Logistic Regression

Patrick Breheny

February 21

Introduction

- Binary outcomes are quite common in medicine and public health: alive/dead, diseased/healthy, infected/not infected, case/control
- Assuming that these outcomes follow a normal distribution is clearly wrong; the binomial (Bernoulli) distribution is much more natural
- The binomial distribution is what we use to model things like coin flips, and that's essentially how we're modeling disease/survival/infection here

Setting up the GLM

- The difference, however, is that we are not assuming that each person's coin has the same probability of landing heads
- The probability that each person will die/succumb to disease/get infected varies depending on the explanatory variable in a way that is explicitly modeled by the GLM:
  - Random component: $Y_i \sim \text{Binom}(1, \pi_i)$
  - Systematic component: $\eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$
- What about the link?

The link

- The most natural choice is the canonical link:

$$\eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$$

$$\pi_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

- The first function is called the "logit" of π_i, and the generalized linear model based on this link is called *logistic regression*

The Donner party

- Our example data set from today involves the survival of the members of the Donner party
- In the spring of 1846, a group of American pioneers set out for California
- However, they suffered a series of setbacks and did not arrive at the Sierra Nevada mountains until October
- While crossing the mountains, they became trapped by an early snowfall, and had to spend the winter there

The Donner party (cont'd)

- Conditions were harsh, food supplies ran low, and 40 of the 87 members of the Donner party died before they were finally rescued
- The data set `donner.txt` contains the following information regarding the adult (over 15 years old) members of the Donner party:
  - `Age`
  - `Sex`
  - `Status`: either `Died` or `Survived`

Fitting logistic regression models in SAS

- In SAS, generalized linear models can be fit using `PROC GENMOD`
- Logistic regression models can also be fit using `PROC LOGISTIC` with similar syntax, although some of the default settings and output options are different
- Don't use `PROC GLM`, however, because it doesn't actually fit GLMs (go figure)

Syntax: SAS

- Specifying models in `PROC GENMOD` is very similar to using `PROC REG`, although now we have to specify the distribution of the outcome:

```
PROC GENMOD DATA=donner;
  CLASS Status Sex;
  MODEL Status = Sex|Age / DIST=Binomial;
RUN;
```

- In principle, we could specify the link as well, although by default SAS (and `R`) use the canonical link

Syntax: R

- The output, in estimate/standard error/test statistic/p-value format, is quite similar to what we're used to seeing from linear regression, although recall that these numbers are only approximate
- In R, the glm function accomplishes the same thing:

```
fit <- glm(Status~Age*Sex,donner,family=binomial)
summary(fit)
```

Fitting the model

- To ensure that we all understand where these numbers are coming from, let's calculate them directly
- First, we note that for the binomial distribution, $W(\mu_i) = \mu_i(1 - \mu_i)$
- Thus, $\mathbf{W}$ is a diagonal matrix with ith entry $\pi_i(1 - \pi_i)$

Fitting the model (cont'd)

We can then implement the IRLS algorithm as discussed in Tuesday's lecture:

```
repeat {
  old <- b
  eta <- X%*%b
  pi <- exp(eta)/(1+exp(eta))
  W <- diag(as.numeric(pi*(1-pi)))
  z <- eta + solve(W)%*%(y-pi)
  b <- solve(t(X)%*%W%*%X)%*%t(X)%*%W%*%z
  if (converged(b,old)) break
}
```

Inferential results

- We can obtain the inferential results via

```
VarB <- solve(t(X)%*%W%*%X)
SE <- sqrt(diag(VarB))
z <- b/SE
p <- 2*pnorm(-abs(z))
```

- As one would hope, we get all of the same results whether we use SAS, R, or calculate the results ourselves

Standard results

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.3183	1.1310	0.2815	0.7784
Age	-0.0325	0.0353	-0.9209	0.3571
Female	6.9280	3.3989	2.0383	0.0415
Age:Female	-0.1616	0.0943	-1.7143	0.0865

Interpreting the coefficients

- Now let's talk a little more about what these results mean
- The model coefficients are somewhat difficult to interpret directly, as (a) we've fit a model with an interaction, and (b) they must go through the link function before they are on the correct scale
- To help gain some familiarity with what the model coefficients mean, let's calculate survival probabilities for various kinds of people

Males

- Let's write out the model as:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \texttt{Age} + \beta_2 \texttt{Female} + \beta_3 \texttt{Age·Female}$$

- What is the survival probability for a 20 year old male?

$$\begin{aligned}
\eta &= \beta_0 + 20\beta_1 \\
\hat{\eta} &= -0.3312 \\
\pi &= \frac{e^{\eta}}{1+e^{\eta}} \\
\hat{\pi} &= .418
\end{aligned}$$

Males (cont'd)

- So, a 20-year old male in the Donner party had a 41.8% chance of surviving
- By a similar calculation, a 40-year male had a 27.3% chance of surviving and a 60-year old male a 16.4% chance

Females

- What about for a 20-year-old female?

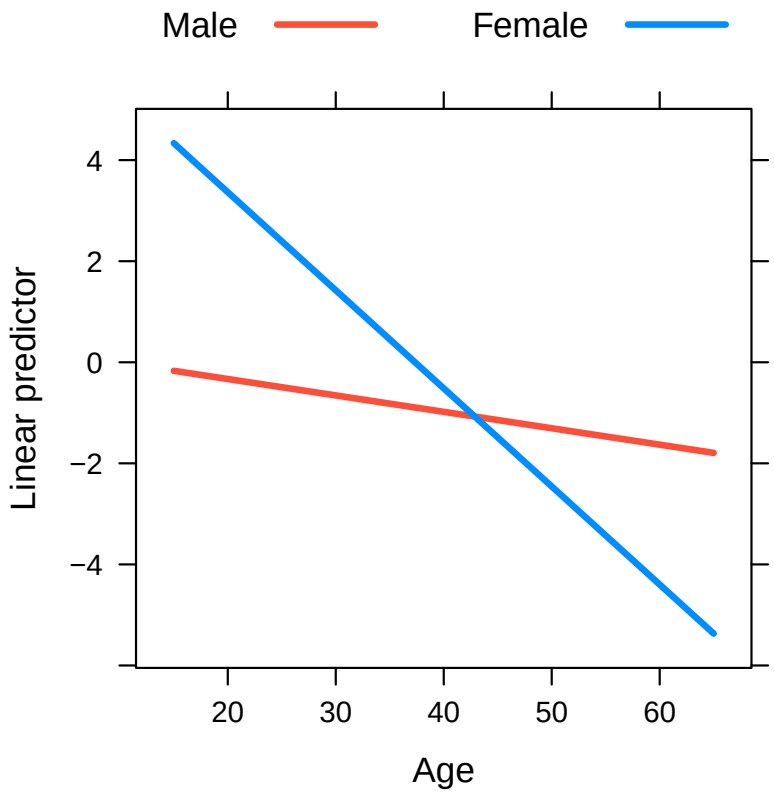
$$\eta = \beta_0 + 20\beta_1 + \beta_2 + 20\beta_3$$
$$\hat{\eta} = 3.3649$$
$$\hat{\pi} = .967$$

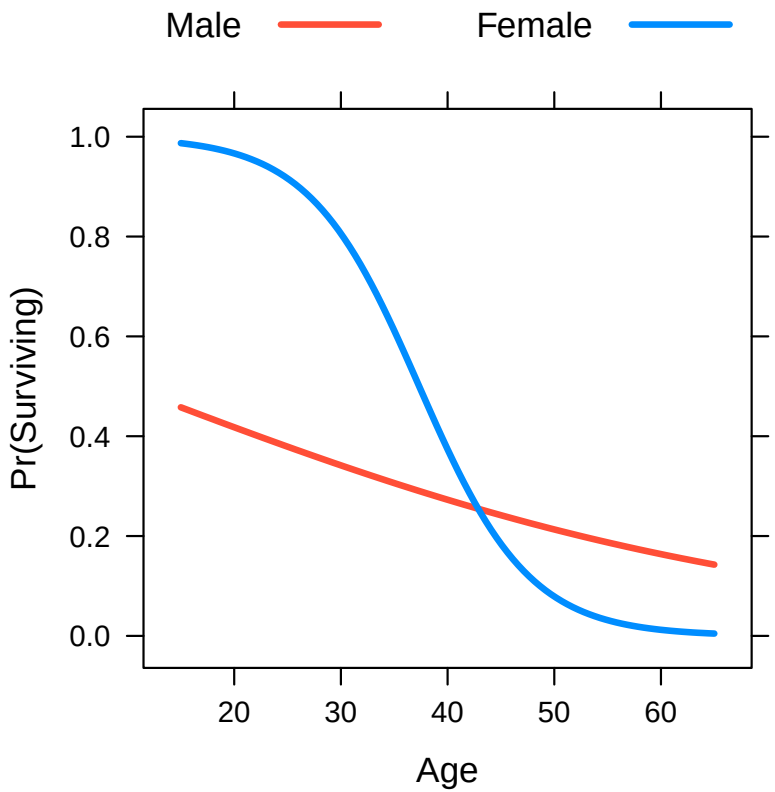
- So, 96.7% for a 20-year-old female; further calculation shows 37.4% for a 40-year old female, and just 1.2% for a 60-year old female
- To summarize the trend: young adult females had a much higher survival probability than young adult males, but their survival probability decreases much more rapidly with age, to the point where older females were much less likely to survive than older males

Automating these calculations

- We don't have to do all those calculations by hand; we can use an `ESTIMATE` statement in SAS or the `predict` function in `R`
- In particular, `R` and `PROC LOGISTIC` also allows you to return those predictions on the scale of the response
- This makes it easy to do things like plot the estimated probabilities

Logistic regression: Fit



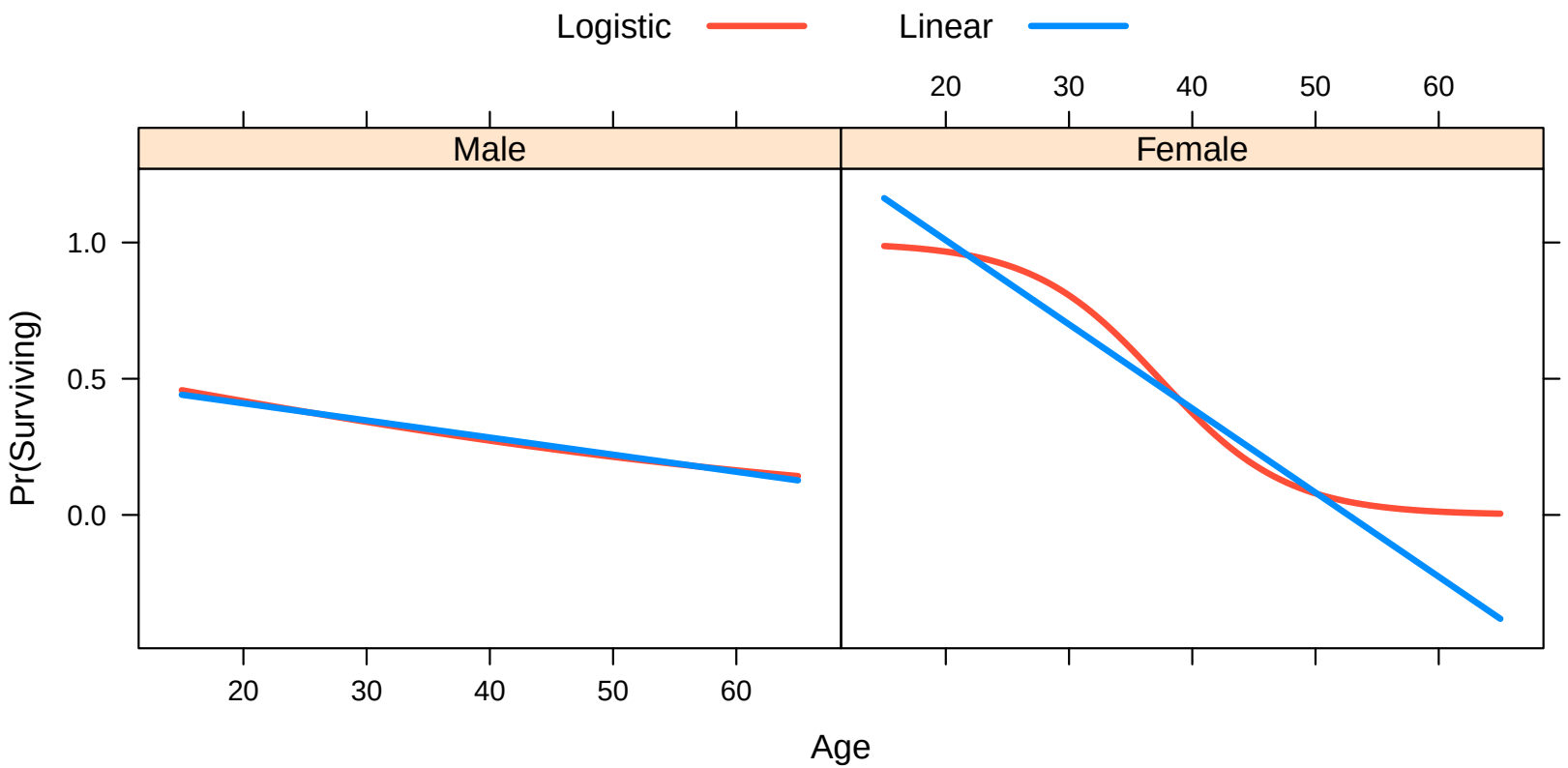


Sigmoidal curves

- As you can see (most clearly for the females), logistic regression produces fitted probabilities that take on an "S" shape (or *sigmoidal curve*)
- This comes from the logit link function, which as we have demonstrated earlier, constrains the fitted probabilities to lie within $[0, 1]$
- Once again, we see the utility of the link function: it allows us to obtain a nonlinear fit from a linear model

Linear vs. logistic regression

Linear regression is still the Best Linear Unbiased Estimator (BLUE), and in reasonable agreement for males, but the linearity requirement can lead to unrealistic estimates:



Thinking about the linearity assumption

- Finally, all of the same questions one asks in linear regression about the systematic part of the model are still relevant to GLMs
- For example, we have assumed a linear effect – is this reasonable?
- Perhaps the probability of survival is low for young adults and the elderly, and at its maximum in middle age

Quadratic effect: results

Including a quadratic effect for age in the systematic component of the model, we see that there may be some justification for this idea: