GENERALIZED LINEAR MODELS

Consider the normal theory Gauss-Markov linear model

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}, \quad \underset{\sim}{\varepsilon} \sim N(0, \sigma^2 I).$$

Another way to write this model is $Y_i \sim N(\mu_i, \sigma^2)$, where

$\mu_i = \underset{\sim}{x}_i' \underset{\sim}{\beta}$ for all $i = 1, \ldots, n$ and

$Y_1, \ldots, Y_n$ independent.

This is a special case of what is known as a generalized linear model.

Here is another special case:

$$Y_i \sim \text{Bernoulli}(\pi_i), \text{ where}$$

$$\pi_i = \frac{\exp(\underset{\sim}{x}_i'\underset{\sim}{\beta})}{1 + \exp(\underset{\sim}{x}_i'\underset{\sim}{\beta})} \quad \text{for all } i = 1, \ldots, n \text{ and}$$

$Y_1, \ldots, Y_n$ are independent.

In each example, all responses are independent and each response is a draw from one type of distribution whose parameters may depend on explanatory variables through a linear predictor $\underset{\sim}{x}_i'\underset{\sim}{\beta}$.

The second model, for the case of a binary response, is often called a logistic regression model.

Binary responses are common (success/failure, survive/die, good customer/bad customer, win/lose, etc.)

The logistic regression model can help us understand how explanatory variables are related to the probability of "success."

Disease Outbreak Study from *Applied Linear Statistical Models*, fourth edition, by Neter, Kutner, Nachtsheim, Wasserman (1996)

In health study to investigate an epidemic outbreak of a disease that is spread by mosquitoes, individuals were randomly sampled within two sectors in a city to determine if the person had recently contracted the disease under study.

y_i = 1 (person i has the disease)

y_i = 0 (person does not have the disease)

Potential explanatory variables include

age in years

socioeconomic status (1 = upper,
2 = middle,
3 = lower)

sector (1 or 2)

These variables were recorded for 196 randomly selected individuals.

Are any of these variables associated with the probability of disease and if so how?

We will demonstrate how to use R to fit a logistic regression model to this data set.

Before delving more deeply into logistic regression, we will review the basic facts of the Bernoulli distribution.

$y \sim \text{Bernoulli}(\pi)$ has probability mass function

$$f(y) = \begin{cases} \pi^y (1-\pi)^{1-y} & \text{for } y \in \{0, 1\} \\ 0 & \text{otherwise.} \end{cases}$$

Thus, $\Pr(y=0) = \pi^0 (1-\pi)^{1-0} = 1-\pi$

and $\Pr(y=1) = \pi^1 (1-\pi)^{1-1} = \pi$.

$$E(y) = \sum_y y f(y) = 0 \cdot (1-\pi) + 1 \cdot \pi = \pi.$$

$$E(y^2) = \sum_y y^2 f(y) = 0^2 \cdot (1-\pi) + 1^2 \cdot \pi = \pi.$$

$$\begin{aligned} Var(y) &= E(y) - \{E(y)\}^2 = \pi - \pi^2 \\ &= \pi(1-\pi). \end{aligned}$$

Note that $Var(y)$ is a function of $E(y)$.

The Logistic Regression Model

For $i = 1, \ldots, n$; $y_i \sim \text{Bernoulli}(\pi_i)$,

where $\pi_i = \dfrac{\exp(\underset{\sim}{x}_i' \underset{\sim}{\beta})}{1 + \exp(\underset{\sim}{x}_i' \underset{\sim}{\beta})}$ and

$y_1, \ldots, y_n$ are independent.

The function $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$ is called the logit function.

The logit function maps the interval $(0,1)$ to the real line $(-\infty, \infty)$.

π is a probability, so $\log\left(\frac{\pi}{1-\pi}\right)$ is the log ("odds").

(Odds of event $A \equiv \frac{Pr(A)}{1-Pr(A)}$.)

Note that

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

$$= \log\left[\frac{\exp(\underset{\sim}{x}_i'\underset{\sim}{\beta})}{1+\exp(\underset{\sim}{x}_i'\underset{\sim}{\beta})} \Big/ \frac{1}{1+\exp(\underset{\sim}{x}_i'\underset{\sim}{\beta})}\right]$$

$$= \log\left[\exp(\underset{\sim}{x}_i'\underset{\sim}{\beta})\right] = \underset{\sim}{x}_i'\underset{\sim}{\beta}.$$

Thus, the logistic regression model says that

$$y_i \sim \text{Bernoulli}(\pi_i) \text{ where}$$

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \underset{\sim}{x}_i' \underset{\sim}{\beta}.$$

In Generalized Linear Models terminology, the logit is called the <u>link function</u> because it "links" the mean of y_i (π_i) to the linear predictor $\underset{\sim}{x}_i' \underset{\sim}{\beta}$.

For Generalized Linear Models, it is not necessarily that the mean of y_i be a linear function of $\underset{\sim}{\beta}$.

Rather, some function of the mean of y_i is a linear function of $\underset{\sim}{\beta}$.

For logistic regression, that function is

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \underset{\sim}{x}_i'\underset{\sim}{\beta}.$$

When the response is Bernoulli or, more generally, binomial, the logit link function is one natural choice. However, other link functions can be considered.

Some common choices (that are also available in R) include the following:

logit: $\log\left(\frac{\pi}{1-\pi}\right) = \underset{\sim}{x}'\underset{\sim}{\beta}$

probit: $\Phi^{-1}(\pi) = \underset{\sim}{x}'\underset{\sim}{\beta}$

↑ Inverse of $N(0,1)$ CDF

complementary log-log (cloglog in R):

$$\log(-\log(1-\pi)) = \underset{\sim}{x}'\underset{\sim}{\beta}$$

Although any of these link functions (or others) can be used, the logit link has some advantages when it comes to interpreting the results (as we will discuss later).

Thus, the logit link is a good choice if it can provide a good fit to the data.

The likelihood function for logistic regression is

$$\ell(\underset{\sim}{\beta} | \underset{\sim}{y}) = \sum_{i=1}^{n} \log\left[\pi_i^{y_i} (1-\pi_i)^{1-y_i} \right]$$

$$= \sum_{i=1}^{n} \left[y_i \log(\pi_i) + (1-y_i) \log(1-\pi_i) \right]$$

$$= \sum_{i=1}^{n} \left[y_i \left\{ \log(\pi_i) - \log(1-\pi_i) \right\} + \log(1-\pi_i) \right]$$

$$= \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i} \right) + \log(1-\pi_i) \right]$$

$$= \sum_{i=1}^{n} \left[y_i \underset{\sim}{x}_i' \underset{\sim}{\beta} - \log\left(1 + \exp\{ \underset{\sim}{x}_i' \underset{\sim}{\beta} \} \right) \right].$$

For Generalized Linear Models, Fisher's Scoring Method is typically used to obtain an MLE for $\underset{\sim}{\beta}$, denoted by $\hat{\underset{\sim}{\beta}}$.

Fisher's Scoring Method is a variation of the Newton-Raphson algorithm in which the Hessian matrix (matrix of second partial derivatives) is replaced by its expected value (- Fisher Information matrix).

For Generalized Linear Models, Fisher's scoring method results in an iterative Weighted least squares procedure.

The algorithm is presented for the general case in Section 2.5 of Generalized Linear Models 2nd Edition (1989) by McCullagh and Nelder.

For sufficiently large samples,

$\hat{\underset{\sim}{\beta}}$ is approximately normal with

mean $\underset{\sim}{\beta}$ and a variance-covariance

matrix that can be approximated

by the estimated inverse of the

Fisher information matrix.

Inference can be conducted using the Wald approach or via likelihood ratio testing as discussed in our slides on likelihood-related topics.

Interpretation of Logistic Regression Parameters:

Let $\tilde{\underset{\sim}{x}} = (x_1, x_2, \ldots, x_{j-1}, x_j + 1, x_{j+1}, \ldots, x_p)'$.

In other words, $\tilde{\underset{\sim}{x}}$ is the same as $\underset{\sim}{x}$ except that the j^{th} explanatory variable has been increased by one unit.

Let $\pi = \dfrac{\exp(\underset{\sim}{x}'\underset{\sim}{\beta})}{1 + \exp(\underset{\sim}{x}'\underset{\sim}{\beta})}$ and $\tilde{\pi} = \dfrac{\exp(\tilde{\underset{\sim}{x}}'\underset{\sim}{\beta})}{1 + \exp(\tilde{\underset{\sim}{x}}'\underset{\sim}{\beta})}$.

The odds ratio

$$\frac{\frac{\tilde{\pi}}{1-\tilde{\pi}}}{\frac{\pi}{1-\pi}} = \exp\left\{\log\left(\frac{\frac{\tilde{\pi}}{1-\tilde{\pi}}}{\frac{\pi}{1-\pi}}\right)\right\}$$

$$= \exp\left\{\log\left(\frac{\tilde{\pi}}{1-\tilde{\pi}}\right) - \log\left(\frac{\pi}{1-\pi}\right)\right\}$$

$$= \exp\left\{\tilde{x}'\beta - x'\beta\right\}$$

$$= \exp\left\{(x_j + 1)\beta_j - x_j\beta_j\right\}$$

$$= \exp\left\{\beta_j\right\}.$$

Thus, $\frac{\tilde{\pi}}{1-\tilde{\pi}} = \exp(\beta_j)\frac{\pi}{1-\pi}$.

All other explanatory variables held constant, the odds of success at $x_j + 1$ are $\exp(\beta_j)$ times the odds of success at x_j.

This is true regardless of the initial value x_j.

A 1 unit increase in the jth explanatory variable (with all other explanatory variables held constant) is associated with a multiplicative change in the odds of success by the factor $\exp(\beta_j)$.

If (L_j, U_j) is a $100(1-\alpha)\%$ confidence interval for β_j, then

$$(\exp\{L_j\}, \exp\{U_j\})$$

is a $100(1-\alpha)\%$ confidence interval for $\exp\{\beta_j\}$.

Also, note that

$$\pi = \frac{\exp(\underset{\sim}{x}'\underset{\sim}{\beta})}{1+\exp(\underset{\sim}{x}'\underset{\sim}{\beta})} = \frac{1}{\frac{1}{\exp(\underset{\sim}{x}'\underset{\sim}{\beta})}+1}$$

$$= \frac{1}{1+\exp(-\underset{\sim}{x}'\underset{\sim}{\beta})}.$$

Thus, if (L_j, U_j) is a $100(1-\alpha)\%$ confidence interval for $\underset{\sim}{x}'\underset{\sim}{\beta}$, then a $100(1-\alpha)\%$ confidence interval for π is $\left(\frac{1}{1+\exp(-L_j)}, \frac{1}{1+\exp(-U_j)}\right)$.

```
d=read.delim(
"http://www.public.iastate.edu/~dnett/S511/Disease.txt")

head(d)

  id age ses sector disease savings
1  1  33   1      1       0       1
2  2  35   1      1       0       1
3  3   6   1      1       0       0
4  4  60   1      1       0       1
5  5  18   3      1       1       0
6  6  26   3      1       0       0

d$ses=as.factor(d$ses)
d$sector=as.factor(d$sector)
```

```
o=glm(disease~age+ses+sector,
       family=binomial(link=logit),
       data=d)



summary(o)

Call:
glm(formula = disease ~ age + ses + sector, family =
binomial(link = logit),
    data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6576  -0.8295  -0.5652   1.0092   2.0842
```

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.293933   0.436769  -5.252 1.50e-07 ***
age          0.026991   0.008675   3.111 0.001862 **
ses2         0.044609   0.432490   0.103 0.917849
ses3         0.253433   0.405532   0.625 0.532011
sector2      1.243630   0.352271   3.530 0.000415 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 236.33  on 195  degrees of freedom
Residual deviance: 211.22  on 191  degrees of freedom
AIC: 221.22

Number of Fisher Scoring iterations: 3
```

```
coef(o)
(Intercept)         age        ses2        ses3     sector2
-2.29393347  0.02699100  0.04460863  0.25343316  1.24363036

v=vcov(o)
round(v,3)
            (Intercept)    age   ses2   ses3 sector2
(Intercept)       0.191 -0.002 -0.083 -0.102  -0.080
age              -0.002  0.000  0.000  0.000   0.000
ses2             -0.083  0.000  0.187  0.072   0.003
ses3             -0.102  0.000  0.072  0.164   0.039
sector2          -0.080  0.000  0.003  0.039   0.124


confint(o)
Waiting for profiling to be done...
                   2.5 %      97.5 %
(Intercept) -3.19560769 -1.47574975
age          0.01024152  0.04445014
ses2        -0.81499026  0.89014587
ses3        -0.53951033  1.05825383
sector2      0.56319260  1.94992969
```

```
oreduced=glm(disease~age+sector,
       family=binomial(link=logit),
       data=d)

anova(oreduced,o,test="Chisq")
Analysis of Deviance Table

Model 1: disease ~ age + sector
Model 2: disease ~ age + ses + sector
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1       193     211.64
2       191     211.22  2   0.4193    0.8109
```

```
o=oreduced

anova(o,test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: disease

Terms added sequentially (first to last)


       Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     195     236.33
age     1   12.013       194     224.32 0.0005283 ***
sector  1   12.677       193     211.64 0.0003702 ***
```

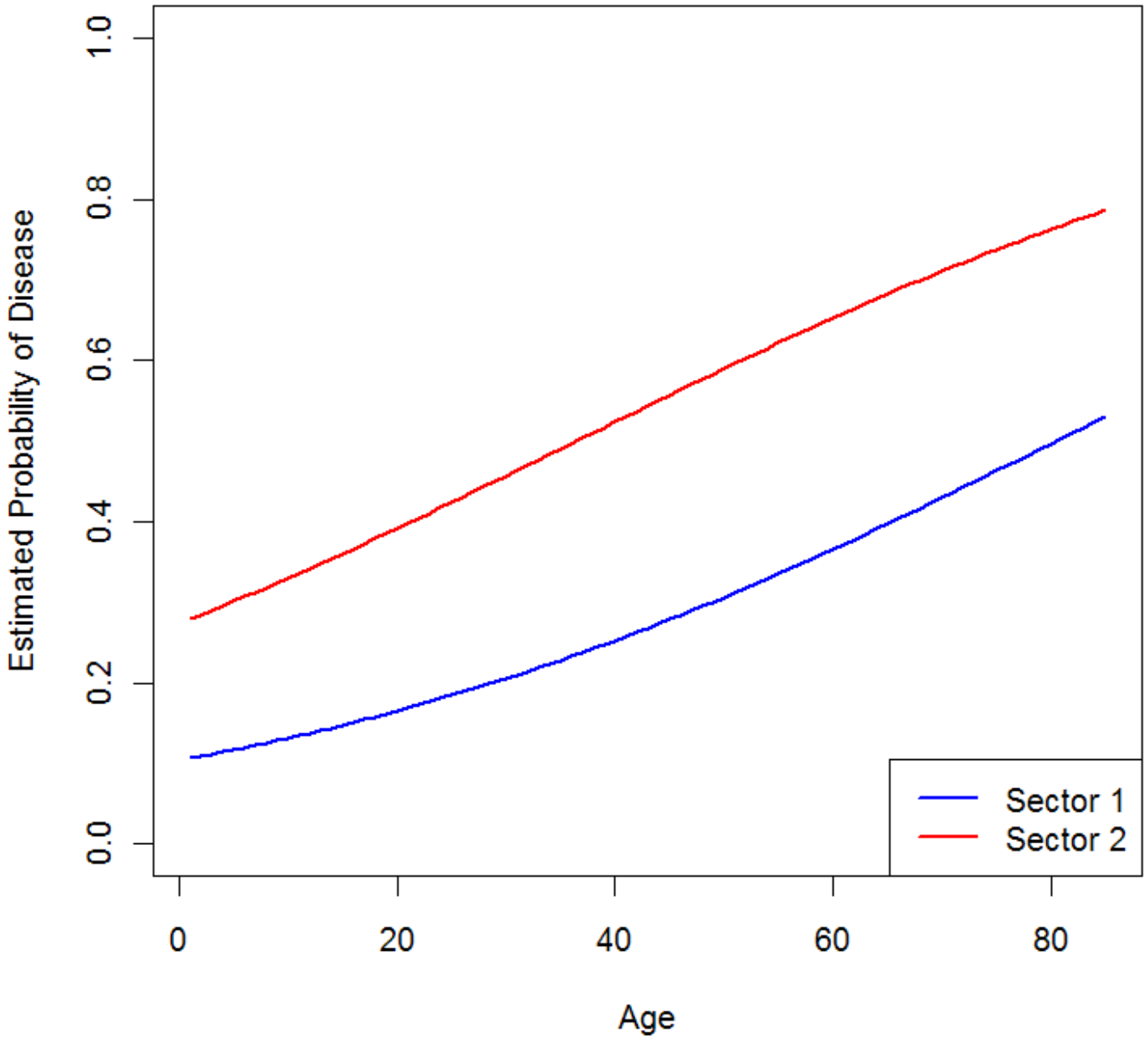
```
head(model.matrix(o))
  (Intercept) age sector2
1           1  33       0
2           1  35       0
3           1   6       0
4           1  60       0
5           1  18       0
6           1  26       0

b=coef(o)
b
(Intercept)         age     sector2
-2.15965912  0.02681289  1.18169345

ci=confint(o)
Waiting for profiling to be done...
ci
                   2.5 %      97.5 %
(Intercept) -2.86990940 -1.51605906
age          0.01010532  0.04421365
sector2      0.52854584  1.85407936
```

```
#How should we interpret our estimate of
#the slope coefficient on age?

exp(b[2])
     age
1.027176

#All else equal, the odds of disease are
#about 1.027 times greater for someone age
#x+1 than for someone age x. An increase of
#one year in age is associated with an
#increase in the odds of disease by about 2.7%.
#A 95% confidence interval for the multiplicative
#increase factor is

exp(ci[2,])
   2.5 %   97.5 %
1.010157 1.045206
```

```
#How should we interpret our estimate of
#the slope coefficient on sector?

exp(b[3])
sector2
3.25989

#All else equal, the odds of disease are
#about 3.26 times greater for someone
#living in sector 2 than for someone living
#in sector 1.

#A 95% confidence interval for the multiplicative
#increase factor is

exp(ci[3,])
   2.5 %   97.5 %
1.696464 6.385816
```

```
#Estimate the probability that a randomly
#selected 40-year-old living in sector 2
#has the disease.

x=c(1,40,1)

1/(1+exp(-t(x)%*%b))
          [,1]
[1,] 0.5236198

#Approximate 95% confidence interval
#for the probability in question.

sexb=sqrt(t(x)%*%vcov(o)%*%x)

cixb=c(t(x)%*%b-2*sexb,t(x)%*%b+2*sexb)

1/(1+exp(-cixb))

[1] 0.3965921 0.6476635
```

```
#Plot estimated probabilities as a function
#of age for each sector.

x=1:85
plot(x,1/(1+exp(-(b[1]+b[2]*x))),ylim=c(0,1),
     type="l",col=4,lwd=2,xlab="Age",
     ylab="Estimated Probability of Disease")
lines(x,1/(1+exp(-(b[1]+b[2]*x+b[3]))),col=2,lwd=2)
legend("bottomright",legend=c("Sector 1","Sector 2"),
       col=c(4,2),lwd=2)
```


Estimated Probability of Disease
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
Age
Sector 1
Sector 2

Now suppose that instead of a Bernoulli response, we have a binomial response for each unit in an experiment or an observational study.

As an example, consider the trout data set discussed on page 641 of *The Statistical Sleuth*, second edition, by Ramsey and Schafer.

Five doses of toxic substance were assigned to a total of 20 fish tanks using a completely randomized design with four tanks per dose.

For each tank, the total number of fish and the number of fish that developed liver tumors were recorded.

```
d=read.delim(
"http://www.public.iastate.edu/~dnett/S511/Trout.txt")

d
    dose tumor total
1  0.010     9    87
2  0.010     5    86
3  0.010     2    89
4  0.010     9    85
5  0.025    30    86
6  0.025    41    86
7  0.025    27    86
8  0.025    34    88
9  0.050    54    89
10 0.050    53    86
11 0.050    64    90
12 0.050    55    88
13 0.100    71    88
14 0.100    73    89
15 0.100    65    88
16 0.100    72    90
17 0.250    66    86
18 0.250    75    82
19 0.250    72    81
20 0.250    73    89
```

One way to analyze this data would be to convert the binomial counts and totals into Bernoulli responses.

For example, the first line of the data set could be converted into 9 ones and 87-9=78 zeros. Each of these 87 observations would have dose 0.01 as their explanatory variable value.

We could then use the logistic regression modeling strategy for Bernoulli response as described above.

A simpler and equivalent way to deal with this data is to consider a logistic regression model for the binomial counts directly.

Logistic Regression Model for Binomial Count Data:

For all $i = 1, \ldots, n$; $Y_i \sim \text{Binomial}(m_i, \pi_i)$,

where m_i is a known number of trials for observation i,

$$\pi_i = \frac{\exp(\underset{\sim}{x}_i'\underset{\sim}{\beta})}{1 + \exp(\underset{\sim}{x}_i'\underset{\sim}{\beta})}, \quad \text{and}$$

$Y_1, \ldots, Y_n$ are independent.

Recall that for $y_i \sim \text{Binomial}(m_i, \pi_i)$,

$$E(y_i) = m_i \pi_i \qquad \text{Var}(y_i) = m_i \pi_i (1-\pi_i)$$

$$f(y_i) = \binom{m_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{m_i - y_i} \quad \text{for } y_i \in \{0, \ldots, m_i\}.$$

$$\ell(\underset{\sim}{\beta} \mid \underset{\sim}{y}) = \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + m_i \log(1-\pi_i) \right] + \text{const}$$

$$= \sum_{i=1}^{n} \left[y_i \underset{\sim}{x}_i' \underset{\sim}{\beta} - m_i \log\left(1 + \exp\{-\underset{\sim}{x}_i' \underset{\sim}{\beta}\}\right) \right]$$

$$+ \text{const.}$$

The function $\ell(\underset{\sim}{\beta}/\underset{\sim}{y})$ can be maximized over $\underset{\sim}{\beta} \in \mathbb{R}^p$ as discussed previously to obtain an MLE $\hat{\underset{\sim}{\beta}}$.

We can compare the fit of a logistic regression model to what is known as a "saturated" model.

The saturated model uses one parameter for each observation.

In this case, there is one π_i parameter for each y_i.

Logistic Regression Model

$Y_i \sim \text{Binomial}(m_i, \pi_i)$

$Y_1, \ldots, Y_n$ independent

$$\pi_i = \frac{\exp(\underline{x}_i' \underline{\beta})}{1 + \exp(\underline{x}_i' \underline{\beta})}$$

for some $\underline{\beta} \in \mathbb{R}^p$

p parameters

Saturated Model

$Y_i \sim \text{Binomial}(m_i, \pi_i)$

$Y_1, \ldots, Y_n$ independent

$\pi_i \in [0,1]$ for $i = 1, \ldots, n$ with no other restrictions.

n parameters

Let $\hat{\pi}_i = \dfrac{\exp(\underset{\sim}{x}_i'\underset{\sim}{\beta})}{1+\exp(\underset{\sim}{x}_i'\underset{\sim}{\beta})}$ denote the MLE of π_i under the logistic regression model $\forall\ i=1,\ldots,n$.

Under the saturated model, the MLE of π_i is y_i/m_i $\forall\ i=1,\ldots,n$.

Then the likelihood ratio statistic for testing the logistic regression model as the reduced model vs. the saturated model as the full model is

$$\sum_{i=1}^{n} 2\left[y_i \log\left(\frac{y_i/m_i}{\hat{\pi}_i}\right) + (m_i - y_i)\log\left(\frac{1-y_i/m_i}{1-\hat{\pi}_i}\right)\right]$$

This statistic is sometimes called the Deviance Statistic, the Residual Deviance, or just the Deviance.

The statistic can be compared to the χ^2_{n-p} distribution to check the goodness of fit of the logistic regression model.

The χ^2 approximation to the null distribution works reasonably well if $m_i \geq 5$ for most i.

The term

$$d_i \equiv \text{sign}(y_i - m_i\hat{\pi}_i)\sqrt{2\left[y_i \log\left(\frac{y_i}{m_i\hat{\pi}_i}\right) + (m_i - y_i)\log\left(\frac{m_i - y_i}{m_i - m_i\hat{\pi}_i}\right)\right]}$$

is called a <u>deviance residual</u>.

Note that the residual deviance statistic $= \sum_{i=1}^{n} d_i^2$.

Another goodness of fit statistic that is approximately χ^2_{n-p} under the null is Pearson's Chi-Square Statistic

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{y_i - m_i \hat{\pi}_i}{\sqrt{m_i \hat{\pi}_i (1 - \hat{\pi}_i)}} \right)^2$$

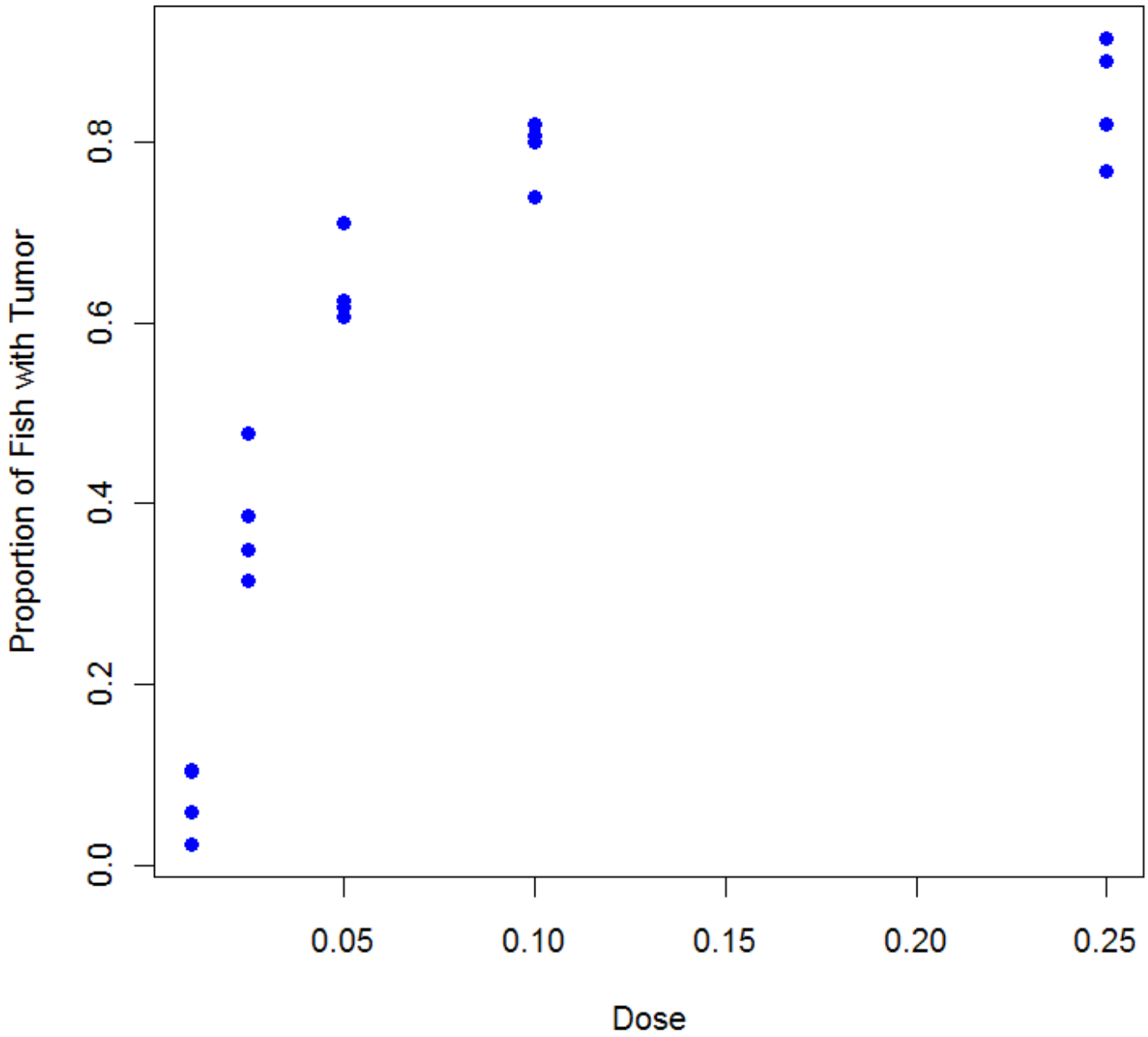
$$= \sum_{i=1}^{n} \left(\frac{y_i - \hat{E}(y_i)}{\sqrt{\hat{Var}(y_i)}} \right)^2 .$$

$$r_i = \frac{y_i - m_i \hat{\pi}_i}{\sqrt{m_i \hat{\pi}_i (1 - \hat{\pi}_i)}}$$ is known as a Pearson residual.

$$X^2 = \sum_{i=1}^{n} r_i^2.$$

For large m_i's, both d_i and r_i should behave like standard normal random variables if the logistic regression model is correct.

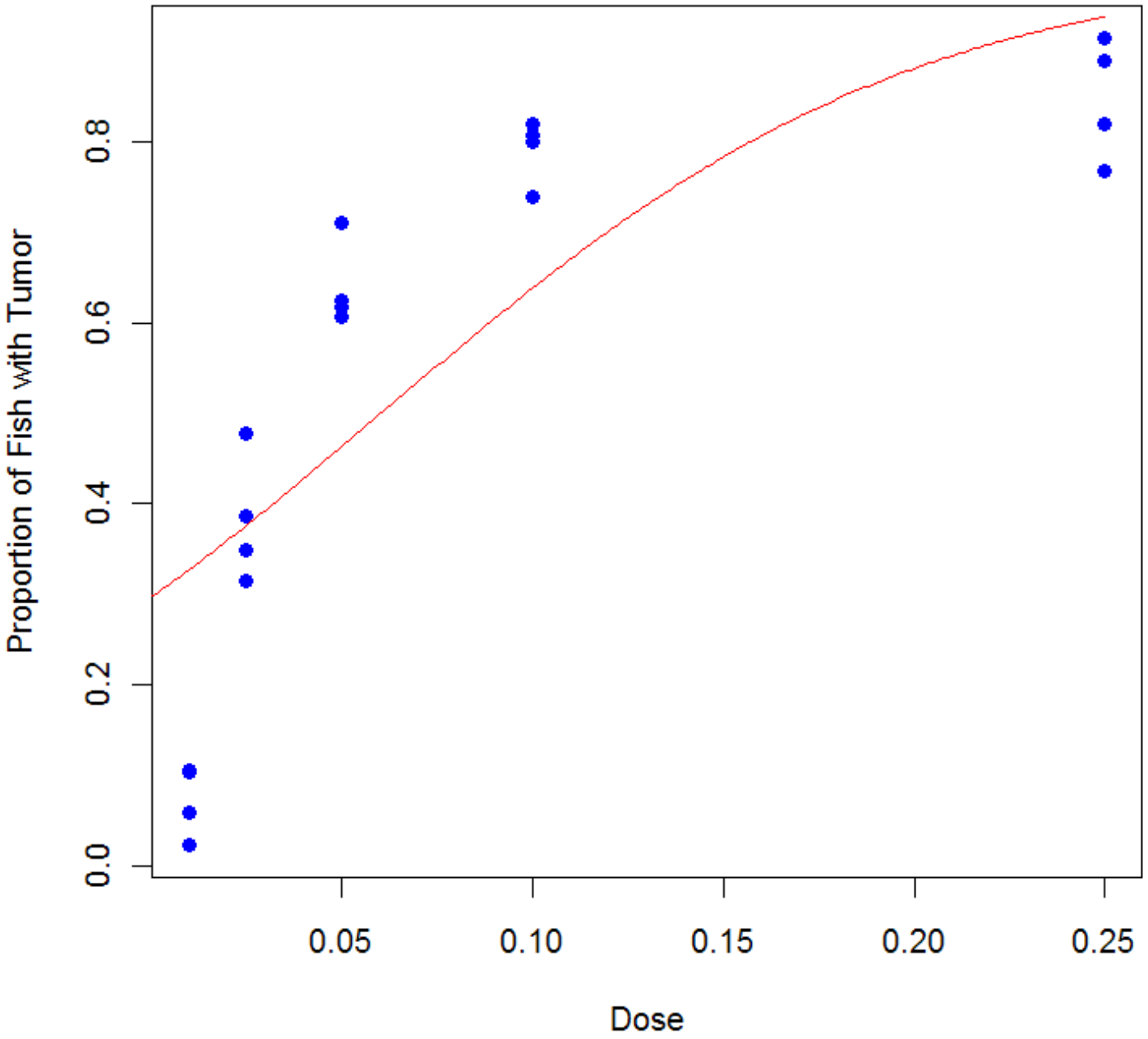
```
#Let's plot observed tumor proportions
#for each tank.

plot(d$dose,d$tumor/d$total,col=4,pch=19,
      xlab="Dose",
      ylab="Proportion of Fish with Tumor")
```


Proportion of Fish with Tumor
0.0
0.2
0.4
0.6
0.8
0.05
0.10
0.15
0.20
0.25
Dose

```
#Let's fit a logistic regression model
#dose is a quantitative explanatory variable.

o=glm(cbind(tumor,total-tumor)~dose,
       family=binomial(link=logit),
       data=d)

summary(o)

Call:
glm(formula = cbind(tumor, total - tumor) ~ dose,
family = binomial(link = logit),
    data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-7.3577  -4.0473  -0.1515   2.9109   4.7729
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.86705    0.07673  -11.30   <2e-16 ***
dose        14.33377    0.93695   15.30   <2e-16 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 667.20  on 19  degrees of freedom
Residual deviance: 277.05  on 18  degrees of freedom
AIC: 368.44

Number of Fisher Scoring iterations: 5
```

```
#Let's plot the fitted curve.

b=coef(o)
u=seq(0,.25,by=0.001)
xb=b[1]+u*b[2]
pihat=1/(1+exp(-xb))
lines(u,pihat,col=2,lwd=1.3)
```


Proportion of Fish with Tumor
0.0
0.2
0.4
0.6
0.8
0.05
0.10
0.15
0.20
0.25
Dose

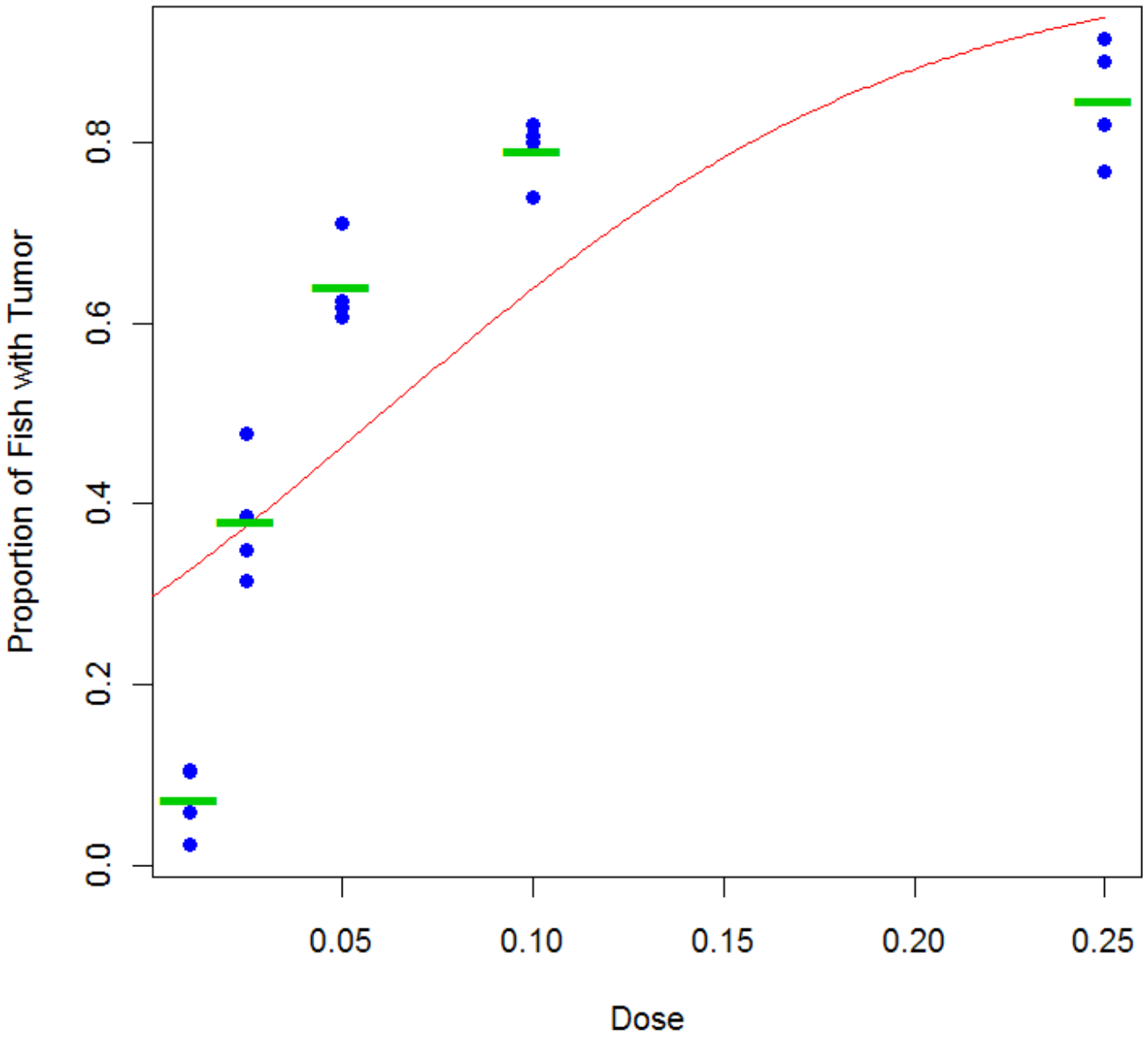
```
#Let's use a reduced versus full model
#likelihood ratio test to test for
#lack of fit relative to the
#saturated model.

1-pchisq(deviance(o),df.residual(o))
[1] 0


#We could try adding higher-order
#polynomial terms, but let's just
#skip right to the model with dose
#as a categorical variable.
```

```
d$dosef=gl(5,4)
d
    dose tumor total dosef
1  0.010     9    87     1
2  0.010     5    86     1
3  0.010     2    89     1
4  0.010     9    85     1
5  0.025    30    86     2
6  0.025    41    86     2
7  0.025    27    86     2
8  0.025    34    88     2
9  0.050    54    89     3
10 0.050    53    86     3
11 0.050    64    90     3
12 0.050    55    88     3
13 0.100    71    88     4
14 0.100    73    89     4
15 0.100    65    88     4
16 0.100    72    90     4
17 0.250    66    86     5
18 0.250    75    82     5
19 0.250    72    81     5
20 0.250    73    89     5
```

```
o=glm(cbind(tumor,total-tumor)~dosef,
       family=binomial(link=logit),
       data=d)

summary(o)

Call:
glm(formula = cbind(tumor, total - tumor) ~ dosef,
family = binomial(link = logit),
    data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0966  -0.6564  -0.1015   1.0793   1.8513
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.5557     0.2076 -12.310   <2e-16 ***
dosef2        2.0725     0.2353   8.809   <2e-16 ***
dosef3        3.1320     0.2354  13.306   <2e-16 ***
dosef4        3.8900     0.2453  15.857   <2e-16 ***
dosef5        4.2604     0.2566  16.605   <2e-16 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 667.195  on 19  degrees of freedom
Residual deviance:  25.961  on 15  degrees of freedom
AIC: 123.36

Number of Fisher Scoring iterations: 4
```

```
#Let's add the new fitted values to our plot.

fitted(o)
        1         2         3         4         5         6         7
0.07204611 0.07204611 0.07204611 0.07204611 0.38150289 0.38150289 0.38150289
        8         9        10        11        12        13        14
0.38150289 0.64022663 0.64022663 0.64022663 0.64022663 0.79154930 0.79154930
       15        16        17        18        19        20
0.79154930 0.79154930 0.84615385 0.84615385 0.84615385 0.84615385

points(d$dose,fitted(o),pch="_",cex=3,col=3)
```


Proportion of Fish with Tumor
0.0
0.2
0.4
0.6
0.8
0.05
0.10
0.15
0.20
0.25
Dose

```
#The fit looks good, but let's formally
#test for lack of fit.

1-pchisq(deviance(o),df.residual(o))
[1] 0.03843272

#There is still a significant lack of fit
#when comparing to the saturated model.

#The problem is over dispersion, otherwise
#known in this case as extra binomial variation.
```

Overdispersion:

In the Generalized Linear Models framework, it's often the case that $Var(y_i)$ is a function of $E(y_i)$.

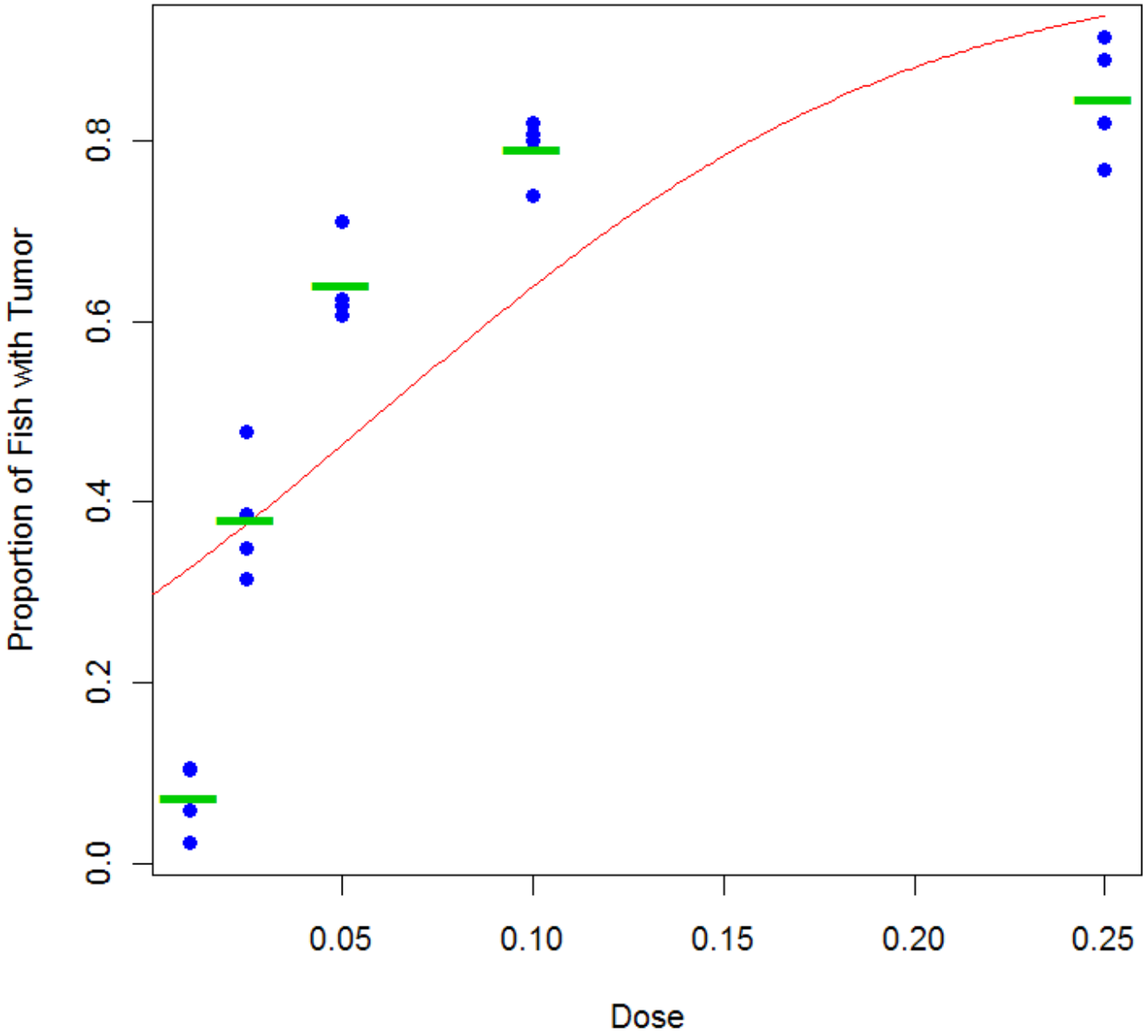
That is the case for logistic regression where

$$Var(y_i) = m_i \pi_i (1-\pi_i) = m_i \pi_i - \frac{(m_i \pi_i)^2}{m_i}$$

$$= E(y_i) - [E(y_i)]^2 / m_i$$

Thus, when we fit a logistic regression model and obtain estimates of the mean of the response, we get estimates of the variance of the response as well.

If the variability of our response is greater than we should expect based on our estimates of the mean, we say that there is <u>overdispersion.</u>


Proportion of Fish with Tumor
0.0
0.2
0.4
0.6
0.8
0.05
0.10
0.15
0.20
0.25
Dose

If either the likelihood ratio-based or the Pearson Chi Square-based test of goodness of fit (or lack of fit), suggests a lack of fit that cannot be explained by other reasons (e.g., poor model for the Mean or a few extreme outliers) overdispersion may be the problem.

If there is overdispersion, a quasi-likelihood approach may be used.

In the binomial case we make all the same assumptions as before except that we assume $Var(y_i) = \phi m_i \pi_i (1-\pi_i)$ for some unknown dispersion parameter $\phi > 1$.

The dispersion parameter ϕ can be estimated

by $\dfrac{\sum_{i=1}^{n} di^2}{n-p}$ or $\dfrac{\sum_{i=1}^{n} r_i^2}{n-p}$

↑ Residual Deviance Statistic

↑ Pearson Chi-Square Statistic

All analyses are as before except that

1. The estimated variance of $\hat{\underset{\sim}{\beta}}$ is multiplied by $\hat{\phi}$.
2. For Wald type inferences, the standard normal null distribution is replaced by t with $n-p$ degrees of freedom.
3. A test statistic T that was assumed χ^2_q under H_0 is replaced with $T/(q\hat{\phi})$ and compared to an F distribution with q and $n-p$ degrees of freedom.

These changes to the inference strategy in the presence of overdispersion are analogous to the changes that would take place in normal theory Gauss-Markov linear model analysis if we switched from assuming σ^2 was known to be 1 to assuming σ^2 was unknown and estimating it with MSE.

(Here ϕ is like σ^2 and $\hat{\phi}$ is like MSE.)

Whether there is overdispersion or not, all the usual ways of conducting generalized linear models Inference are approximate except for the special case of normal theory linear models.

```
#Let's estimate the dispersion parameter.

phihat=deviance(o)/df.residual(o)
phihat
[1] 1.730745

#We can obtain the same estimate by using
#the deviance residuals.

di=residuals(o,type="deviance")
sum(di^2)/df.residual(o)
[1] 1.730745

#We can obtain an alternative estimate by
#using the Pearson residuals.

ri=residuals(o,type="pearson")
phihat=sum(ri^2)/df.residual(o)
phihat
[1] 1.671226
```

```
#Now we will conduct a quasilikelihood analysis
#that accounts for overdispersion.

oq=glm(cbind(tumor,total-tumor)~dosef,
       family=quasibinomial(link=logit),
       data=d)

summary(oq)

Call:
glm(formula = cbind(tumor, total - tumor) ~ dosef,
family = quasibinomial(link = logit),
    data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0966  -0.6564  -0.1015   1.0793   1.8513
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.5557     0.2684  -9.522 9.48e-08 ***
dosef2        2.0725     0.3042   6.814 5.85e-06 ***
dosef3        3.1320     0.3043  10.293 3.41e-08 ***
dosef4        3.8900     0.3171  12.266 3.20e-09 ***
dosef5        4.2604     0.3317  12.844 1.70e-09 ***

(Dispersion parameter for quasibinomial family taken
to be 1.671232)

    Null deviance: 667.195  on 19  degrees of freedom
Residual deviance:  25.961  on 15  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

```
#Test for the effect of dose on the response.

drop1(oq,test="F")
Single term deletions

Model:
cbind(tumor, total - tumor) ~ dosef
       Df Deviance F value     Pr(F)
<none>       25.96
dosef   4   667.20  92.624 2.187e-10 ***


#There is strong evidence that
#the probability of tumor formation
#is different for different doses
#of the toxicant.
```

```
#Let's test for a difference between
#the top two doses.

b=coef(oq)
b
(Intercept)      dosef2      dosef3      dosef4      dosef5
  -2.555676    2.072502    3.132024    3.889965    4.260424

v=vcov(oq)
v
            (Intercept)      dosef2      dosef3      dosef4      dosef5
(Intercept)   0.0720386 -0.07203860 -0.07203860 -0.0720386 -0.0720386
dosef2       -0.0720386  0.09250893  0.07203860  0.0720386  0.0720386
dosef3       -0.0720386  0.07203860  0.09259273  0.0720386  0.0720386
dosef4       -0.0720386  0.07203860  0.07203860  0.1005702  0.0720386
dosef5       -0.0720386  0.07203860  0.07203860  0.0720386  0.1100211
```

```
se=sqrt(t(c(0,0,0,-1,1))%*%v%*%c(0,0,0,-1,1))

tstat=(b[5]-b[4])/se

pval=2*(1-pt(abs(tstat),df.residual(oq)))

pval
0.1714103
```

We have discussed the case of Bernoulli or binomial response, where logistic regression modeling is a natural generalized linear modeling strategy.

Another commonly encountered special case of generalized linear modeling involves Poisson response.

We begin with a review of the basics of the Poisson distribution.

$$y \sim \text{Poisson}(\mu) \Longrightarrow$$

$$f(y) = \begin{cases} \dfrac{\mu^y e^{-\mu}}{y!} & \text{for } y = 0, 1, 2, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

$$E(y) = \mu \qquad \text{Var}(y) = \mu$$

The usual Generalized Linear Model for Poisson response:

For all $i = 1, \ldots, n$;

$$Y_i \sim \text{Poisson}(\mu_i), \text{ where}$$

$$\mu_i = \exp(\underset{\sim}{x}_i' \underset{\sim}{\beta}) \text{ and}$$

$Y_1, \ldots, Y_n$ are independent.

Note that

$$\mu_i = \exp(\underset{\sim}{x}_i' \underset{\sim}{\beta}) \iff \log(\mu_i) = \underset{\sim}{x}_i' \underset{\sim}{\beta}.$$

Thus, log is the link function in this case.

All the subsequent details for the Poisson case are analogous to those we discussed for the binomial response case.

The general case: For $i = 1, \ldots, n$,

suppose Y_i has density (or p.m.f.)

$$\exp\left\{ \left(Y_i \theta_i - b(\theta_i) \right) / a(\phi) + c(Y_i, \phi) \right\},$$

where $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are known functions and θ_i is an unknown parameter and ϕ is either a known or unknown parameter depending on the special case.

For all $i = 1, \ldots, n$;

let $\mu_i = E(y_i)$ and assume

that $g(\mu_i) = \underset{\sim}{x}_i' \underset{\sim}{\beta}$ for some

link function $g(\cdot)$, known vector

of explanatory variables $\underset{\sim}{x}_i$, and

unknown parameter vector $\underset{\sim}{\beta} \in \mathbb{R}^p$.

Finally, suppose $y_1, \ldots, y_n$ independent.

Analysis Strategy:

1. Find MLE for $\underset{\sim}{\beta}$ using the method of Fisher Scoring which results in an iterative weighted least squares approach in this case.

2. Obtain an estimate of the inverse Fisher information matrix that can be used for Wald type inference concerning $\underset{\sim}{\beta}$ and/or conduct likelihood ratio based inference of reduced vs. full models.