H_0; conclude that the model fits the data reasonably well

- Conclusion matches the other methods
  - Scatter plots showed same relationship as model
  - the observed and predicted probabilities matched
    - method 2: straight line
  - the observed and predicted data matched
    - method 3: $p>0.05$

Summary: logistic regression model diagnostics

- There are no easy graphs for looking at binary outcome data
  - use lowess
  - split according to binary/categorical covariates to see how relationship between outcome and primary predictor varies
- Assessing model fit: 3 methods
  - look at tables and graphs
  - compare graph of observed vs. predicted p
  - X^2 Test of Goodness of Fit: want large p-value

How do we add
Flexibility in logistic regression?

Same methods as in linear regression!

- Splines
  - are used to allow the "line" to bend

- Interaction
  - is used to allow different effects (difference in log odds ratio) for different groups

Example: Back to breastfeeding example

- **Outcome**: breastfeeding (1=yes, 0=no)
- **Primary predictor**: gender (1=F, 0=M)
- **Secondary predictors**:
  - Child's age (0 to 76 months)
  - Mother's age (17 to 52) – *need to center*
  - Number of children (parity) (1 to 14) – *need to center*

Model A: gender

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1(Gender) \quad \Rightarrow \quad \log\left(\frac{p}{1-p}\right)=\text{-}0.37+0.04(Gender)$$

```
Logit estimates                                   Number of obs   =        472
                                                  LR chi2(1)      =       0.04
                                                  Prob > chi2     =     0.8352
Log likelihood = -319.98468                       Pseudo R2       =     0.0001

------------------------------------------------------------------------------
          bf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |   .0389756   .1873558     0.21   0.835    -.3282351    .4061863
 (Intercept) |  -.3692173   .1281411    -2.88   0.004    -.6203693   -.1180653
------------------------------------------------------------------------------
```

baby's gender (1=F, 0=M)

Model B: gender and mother's age

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1(Gender)+\beta_2(Age_{mom}-25)$$

$$\Rightarrow \quad \log\left(\frac{p}{1-p}\right)=\text{-}0.16+0.06(Gender)+\text{-}0.06(Age_{mom}-25)$$

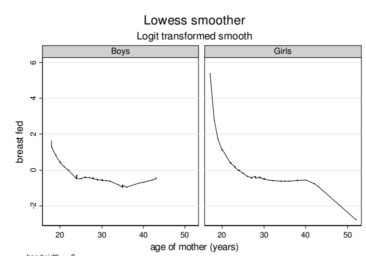
```
Logit estimates                                   Number of obs   =        472
                                                  LR chi2(2)      =      16.50
                                                  Prob > chi2     =     0.0003
Log likelihood = -311.75482                       Pseudo R2       =     0.0258

------------------------------------------------------------------------------
          bf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |   .0620916   .1907094     0.33   0.745     -.311692    .4358751
     age_momc |  -.0615396   .0156442    -3.93   0.000    -.0922016   -.0308776
 (Intercept) |  -.1573215     .13957    -1.13   0.260    -.4308736    .1162307
------------------------------------------------------------------------------
```

baby's gender (1=F, 0=M)

Possible modification – add a spline

- A plot of the log odds of the lowess smooth of breastfeeding versus mother's age reveals
  - There may be a bend in the line at approximately mother's age = 25
  - We'll add a spline for mother's age>25



Possible modification – add a spline

- For mother's age > 25
  - we center mother's age at 25 also, for convenience
  - The spline is a new variable:

$$(age_{mom} - 25)^+$$
$$= 0 \qquad \text{if age} < 25$$
$$= (age_{mom} - 25) \qquad \text{if age} > 25$$

Model C: gender and mother's age with spline

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(Gender) + \beta_2(Age_{mom} - 25) + \beta_3(Age_{mom} - 25)^+$$

$$\Rightarrow \log\left(\frac{p}{1-p}\right) = -0.55 + 0.08(Gender) + -0.25(Age_{mom} - 25) + 0.23(Age_{mom} - 25)^+$$

```
Logit estimates                                   Number of obs   =        472
                                                  LR chi2(3)      =      26.49
                                                  Prob > chi2     =     0.0000
Log likelihood = -306.76341                       Pseudo R2       =     0.0414

------------------------------------------------------------------------------
          bf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |   .0821887   .1928521     0.43   0.670    -.2957946    .4601719
    age_momc |  -.2467804   .0627557    -3.93   0.000    -.3697794   -.1237814
  age_mom_sp |   .2306511    .074613     3.09   0.002     .0844122    .3768899
 (Intercept) |  -.5487527   .1888302    -2.91   0.004    -.9188531   -.1786522
------------------------------------------------------------------------------
```

baby's gender (1=F, 0=M)

Understanding the equation
Write separate equations by age group

log(odds) = -0.55 + 0.08(Gender)– 0.25(Age-25) + 0.23(Age-25)$^+$

- For those with mothers under 25

  -0.55 + 0.08(Gender) – 0.25(Age-25)

- For those with mothers over 25

  -0.55+0.08(Gender)–0.25(Age-25) + 0.23(Age-25)
  = -0.55 + 0.08(Gender)+(-0.25+0.23)(Age-25)
  = -0.55 + 0.08(Gender)+ -0.02 (Age-25)

Model C: Interpretation

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(Gender) + \beta_2(Age_{mom} - 25) + \beta_3(Age_{mom} - 25)^+$$

- β_0: The log odds of breastfeeding for boys with 25-year-old mothers is -0.55

baby's gender (1=F, 0=M)

- β_1: Adjusting for mother's age, the log odds ratio of breastfeeding for girls vs. boys is 0.08
- β_2: Adjusting for gender, the log odds ratio of breastfeeding corresponding to a one year difference in mother's age for mothers **under 25** years is -0.25

Model C: Interpretation

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1(Gender)+\beta_2(Age_{mom}-25)\ +\beta_3(Age_{mom}-25)^+$$

- $\beta_2+\beta_3$: Adjusting for gender, the log odds ratio of breastfeeding corresponding to a one year difference in mother's age for mothers **over 25 years** is -0.25 + 0.23
- β_3: Adjusting for gender, the difference in the log odds ratio of breastfeeding corresponding to a one year difference in mother's age for mothers over 25 years compared with mothers under 25 years is 0.23

Tough both to put in words and to understand, can be easier to understand mathematically!

Model C: Is the difference in the log odds ratio for mother's age statistically significant?

$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1(Gender)+\beta_2(Age_{mom}-25)\ +\beta_3(Age_{mom}-25)^+$$

- H_0: β_3 = 0 in the population
  - i.e., the change in slope is 0, and the line does not bend in the population
- One variable added: use the Wald test
- Z=3.09, p=0.002, CI for β_3 = (0.08, 0.38)
- Reject H_0
- Conclude that Model C is better than Model B

Breastfeeding example conclusion

- For boys and girls with mothers under 25 years of age, the odds that the mother will breastfeed the child decreases by a factor of $\exp(\beta_2)=\exp(-.24)=0.78$ for each additional year of mother's age (95% CI: 0.69, 0.88)
- This relationship is significantly different for boys and girls with mothers over 25 years of age:
  - for these children, the odds that the mother will breastfeed the child is approximately the same for each year of mother's age; the odds decreases by a factor of only $\exp(\beta_2+\beta_3)=0.98$ for each additional year of mother's age (95% CI: 0.95, 1.02)

Model D: gender and number of children (parity)

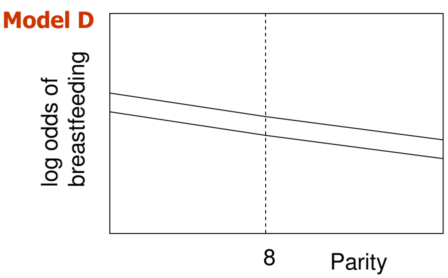
```
Logit estimates                                   Number of obs   =        472
                                                  LR chi2(2)      =       9.99
                                                  Prob > chi2     =     0.0068
Log likelihood = -315.01027                       Pseudo R2       =     0.0156

------------------------------------------------------------------------------
          bf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |   .0622939   .1894771     0.33   0.742    -.3090744    .4336622
     parityc |  -.1180777   .0384221    -3.07   0.002    -.1933837   -.0427718
 (Intercept) |  -.8009664   .1937284    -4.13   0.000    -1.180667   -.4212659
------------------------------------------------------------------------------
```

Sketch of Model D

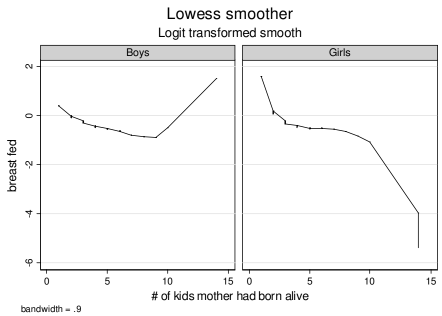
$$\log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1(Gender)$$
$$+\beta_2(Parity-8)$$

baby's gender (1=F, 0=M)



Asessing the relationship in the data

- The relationship between logit(bf) and parity is very different for boys and girls
  - Mothers of **more** children tend to
    - breastfeed boys more
    - breastfeed girls less



- The relationship is about the same for boys and girls whose mothers have about 8 or fewer kids
  - Could add a spline and an interaction term for only parity > 8 so that the slopes only differ then
  - First we'll just add a spline

Model E: gender, parity, and parity spline

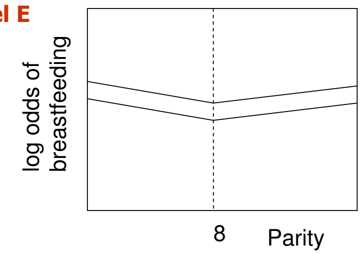
```
Logit estimates                                   Number of obs   =        472
                                                  LR chi2(3)      =      14.18
                                                  Prob > chi2     =     0.0027
Log likelihood = -312.91444                       Pseudo R2       =     0.0222

------------------------------------------------------------------------------
          bf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |   .0666432   .1903717     0.35   0.726    -.3064785     .439765
     parityc |  -.1718923   .0465719    -3.69   0.000    -.2631716    -.080613
   parity_sp |   .3281222   .1562619     2.10   0.036     .0218545    .6343899
 (Intercept) |  -1.045415   .2291123    -4.56   0.000    -1.494466   -.5963627
------------------------------------------------------------------------------
```

Sketch of Model E

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(Gender) + \beta_2(Parity-8) + \beta_3(Parity-8)^{+}$$

baby's gender (1=F, 0=M)



Understanding the equation
Write separate equations by parity group

log(odds) = -1.05 + 0.07(Gender) – 0.17(Parity-8) + 0.33(Parity-8)$^{+}$

- For those with mothers with less than 8 children

  -1.05 + 0.07(Gender) – 0.17(Parity-8)

- For those with mothers with at least 8 children

  -1.05 + 0.07(Gender) – 0.17(Parity-8) + 0.33(Parity-8)
  = -1.05 + 0.07(Gender) + (-0.17+0.33)(Parity-8)
  = -1.05 + 0.07(Gender) + 0.16(Parity-8)

Problem with the parity spline

- Model E forces the "slope" to be the same for boys and girls
- The lowess curve suggests slope should differ for boys and girls whose mothers had more than around 8 children
  - Add an interaction term between the spline and gender
  - that allows the slope to differ by gender only for those whose mothers have 8 or more children

The new variable

- Gender = 0 for boys
- $(Parity - 8)^+$ = 0 for children of low parity families

- $(Gender) \times (Parity - 8)^+$
  - = 0 for boys
  - = 0 for parity < 8
  - = (Parity – 8) for girls with parity >=8

baby's gender (1=F, 0=M)

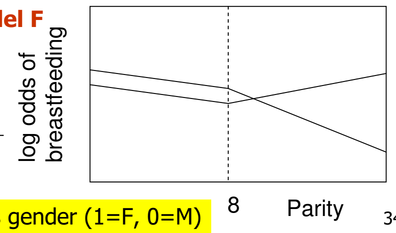
Model F: spline + interaction with spline

```
Logit estimates                                   Number of obs   =        472
                                                  LR chi2(4)      =      21.75
                                                  Prob > chi2     =     0.0002
Log likelihood = -309.12925                       Pseudo R2       =     0.0340

------------------------------------------------------------------------------
          bf |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |   .1806766   .1953877     0.92   0.355    -.2022763    .5636294
     parityc |  -.1737844   .0473172    -3.67   0.000    -.2665244   -.0810445
   parity_sp |    .734593   .2786475     2.64   0.008     .1884539    1.280732
parity_sp_~r |  -.8665087   .3966433    -2.18   0.029    -1.643915   -.0891021
 (Intercept) |  -1.106983   .2343301    -4.72   0.000    -1.566261    -.647704
------------------------------------------------------------------------------
```

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(Gender)$$
$$+\beta_2(Parity-8) + \beta_3(Parity-8)^+$$
$$+\beta_4 Gender \times (Parity-8)^+$$

Sketch of Model F



baby's gender (1=F, 0=M)

Understanding the equation
Write separate equations by parity and gender

log(odds) = -1.11 + 0.18(Gender) – 0.17(Parity-8) + 0.73(Parity-8)$^+$
- 0.87(Gender)x(Parity-8)$^+$

baby's gender (1=F, 0=M)

- For those with mothers with less than 8 children
  -1.11 + 0.18(Gender) – 0.17(Parity-8)

- For ***boys*** with mothers with at least 8 children
  -1.11 + 0.18(Gender) – 0.17(Parity-8) + 0.73(Parity-8)
  = -1.11 + (-0.17+0.73)(Parity-8)
- For ***girls*** with mothers with at least 8 children
  -1.11 + 0.18(Gender) – 0.17(Parity-8) + 0.73(Parity-8)
  - 0.87(Gender)x(Parity-8)
  = (-1.11 + 0.18) +(-0.17 + 0.73 - 0.87)(Parity-8)

Interpretation – Model F

- $\exp(\beta_0)$: The odds of breastfeeding for boys of mothers with 8 children is exp(-1.11) = 0.33

- $\exp(\beta_1)$: Adjusting for mother's parity, the odds ratio of breastfeeding for girls vs. boys is 1.20 for children of mothers with less than 8 children

- $\exp(\beta_2)$: Adjusting for gender, the odds ratio of breastfeeding corresponding to a one child difference in parity for mothers with fewer than 8 children is .84

Interpretation – Model F

- $\exp(\beta_2+\beta_3)$: ***Among boys***, the odds ratio of breastfeeding corresponding to a one child difference in parity for mothers with at least 8 children is 1.75
- $\exp(\beta_2+\beta_3+\beta_4)$: ***Among girls***, the odds ratio of breastfeeding corresponding to a one child difference in parity for mothers with at least 8 children is 0.74

Interpretation – Model F

- **Complicated to interpret the components on their own – read on your own if you want!**
- $\exp(\beta_3)$: The odds ratio of breastfeeding corresponding to a one child difference in parity is 2.08 times higher for ***boys whose mothers have at least 8 children*** than for ***boys whose mothers have fewer than 8 children***
- $\exp(\beta_3+\beta_4)$: The odds ratio of breastfeeding corresponding to a one child difference in parity is 0.74 times lower for ***girls whose mothers have at least 8 children*** than for ***girls whose mothers have fewer than 8 children***
- $\exp(\beta_4)$: The odds ratio of breastfeeding corresponding to a one child difference in parity is 0.42 times lower for ***boys whose mothers have at least 8 children*** than for ***girls whose mothers have at least 8 children***

Is the difference in the log odds ratio for parity by gender statistically significant?

- H_0: $\beta_4 = 0$ in the population
  - i.e. the change in slope for parity > 8 is the same for boys and girls in the population
- One variable added: use the Wald test
  - Z=-2.18, p=0.029, CI for $\exp(\beta_3)$ = (0.19, 0.91)
  - Reject H_0
- Conclude that Model F is better than Model E

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1(Gender)$$
$$+\beta_2(Parity-8) \ +\beta_3(Parity-8)^+$$
$$+\beta_4 Gender \times (Parity-8)^+$$

Conclusion – Model F

- For children whose mothers have fewer than 8 children, the odds that the mother will breastfeed the child is about the same for boys and girls and decreases by a factor of $\exp(\beta_2)$=0.84 for each additional year of mother's age (95% CI: 0.77, 0.92).
- This relationship is significantly different for both boys and girls whose mothers have more than 8 children:
  - For ***boys*** whose mothers have more than 8 children, the odds that the mother will breastfeed increases by a factor of $\exp\{\beta_2+\beta_3\}$=1.75 for each additional year of mother's age (95% CI: 1.05, 2.93).
  - For ***girls*** whose mothers have more than 8 children, the odds that the mother will breastfeed decreases by a factor of $\exp\{\beta_2+\beta_3+\beta_4\}$=0.74 for each additional year of mother's age (95% CI: 0.40, 1.37).

Comparing the models

	Odds Ratio for Model					
Variables	A	B	C	D	E	F
Reference*	0.69	0.85	0.58	0.45	0.35	0.33
Gender	1.04	1.06	1.09	1.06	1.07	1.20
Age-25		0.94	0.78			
$(\text{Age-25})^+$			1.26			
Parity – 8				0.89	0.84	0.84
$(\text{Parity-8})^+$					1.39	2.08
$(\text{Gender})_x (\text{Parity-8})^+$						0.42
Deviance	640.0	623.5	613.5	630.0	625.8	618.3

The table value for the reference group is the odds, not the odds ratio

Comparing the models

- Models C and F are both nested in Model A
- Models C and F cannot be directly compared to one another, but we can see which has a smaller p-value when compared to Model A
  - C vs. A: $X^2 = 26.5$ with 2 df
  - F vs. A: $X^2 = 21.7$ with 3 df
    - Both p-values are very small <.0001, but the p-value for model C is slightly smaller

What next?

- Model C improves prediction beyond gender alone (Model A) more than Model F.
- Model C should be the next parent model, and we should test the new variables in Model F to see if they continue to improve prediction within the context of Model C
- When a tentative final model is identified, the assumptions of logistic regression should be checked

Summary of lecture 16

- Logistic regression assumptions
  - L – the model fits the data
  - I – the observations are all independent
- Logistic regression diagnostics
  - "Look" at the data: tables or logits of lowess curves
  - Graph observed probability vs. the predicted probability
  - Use the X^2 Test of Goodness of Fit to assess the predicted probabilities
- Splines and interactions add flexibility to the model
- When comparing nested models, a table of
  - the coefficients and their CI's, or
  - the odds ratios and their CI's

  helps the reader quickly compare models
- Two models ***not*** nested in one another cannot be directly compared
- One can identify a ***new parent model*** by comparing statistical significance