Lecture 13: Introduction to Logistic Regression

Sandy Eckel
seckel@jhsph.edu

13 May 2008

Logistic Regression

Basic Idea:

- **Logistic regression** is the type of regression we use for a response variable (Y) that follows a binomial distribution
- Linear regression is the type of regression we use for a continuous, normally distributed response (Y) variable

- Remember the Binomial Distribution?

Review of the Binomial Model

- $Y \sim \text{Binomial}(n,p)$
- n independent trials
  - (e.g., coin tosses)
- p = probability of success on each trial
  - (e.g., $p = \frac{1}{2}$ = Pr of heads)
- Y = number of successes out of n trials
  - (e.g., Y= number of heads)

Binomial Distribution Example

Binomial probability density function (pdf):

$$P(Y = y) = \binom{n}{y} p^{y} (1 - p)^{n-y}$$

Example:

n=7, p=.5

y	0	1	2	3	4	5	6	7
Pr (Y=y)	.008	.055	.164	.273	.273	.164	.054	.008

Why can't we use Linear Regression to model **binary** responses?

- The response (Y) is NOT normally distributed
- The variability of Y is NOT constant
  - Variance of Y depends on the expected value of Y
  - For a $Y \sim \text{Binomial}(n,p)$ we have $\text{Var}(Y)=pq$ which depends on the expected response, $E(Y)=p$
- The model must produce predicted/fitted probabilities that are between 0 and 1
  - Linear models produce fitted responses that vary from $-\infty$ to ∞

Binomial Y example

- Consider a phase I clinical trial in which 35 independent patients are given a new medication for pain relief. Of the 35 patients, 22 report “significant” relief one hour after medication

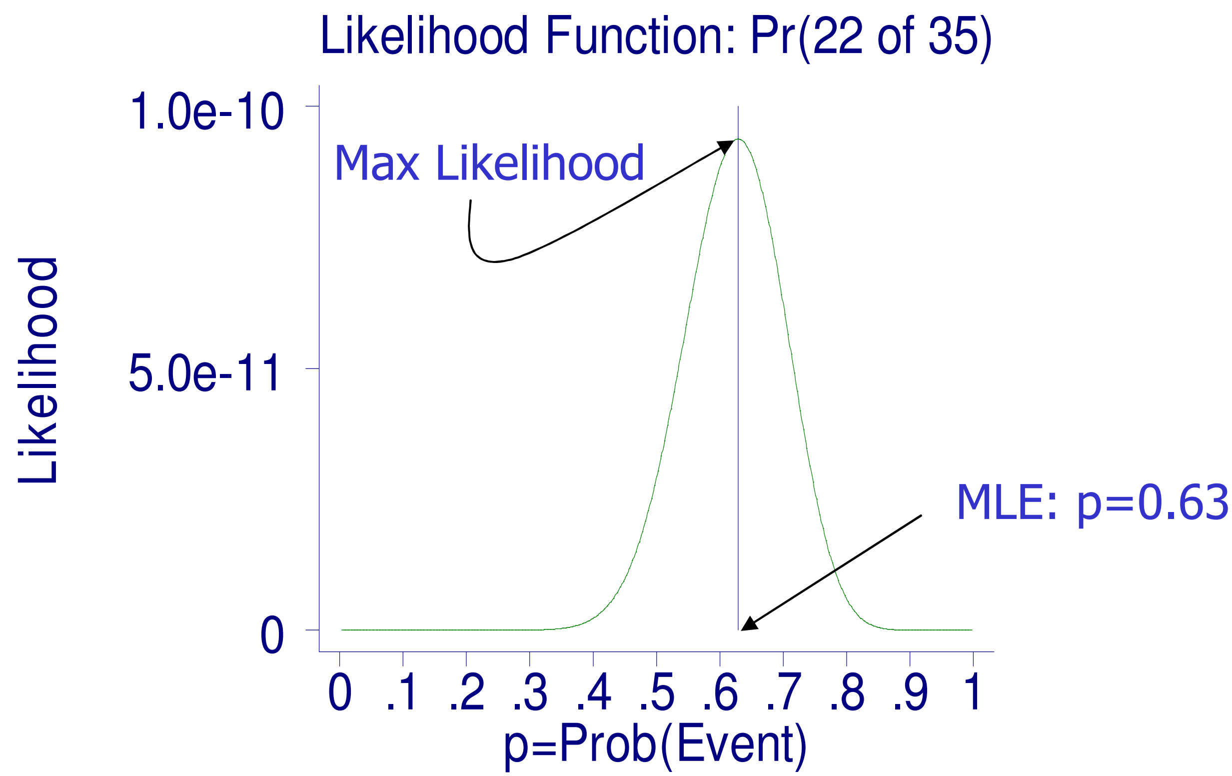
- Question: How effective is the drug?

Model

- Y = # patients who get relief
- n = 35 patients (trials)
- p = probability of relief for any patient
  - The truth we seek in the population
- How effective is the drug? $\Longleftrightarrow$ What is p?
  - Want a method to
    - Get best estimate of p given data
    - Determine range of plausible values for p

How do we estimate p? Maximum Likelihood Method

The method of maximum likelihood estimation chooses values for parameter estimates which make the observed data "maximally likely" under the specified model



Maximum Likelihood
Clinical trial example

- Under the binomial model, 'likelihood' for observed Y=y

$$P(Y=y)=\binom{n}{y}p^{y}(1-p)^{n-y}$$

- So for this example the likelihood function is:

$$P(Y=y)=\binom{35}{22}p^{22}(1-p)^{13}$$

- So, estimate p by choosing the value for p which makes observed data "maximally likely"
  - i.e., choose p that makes the value of *Pr (Y=22)* maximal
- The ML estimate of p is y/n

  = 22/35

  = 0.63

The estimated proportion of patients who will experience relief is 0.63

Confidence Interval (CI) for p

- Recall the general form of any CI:

  Estimate ± (something near 2) x SE(estimate)

- Variance of $\hat{p}$: Var($\hat{p}$)= $\frac{p(1-p)}{n} = \frac{pq}{n}$

- "Standard Error" of $\hat{p}$: $\sqrt{\frac{pq}{n}}$

- Estimate of "Standard Error" of $\hat{p}$: $\sqrt{\frac{\hat{p}\hat{q}}{n}}$

Confidence Interval for p

- 95% Confidence Interval for the 'true' proportion, p:

$$\hat{p} \pm 1.96\sqrt{\frac{\hat{p}\hat{q}}{n}} = 0.63 \pm 1.96\sqrt{\frac{(0.63)(0.37)}{35}}$$

→ LB: 0.63-1.96(.082)

UB: 0.63+1.96(.082)

=(0.47, 0.79)

Conclusion

- Based upon our clinical trial in which 22 of 35 patients experience relief, we estimate that 63% of persons who receive the new drug experience relief within 1 hour (95% CI: 47% to 79%)
- Whether 63% (47% to 79%) represents an ‘effective’ drug will depend many things, especially on the science of the problem.
  - Sore throat pain?
  - Arthritis pain?
  - Childbirth pain?

Aside: Review of Probabilities and Odds

- The odds of an event are defined as:

$$\text{odds}(Y=1) = \frac{P(Y=1)}{P(Y=0)} = \frac{P(Y=1)}{1-P(Y=1)}$$

$$= \frac{p}{1-p}$$

- We can go back and forth between odds and probabilities:

$$\text{Odds} = \frac{p}{1-p}$$

$$p = \text{odds}/(\text{odds}+1)$$

Aside: Review of Odds Ratio

- We saw that an odds ratio (OR) can be helpful for comparisons.
- Recall the Vitamin A trial where we looked at the odds ratio of death comparing the vitamin A group to the no vitamin A group:
- OR = $\dfrac{\text{odds(Death | Vit. A)}}{\text{odds(Death | No Vit A.)}}$

Aside: Review of Odds Ratio Interpretation

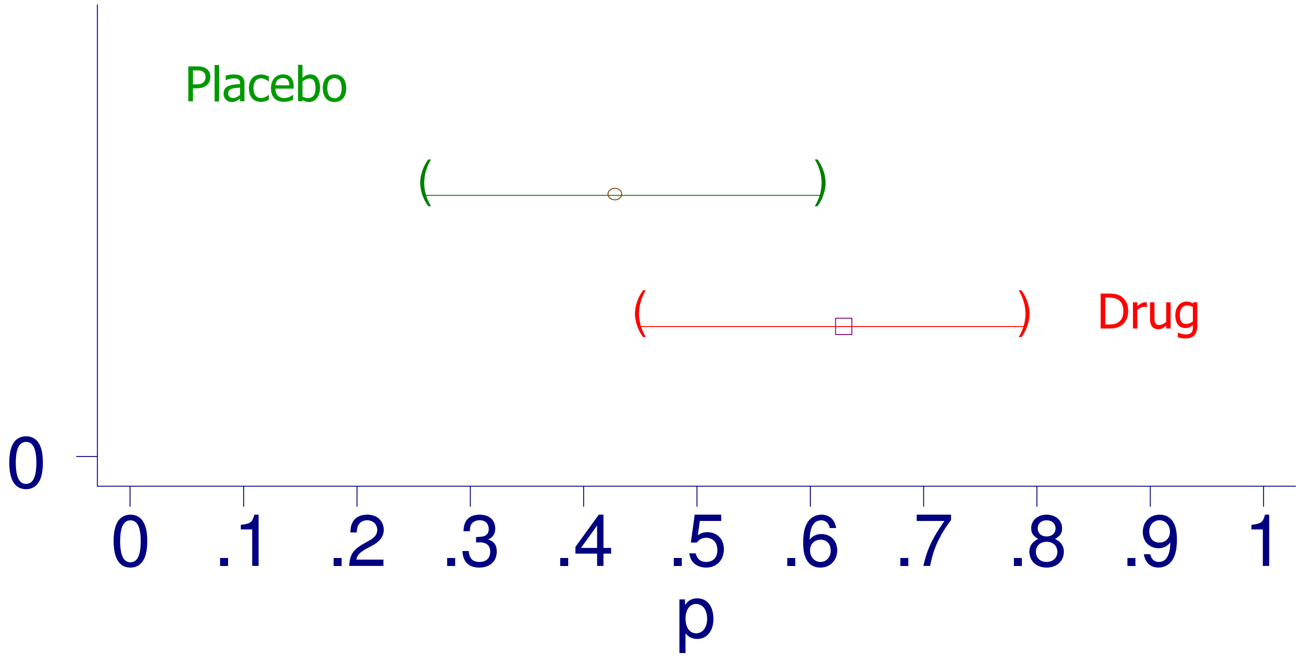
- The OR here describes the benefits of Vitamin A therapy. We saw for this example that:
- OR = 0.59
  - The Vitamin A group had 0.60 times the odds of death of the no Vitamin A group; or
  - An estimated 40% reduction in mortality
- OR is a building block for logistic regression

Logistic Regression

- Suppose we want to ask whether new drug is better than a placebo and have the following observed data:

Relief?	Drug	Placebo
No	13	20
Yes	22	15
Total	35	35

Confidence Intervals for p



Odds Ratio

$$\text{OR} = \frac{\text{odds(Relief} \mid \text{Drug)}}{\text{odds(Relief} \mid \text{Placebo)}}$$

$$= \frac{\text{P(Relief} \mid \text{Drug)} / [1 - \text{P(Relief} \mid \text{Drug)}]}{\text{P(Relief} \mid \text{Placebo)} / [1 - \text{P(Relief} \mid \text{Placebo)}]}$$

$$= \frac{0.63/(1 - 0.63)}{0.45/(1 - 0.45)} = 2.26$$

Confidence Interval for OR

- CI used Woolf's method for the standard error of $\log(\hat{O}R)$ (from lecture 6)
- se($\log(\hat{O}R)$) = $\sqrt{\frac{1}{22}+\frac{1}{13}+\frac{1}{15}+\frac{1}{20}}=0.489$
- find $\log(\hat{O}R)\pm 1.96 se(\log(\hat{O}R))$
- Then (e^L,e^U)

Interpretation

- OR = 2.26
- 95% CI: (0.86 , 5.90)

- The Drug is an estimated 2 ¼ times better than the placebo.
- But could the difference be due to chance alone?
  - YES ! 1 is a 'plausible' true population OR

Logistic Regression

- Can we set up a model for this binomial outcome similar to what we’ve done in regression?

- Idea: model the log odds of the event, (in this example, relief) as a function of predictor variables

A regression model for the log odds

$$\log[\text{odds(Relief} \mid \text{Tx)}] = \log\left(\frac{\text{P(relief} \mid \text{Tx)}}{\text{P(no relief} \mid \text{Tx)}}\right) = \beta_0 + \beta_1 Tx$$

where: $\text{Tx} = \begin{cases} 0 \text{ if Placebo} \\ 1 \text{ if Drug} \end{cases}$

- log(odds(Relief|Drug)) = $\beta_0 + \beta_1$
- log(odds(Relief|Placebo)) = β_0
- log(odds(Relief|D)) – log(odds(Relief|P)) = β_1

And…

- Because of the basic property of logs:

  $\log(\text{ odds(Relief|D)}) - \log(\text{ odds(Relief|P)}) = \beta_1$

$\rightarrow$ $\log\left(\frac{\text{odds}(R \mid D)}{\text{odds}(R \mid P)}\right) = \beta_1$

- And: OR $= \exp(\beta_1) = e^{\beta 1}$!!

- So: $\exp(\beta_1)$ = odds ratio of relief for patients taking the Drug-vs-patients taking the Placebo.

Logistic Regression

```
Logit estimates                                   Number of obs   =         70
                                                  LR chi2(1)      =       2.83
                                                  Prob > chi2     =     0.0926
Log likelihood =  -46.99169                       Pseudo R2       =     0.0292

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        Tx   |   .8137752    .4889211     1.66    0.096    -.1444926    1.772043
 (Intercept) |  -.2876821     .341565    -0.84    0.400    -.9571372    .3817731
------------------------------------------------------------------------------
```

Estimates:

log(odds(relief|Tx)) $= \hat{\beta}_0 + \hat{\beta}_1 Tx$

$= -0.288 + 0.814(Tx)$

Therefore: OR = exp(0.814) = 2.26 !
So 2.26 is the odds ratio of relief for patients taking the Drug compared to patients taking the Placebo

It's the same as the OR we got before!

- So, why go to all the trouble of setting up a linear model?
- What if there is a biologic reason to expect that the rate of relief (and perhaps drug efficacy) is age dependent?
- What if

  Pr(relief) = function of Drug or Placebo AND Age
- We could easily include age in a model such as:

$$\log(\text{ odds(relief) }) = \beta_0 + \beta_1\text{Drug} + \beta_2\text{Age}$$

Logistic Regression

- As in MLR, we can include many additional covariates
- For a Logistic Regression model with r number of predictors:

$$\log\ (\ \text{odds}(Y=1)) = \beta_0 + \beta_1 X_1 + \ldots + \beta_r X_r$$

$$\text{where: } \text{odds}(Y=1) = \frac{\Pr(Y=1)}{1-\Pr(Y=1)} = \frac{\Pr(Y=1)}{\Pr(Y=0)}$$

Logistic Regression

Thus:

$$\log \left(\frac{\Pr(Y=1)}{\Pr(Y=0)} \right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_r X_r$$

- But, why use log(odds)?
- Linear regression might estimate *anything* (-∞, +∞), not just a proportion in the range of 0 to 1
- ***Logistic regression*** is a way to estimate a proportion (between 0 and 1) as well as some related items

Another way to motivate using log(OR) for the lefthand side of logistic regression

- We would like to use something like what we know from linear regression:

$$\text{Continuous outcome} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots$$

- How can we turn a proportion into a continuous outcome?

Transforming a proportion…

- A proportion is a value between 0 and 1
- The ***odds*** are always positive:

$$\text{odds} = \left(\frac{p}{1-p}\right) \Rightarrow [0, +\infty)$$

- The ***log odds*** is continuous:

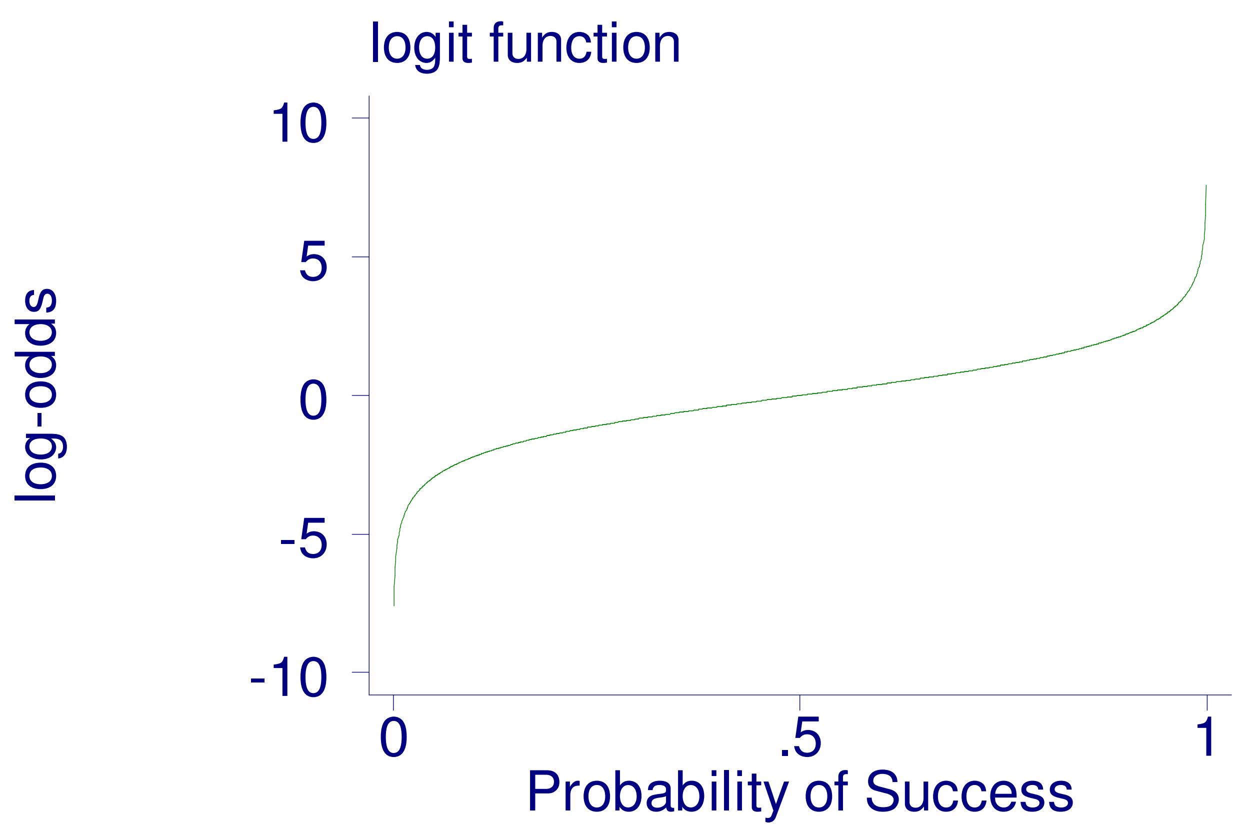
$$\text{Logodds} = \ln\left(\frac{p}{1-p}\right) \Rightarrow (-\infty, +\infty)$$

“Logit” transformation of the probability

Measure	Min	Max	Name
Pr(Y = 1)	0	1	“probability”
$\frac{\Pr(Y=1)}{1-\Pr(Y=1)}$	0	∞	“odds”
$\log\left(\frac{\Pr(Y=1)}{1-\Pr(Y=1)}\right)$	-∞	∞	“log-odds” or “logit”

Logit Function

- Relates log-odds (logit) to $p =$ Pr(Y=1)



Key Relationships

- Relating log-odds, probabilities, and parameters in logistic regression:
- Suppose we have the model:

$$\text{logit}(p) = \beta_0 + \beta_1 X$$

i.e.

$$\log\left(\frac{p}{1\text{-}p}\right) = \beta_0 + \beta_1 X$$

- Take "anti-logs" to get back to OR scale

$$\left(\frac{p}{1\text{-}p}\right) = \exp(\beta_0 + \beta_1 X)$$

Solve for *p as a function of the coefficients*

- $p/(1-p) = \exp(\beta_0 + \beta_1 X)$
- $p = (1 - p)\cdot\exp(\beta_0 + \beta_1 X)$
- $p = \exp(\beta_0 + \beta_1 X) - p \cdot \exp(\beta_0 + \beta_1 X)$

- $p + p \cdot\exp(\beta_0 + \beta_1 X) \quad = \quad \exp(\beta_0 + \beta_1 X)$
- $p \cdot\{1+ \exp(\beta_0 + \beta_1 X)\} \quad = \quad \exp(\beta_0 + \beta_1 X)$
- $$p = \frac{\exp(\beta_0 + \beta_1 X)}{1+\exp(\beta_0 + \beta_1 X)}$$

What's the point of all that algebra?

- Now we can determine the estimated probability of success for a specific set of covariates, X, after running a logistic regression model

Example
Dependence of Blindness on Age

- The following data concern the Aegean island of Kalytos where inhabitants suffer from a congenital eye disease whose effects become more marked with age.
- Samples of 50 people were taken at five different ages and the numbers of blind people were counted

Example: Data

Age	Number blind / 50
20	6 / 50
35	7 / 50
45	26 / 50
55	37 / 50
70	44 / 50

Question

- The scientific question of interest is to determine how the probability of blindness is related to age in this population

Let p_i = Pr(a person in age class$_i$ is blind)

Model 1 – Intercept only model

- $\text{logit}(p_i) = \beta_0^*$

β_0^*= **log-odds** of blindness for all ages
$\exp(\beta_0^*)$ = **odds** of blindness for all ages

- No age dependence in this model

Model 2 – Intercept and age

$$\text{logit}(p_i) = \beta_0 + \beta_1(\text{age}_i - 45)$$

- β_0 = **log-odds** of blindness among 45 year olds
- $\exp(\beta_0)$ = **odds** of blindness among 45 year olds
- β_1 = difference in **log-odds** of blindness comparing a group that is one year older than another
- $\exp(\beta_1)$ = **odds ratio** of blindness comparing a group that is one year older than another

Results

Model 1: logit(p_i) = β_0*

```
Logit estimates                                   Number of obs   =        250
                                                  LR chi2(0)      =       0.00
                                                  Prob > chi2     =          .
Log likelihood = -173.08674                       Pseudo R2       =     0.0000

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 (Intercept) |  -.0800427   .1265924    -0.63   0.527    -.3281593    .1680739
------------------------------------------------------------------------------
```

- logit($\hat{p}_i$) = -0.08 or $\hat{p}_i = \frac{\exp(-.08)}{1+\exp(-.08)} = 0.48$

Results

Model 2: $\text{logit}(p_i) = \beta_0 + \beta_1(\text{age}_i - 45)$

```
Logit estimates                                   Number of obs   =        250
                                                  LR chi2(1)      =      99.30
                                                  Prob > chi2     =     0.0000
Log likelihood = -123.43444                       Pseudo R2       =     0.2869

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.       z     P>|z|      [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0940683    .0119755     7.86    0.000      .0705967    .1175399
 (Intercept) |  -4.356181    .5700966    -7.64    0.000     -5.473549   -3.238812
------------------------------------------------------------------------------
```

$$\text{logit}(\hat{p}_i) = -4.4 + .094(\text{age}_i - 45)$$

or

$$\hat{p}_i = \frac{\exp(-4.4 + 0.094(age_i - 45))}{1 + \exp(-4.4 + 0.094(age_i - 45))}$$

Test of significance

- Is the addition of the age variable in the model important?
- Maximum likelihood estimates:

  $\hat{\beta}_1$ =0.094 s.e.($\hat{\beta}_1$)=0.012
- **z-test**: H_0: $\beta_1 = 0$
  - z=7.855; p-val=0.000
  - 95% C.I. (0.07, 0.12)

What about the Odds Ratio?

- Maximum likelihood estimates:
- OR = exp($\hat{\beta}_1$)= exp(0.094)= 1.10
- SE($\hat{\beta}_1$) =SE(log(OR)) = 0.013
- **Same z-test, reworded for OR scale**:
  Ho: exp(β_1) = 1
  - z = 7.86 p-val = 0.000
  - 95% C.I. for β_1 (1.07, 1.13)

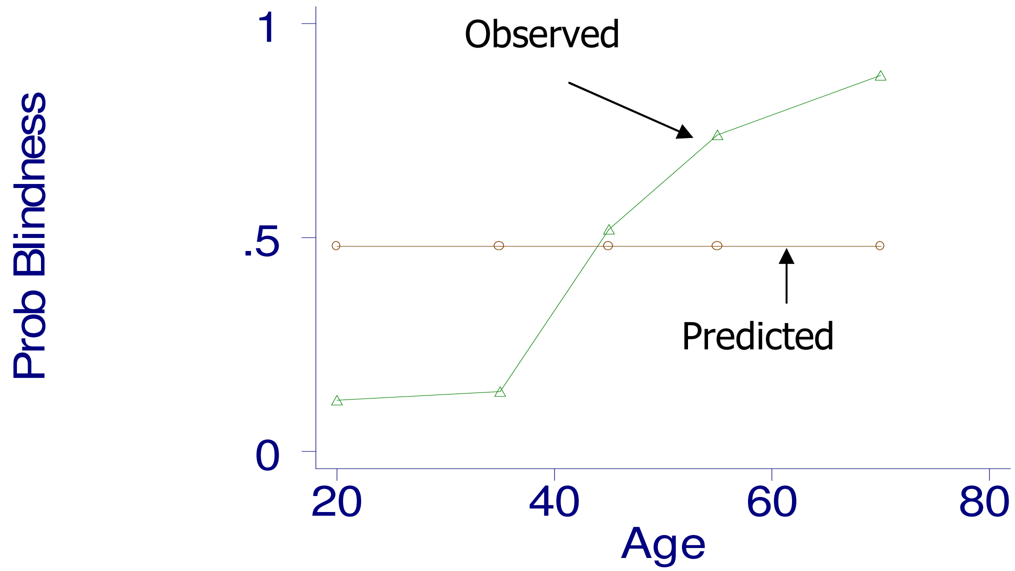
    *(calculated on log scale, then exponentiated!!)

    $e^{(0.094 - 1.96*0.013)}$, $e^{(0.094 + 1.96*0.013)}$

- It appears that blindness is age dependent
- *Note: exp(0) = 1, where is this fact useful?*

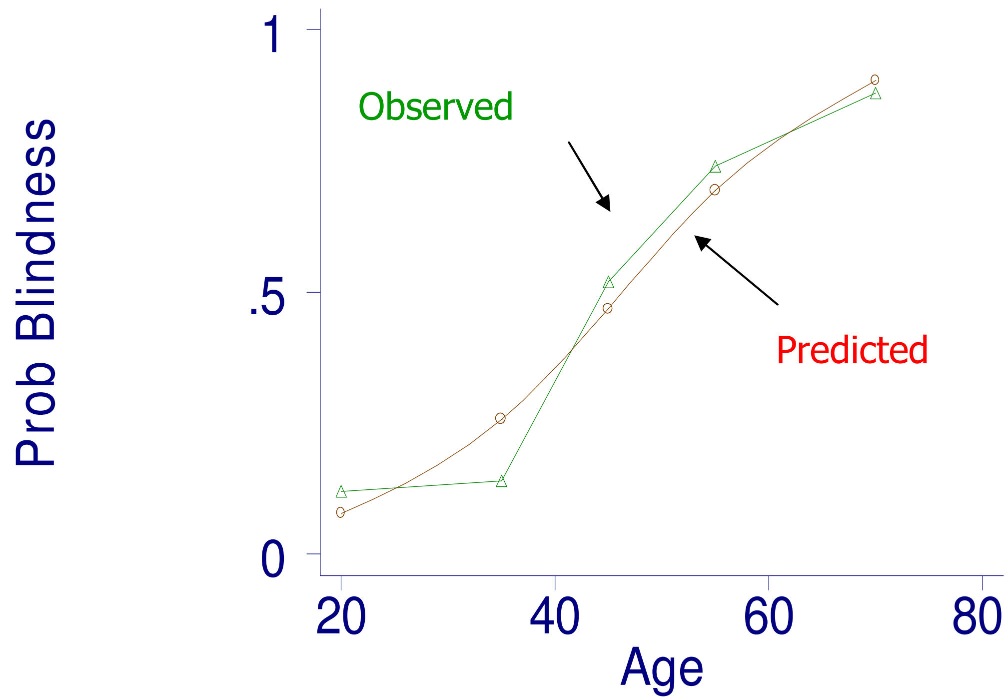
Model 1 fit

- Plot of observed proportion -vs- predicted proportions using an intercept only model



Model 2 fit

- Plot of observed proportion -vs- predicted proportions with age in the model



Conclusion

- Model 2 clearly fits better than Model 1!

- Including age in our model is better than intercept alone.

Lecture 13 Summary

- Logistic regression gives us a framework in which to model binary outcomes
- Uses the structure of linear models, with outcomes modelled as a function of covariates
- As we'll see, many concepts carry over from linear regression
  - Interactions
  - Linear splines
  - Tests of significance for coefficients
- **All coefficients will have different interpretations in logistic regression**
  - **Log odds or Log odds ratios!**

HW 3 Hint

General logistic model specification:

Systematic:

$$\text{logit}(P(Y_i = 1)) = \log(odds(Y_i = 1)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Random:

$$Y_i \sim Binomial(n, p) = Binomial(1, p_i)$$

where p_i depends on the covariates for person i