Contents

Generalized linear model

- Consider the normal theory Gauss-Markov linear model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\boldsymbol{\varepsilon} \sim N(0, \sigma^2 I_n)$
  - Does not have to be written as function + error (additive form)
  - Could be rewritten as "distribution" of $\boldsymbol{y}$ and the parameter for that distribution in terms of $\boldsymbol{X}\boldsymbol{\beta}$ and σ^2, i.e.

$$y_i \overset{i.i.d.}{\sim} N(\mu_i, \sigma^2), \text{ where } \mu_i = \boldsymbol{x}_i'\boldsymbol{\beta} \text{ for } i = 1, \ldots, n$$

- Such a way to specify a model is referred to *generalize linear model*
  - Another example is $y_i \overset{i.i.d.}{\sim} \text{Ber}(\pi_i)$ where $\pi = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})/(1 + \exp(\boldsymbol{x}_i'\boldsymbol{\beta}))$ for all $i = 1, \ldots, n$, which is often called a *logistic regression model*.
  - The logistic regression model can help us understand how explanatory variables are related to the probability of "success"
  - 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 *known* function of a linear predictor $\boldsymbol{x}'\boldsymbol{\beta}$

- The normal and Bernoulli models (and many others) are special cases of a generalized linear model.
- These are models formulated in the following way
  1. **Random Component**: $\{\boldsymbol{y}_i\}_{i=1}^n$ are n independent responses that follow a probability distribution in the *exponential family of probability distributions* that is the pdf reads

$$\exp([\boldsymbol{\eta}(\boldsymbol{\theta})\boldsymbol{T}(y_i) - \boldsymbol{b}(\boldsymbol{\theta})]/a(\phi) + c(y_i, \phi))$$

     where $\boldsymbol{\eta}, \boldsymbol{T}, a, \boldsymbol{b}$ and c are known functions and $\boldsymbol{\theta}$ is a vector of unknown parameters depending on $\boldsymbol{X\beta}$ and ϕ is either known or unknown, $a(\phi)$ models the overdispersion effects.
  2. **Systematic Component**: A linear predictor $\boldsymbol{\eta} = \boldsymbol{X\beta}$ or $w(\boldsymbol{\eta}) = \boldsymbol{X\beta}$
  3. **Link**: A link function g relates the linear predictor to the mean response

$$g(\mathbb{E}(\boldsymbol{y})) = g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{X\beta}$$

- For example, the pdf of a normal distribution can be written as

$$\exp[-y_i^2/(2\sigma^2) + y_i(\mu/(2\sigma^2)) - \mu^2/(2\sigma^2) - \log(2\pi\sigma^2)/2]$$

with $\boldsymbol{\eta}(\boldsymbol{\theta}) = (\mu/\sigma^2, -1/(2\sigma^2))$ and $\boldsymbol{T}(y_i) = (y_i, y_i^2)$

  - Many theoretical results follow immediately from the exponential form: e.g. $\boldsymbol{T}(y_i)$ is the vector of sufficient statistics (STAT530,730)
  - Many theoretical results are much nicer when distribution parameterized in terms of $\boldsymbol{\eta}(\boldsymbol{\theta})$ instead of $\boldsymbol{\theta}$, such as $\log(\pi/(1-\pi))$ for Bernoulli distribution instead of π. This form is called *the canonical or natural form*

- Exponential family can be used model different distribution and data

Data type	Distribution	$\boldsymbol{\eta}(\boldsymbol{\theta})$	$\boldsymbol{T}(y_i)$	$a(\phi)$
Continuous	Normal	$(\mu/\sigma^2, -1/(2\sigma^2))$	(y_i, y_i^2)	1
	Bernoulli	$\log(\pi/(1-\pi))$	y_i	1
Count	Poisson	$\log(\lambda)$	y_i	1
	Overdisp. Poisson	$\log(\lambda)$	y_i	ϕ
Positive Cont.	Gamma	$(-1/\theta, (k-1))$	$(y_i, \log y_i)$	1

Examples: linear model

Example: Ordinary Linear Regression

- **Random Component**:

$$y_i \sim N(\mu_i, \sigma^2)$$

- **Systematic Component**:

$$w(\boldsymbol{\eta}) = -\eta_1/(2\eta_2) = \boldsymbol{x}_i'\boldsymbol{\beta}$$

- **Link**:

$$g(\mu_i) = \mu_i = \boldsymbol{x}_i'\boldsymbol{\beta}$$

Examples: logistic regression

Example: Logistic Regression

- **Random Component**:

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

- **Systematic Component**:

$$\eta = \boldsymbol{x}_i'\boldsymbol{\beta}$$

- **Link**:

$$g(\pi_i) = \text{logit}\{\pi_i\} = \log\left\{\frac{\pi_i}{1-\pi_i}\right\} = \eta = \boldsymbol{x}_i'\boldsymbol{\beta}$$

Review of Bernoulli distribution

- $Y \sim \text{Ber}(\pi)$ has pmf

$$f(y) = \pi^y (1-\pi)^{1-y} \mathbb{I}(y \in \{0,1\})$$

Thus, $\mathbb{P}(Y=0) = 1-\pi$, $\mathbb{P}(Y=1) = \pi$; and

$$\mathbb{E}(Y) = \sum_y y\mathbb{P}(Y=y) = \pi,$$

$$\text{var}(Y) = \mathbb{E}(Y) - (\mathbb{E}(Y))^2 = \pi(1-\pi)$$

Note that var(Y) is a function of $\mathbb{E}(Y)$

- Exponential family is employed in GLM to handle the dependency of mean and variance
- Link function is used in GLM to handle the connection between the mean and the linear predictor; it is not necessarily that the mean of Y is linear in $\boldsymbol{\beta}$ but rather *some function of the mean is linear in $\boldsymbol{\beta}$*
- For logistic regression, the link function need to map the interval $(0,1)$ to the real line $\mathbb{R}$.

Probit link Function

- Probit link function is an alternative one to the logistic link function for logistic regressions
  - Let W_i denote a latent continuous variable following a simple linear regression model

$$W_i = \alpha_0 + \alpha_1 X_i + \nu_i, \quad \nu_i \sim N(0, \sigma_\nu^2)$$

  - Let Y_i denote the binary response such that, for a given threshold value c, $Y_i = \mathbb{I}(W_i \leq c)$
  - It can be shown that

$$\begin{aligned} \pi_i = \mathbb{P}(Y_i = 1) &= \mathbb{P}(W_i \leq c) = \mathbb{P}(\alpha_0 + \alpha_1 X_i + \nu_i \leq c) \\ &= \mathbb{P}\left(\frac{\nu_i}{\sigma_\nu} \leq \frac{c - \alpha_0}{\sigma_\nu} - \frac{\alpha_1}{\sigma_\nu} X_i\right) \\ &= \mathbb{P}(Z \leq \beta_0^* + \beta_1^* X_i) \\ &= \Phi(\beta_0^* + \beta_1^* X_i) \end{aligned}$$

where $\Phi(z) = P(Z \leq z)$ denote the cumulative distribution function of a standard normal distribution.

- The logistic regression model with the probit response function is

$$\mathbb{E}(Y_i) = \pi_i = \Phi(\beta_0^* + \beta_1^* X_i)$$

- The inverse function Φ^{-1} is called the probit transformation.
  - The probit link function is

$$\Phi^{-1}(\pi_i) = \beta_0^* + \beta_1^* X_i.$$

  - When $\beta_1 > 0 (< 0)$, the probit response function is monotone increasing (decreasing).
  - When $|\beta_1|$ increases, the probit response function becomes more sigmoidal shaped.
  - Increasing or decreasing β_0 shifts the probit response horizontally.
  - The probit response function has the symmetry property. When the coding of Y is reversed, the signs of all the coefficients are reversed (from $+$ to $-$ or vice versa).

Complementary Log-Log Link Function

- The complementary log-log link function is an another link function for logistic regressions
  - The complementary log-log response function is

$$\mathbb{E}(Y_i|X_i) = \pi_i = 1 - \exp(-\exp(\beta_0^G + \beta_1^G X_i))$$

    based on the cdf of a Gumbel distribution (for modeling extreme values).
  - The inverse yields a complementary log-log link function

$$\log[-\log(1-\pi_i)] = \beta_0^G + \beta_1^G X_i.$$

Logistic Link Function

- Though the logistic link function logit$(\cdot)$ has been widely used, how do we motivate it?
  - Replace the normal distribution with a logistic distribution $L(\mu, \sigma)$

$$f(x; \mu, \sigma) = \frac{\exp\{-(x-\mu)/\sigma\}}{\sigma[1+\exp\{-(x-\mu)\}]^2}$$

    where $\text{var}\{X\} = (\pi\sigma)^2/3$
  - Such a distribution is bell-shaped but more sharply peaked than Gaussian
  - Suppose $Z \sim L(0, \pi^2/3)$. Then the cumulative distribution function is

$$F_L(z) = \mathbb{P}(Z \leq z) = \frac{\exp(z)}{1+\exp(z)}.$$

- Still consider W_i a latent continuous random variables following a simple linear regression and we let $W_i = \alpha_0 + \alpha_1 X_i + \nu_i$ with $\nu_i \sim L(0, \sigma_\nu^2)$; then

$$\begin{aligned} \pi_i = \mathbb{P}(Y_i = 1) &= \mathbb{P}(W_i \leq c) = \mathbb{P}(\alpha_0 + \alpha_1 X_i + \nu_i \leq c) \\ &= \mathbb{P}\left(\frac{\pi}{\sqrt{3}} \frac{\nu_i}{\sigma_\nu} \leq \frac{\pi}{\sqrt{3}} \beta_0^* + \frac{\pi}{\sqrt{3}} \beta_1^* X_i \right) \\ &= \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)} \end{aligned}$$

- Therefore, the logistic response function is

$$\mathbb{E}(Y_i | X_i) = \pi_i = F_L(\beta_0 + \beta_1 X_i) = \frac{\exp(\beta_0 + \beta_1 X_i)}{1 + \exp(\beta_0 + \beta_1 X_i)}.$$

  where the inverse function F_L^{-1} is called the logit transformation.

- The logit link function is, as we have seen,

$$F_L^{-1}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i} \right) = \beta_0 + \beta_1 X_i.$$

- The ratio

$$\frac{\pi_i}{1 - \pi_i}$$

  is called the *odds and thus logit is referring to log odds.*

An example

- 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.
  - $y_i = 1$ if the subject i has the disease, otherwise $y_i = 0$
  - 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.

Likelihood Function

- Estimation and inference are both based on the likelihood arguments
- Since $Y_i|X_i \sim$ Bernoulli(π_i) and Y_i's are independent, the joint probability distribution function is

$$f(y_1, \ldots, y_n) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i}.$$

- The log-likelihood is therefore

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= \log f(y_1, \ldots, y_n) \\
&= \sum_{i=1}^{n} [y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)] \\
&= \sum_{i=1}^{n} \left[y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + \log(1-\pi_i) \right] \\
&= \sum_{i=1}^{n} \left[y_i \cdot \boldsymbol{x}_i'\boldsymbol{\beta} - \log\left\{1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta})\right\} \right].
\end{aligned}$$

Maximum Likelihood Estimation

- Maximize the log-likelihood function $\ell(\boldsymbol{\beta})$ to obtain the MLE $\widehat{\boldsymbol{\beta}}$
  - To maximize, we take the first-order derivatives and set them to 0.
  - However, there is no closed-form expression in general for GLM.
  - Possible approaches are
    1. Iteratively reweighted least squares
    2. Use nonlinear optimization procedure
- For GLM, in general, Fisher's scoring method is used to obtain an MLE
  - Fisher's scoring method is a variation of the Newton-Raphson algorithm in which the Hessian matrix is replaces by its expected value, which is the Fisher Information Matrix.
  - For GLM, Fisher's scoring method results in an iterative weighted least squares
  - The algorithm is presented for the general case in Section 2.5 of "Generalized Linear Models 2nd Edition" (1989) by McCullagh and Nelder
- In `R`, use `glm`

Large-Sample (Asymptotic) Properties of MLEs: Inference

- The MLEs, in general, have little or no bias.
- For sufficiently large samples, $\widehat{\boldsymbol{\beta}}$ for general GLM is approximately normal with mean $\boldsymbol{\beta}^*$ and a variance-covariance matrix that can be approximated by the estimated inverse of Fisher information matrix
  - Let $\boldsymbol{I}(\boldsymbol{\beta}^*)$ denote the $p \times p$ Fisher information matrix evaluated at $\boldsymbol{\beta}^*$

$$\boldsymbol{I}(\boldsymbol{\beta}) = -\mathbb{E}\left[\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T}\right]$$

  - Under suitable regularity conditions, as $n \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{I}(\boldsymbol{\beta}^*)^{-1})$$

- Use either Wald tests/intervals or likelihood ratio tests/profile likelihood intervals for inference.

- **Wald Test**:
  - Let $\boldsymbol{G}$ denote the $p \times p$ sample version Hessian matrix $\boldsymbol{G}(\boldsymbol{\beta}) = \dfrac{\partial^2 \ell_n(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'}$
  - The approximate variances and covariances of $\hat{\boldsymbol{\beta}}$ are obtained from

$$\widehat{\Sigma}_{\boldsymbol{\beta}} := (\widehat{\sigma}_{\boldsymbol{\beta},jk})_{1 \le j,k \le p} = \{-\boldsymbol{G}^{-1}(\widehat{\boldsymbol{\beta}})\}^{-1}.$$

1. For individual regression parameters, $\frac{\hat{\beta}_k - \beta_k}{\widehat{\sigma}_{\boldsymbol{\beta},kk}} \overset{d}{\sim} N(0,1)$ for each k.

   To test $H_0 : \beta_k = 0$ versus $H_a : \beta_k \neq 0$, use $z^* = \frac{\hat{\beta}_k}{\widehat{\sigma}_{\boldsymbol{\beta},kk}}$ and the decision rule is to reject H_0 if $|z^*| > z(1-\alpha/2)$.

   The test above is also known as the *Wald test.*
2. The approximate $1-\alpha$ confidence interval for β_k is

$$\widehat{\beta}_k \pm z(1-\alpha/2)\widehat{\sigma}_{\boldsymbol{\beta},kk}$$

3. The approximate $1-\alpha$ Bonferroni joint confidence intervals for q parameters are

$$\widehat{\beta}_k \pm z(1-\alpha/2q)\widehat{\sigma}_{\boldsymbol{\beta},kk}$$

- **Likelihood Ratio Test**:
  - Analogous to the general linear test procedure under least squares, consider likelihood ratio test under maximum likelihood.
  - Let $L(F)$ and $L(R)$ denote the likelihood function evaluated at the MLE $\widehat{\boldsymbol{\beta}}_F$ under the full model and at the MLE $\widehat{\boldsymbol{\beta}}_R$ under the reduced model, respectively.
  - The *likelihood ratio test statistic* (STAT530) is defined as

$$G^2 = -2\log\left[\frac{L(R)}{L(F)}\right] = -2[\log L(R) - \log L(F)] \overset{H_0}{\sim} \chi^2_{df_R - df_F}$$

    where $df_R = n - q$, $df_F = n - p$, and thus $df_R - df_F = p - q$.
  - The decision rule is to reject H_0 if

$$G^2 > \chi^2(1-\alpha; p-q).$$

  - Note that the likelihood ratio test can be used to test $H_0 : \beta_k = 0$. However, the results are generally not identical to those obtained from the Wald test.

Fitted Logistic Regression

- Given $\widehat{\boldsymbol{\beta}}$, the fitted logistic response is

$$\widehat{\pi}_i = \widehat{\mathbb{E}(Y_i|X_i)} = \frac{\exp(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})}{1+\exp(\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}})}.$$

  - Loosely speaking, $\widehat{\pi}_i \approx \widehat{y}_i$ as in the linear regression.

- The fitted logit response function

$$\log\left(\frac{\hat{\pi}_i}{1-\hat{\pi}_i}\right) = \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$$

Interpretation of $\widehat{\boldsymbol{\beta}}$

- The odds ratio (OR) is defined by $\pi/(1-\pi)$ and is stated as, e.g., "3" or "3:1"; and OR $= 1:1$ is a coin flip.
  - Example: Event=death, A=placebo group, B=treatment group. Then OR(death, A vs. B)$= 4$ means that the odds of death for placebo group is 4 times greater than for treatment group
- Let $\boldsymbol{x}^+ = (x_1, \ldots, x_{j-1}, x_j + 1, x_{j+1}, \ldots, x_p)'$ that $\boldsymbol{x}^+$ is the same as $\boldsymbol{x}$ except that the jth predictor has been increased by one unit.
- The odds ratio (OR) based on the logit function is

$$\begin{aligned}
[\pi^+/(1-\pi^+)]/[\pi/(1-\pi)] &= \exp\{\log([\pi^+/(1-\pi^+)]/[\pi/(1-\pi)])\} \\
&= \exp\{(\boldsymbol{x}^+)'\boldsymbol{\beta} - \boldsymbol{x}'\boldsymbol{\beta}\} \\
&= \exp\{(x_j+1)\beta_j - x_j\beta_j\} = \exp(\beta_j)
\end{aligned}$$

- Hence,

$$\frac{\pi^+}{1-\pi^+} = \exp(\beta_j)\frac{\pi}{1-\pi}.$$

- All other predictors held constants, *the odds of success at* $x_j + 1$ *are* $\exp(\beta_j)$ *times the odds of success at* x_j, which is true regardless of the initial value of x_j
- A 1 unit increase in the jth predictor with all other held constants is associate with a multiplicative change in the odds of success by the factor $\exp(\beta_j)$

Inference on $\widehat{\boldsymbol{\beta}}$

- 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, notice that

$$\pi = \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'\boldsymbol{\beta})} = \frac{1}{1+\exp(-\boldsymbol{x}'\boldsymbol{\beta})}$$

, thus the $100(1-\alpha)\%$ confidence interval for π is $([1+\exp(-L_j)]^{-1}, [1+\exp(-U_j)]^{-1})$

R demonstration

```
> d=read.table("http://www.stat.colostate.edu/~riczw/teach
              /STAT540_F15/computing/lec13/Disease.txt",header=TRUE)

> 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  1.5e-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 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
(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
```

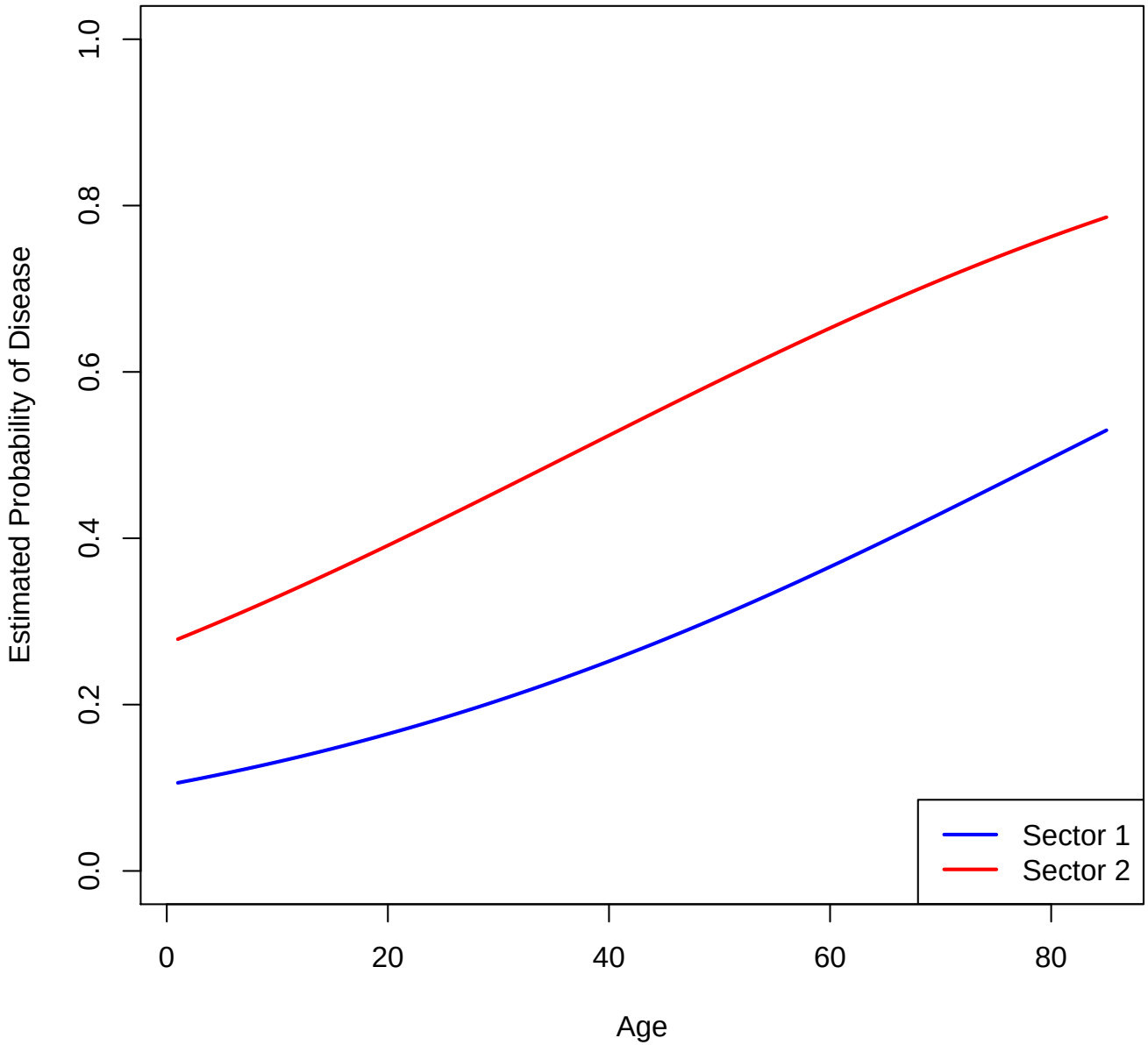
```
 > coef(o)
(Intercept)         age        ses2        ses3     sector2
-2.29393347  0.02699100  0.04460863  0.25343316  1.24363036

> 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 Pr(>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  Pr(>Chi)
NULL                     195     236.33
age     1   12.013       194     224.32 0.0005283 ***
sector  1   12.677       193     211.64 0.0003702 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
 > 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

Repeat Observations–Binomial Outcomes

- 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.table("http://www.stat.colostate.edu/~riczw/teach/
                STAT540_F15/computing/lec13/Trout.txt",header=TRUE)
> 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 for binomial data

- For $i = 1, \ldots, n$, $y_i \sim \text{Bin}(m_i, \pi_i)$ where m_i is a known number of trials for observation i, and

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

  and $y_1, \ldots, y_n$ are independent

- Recall that for $y_i \sim \text{Bin}(m_i, \pi_i)$ then $\mathbb{E}(y_i) = m_i\pi_i$ and $\text{Var}(y_i) = m_i\pi_i(1 - \pi_i)$, and

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

so that

$$\begin{aligned}\ell(\boldsymbol{\beta}|\boldsymbol{y}) &= \sum_{i=1}^{n} [y_i \log\left(\frac{\pi_i}{1 - \pi_i}\right) + m_i \log(1 - \pi_i)] + \text{const} \\ &= \sum_{i=1}^{n} [y_i \cdot \boldsymbol{x}_i'\boldsymbol{\beta} - m_i \log(1 + \exp\{-\boldsymbol{x}_i'\boldsymbol{\beta}\})] + \text{const}\end{aligned}$$

- The function $\ell(\boldsymbol{\beta}|\boldsymbol{y})$ can be maximized over $\boldsymbol{\beta} \in \mathbb{R}^p$ as discussed previously to obtain an MLE $\widehat{\boldsymbol{\beta}}$
- 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 the binomial case, there is one π_i for each y_i

Logistic regression	Saturated model
$y_i \sim \text{Bin}(m_i, \pi_i)$	$y_i \sim \text{Bin}(m_i, \pi_i)$
$y_1, \ldots, y_n$ independent	$y_1, \ldots, y_n$ independent
$\pi_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})/(1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}))$	$\pi_i \in [0,1]$ for $i = 1, \ldots, n$
p parameters	n parameters

Test for Goodness of Fit (GOF)

- Assess the overall fit of a GLM, particularly a logistic regression model, by using
  - Deviance GOF test
  - Pearson χ^2 GOF test
  - Hosmer-Lemeshow test: the idea is to create cells with counts ≥ 5 by grouping data.
- Replications at the same $\boldsymbol{x}$ are required for the Pearson χ^2 and deviance GOF tests, but not for the Hosmer-Lemeshow test.
- GOF tests provide an overall measure of the model fit.
  - GOF test can be used to detect large departures from the model (e.g. not sigmoidal shape).
  - Take logistic regression for an example, the hypothesis of interest is

$$H_0 : \mathbb{E}(Y|\boldsymbol{X}) = \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta})}{1+\exp(\boldsymbol{X}'\boldsymbol{\beta})}.$$

Deviance and GOF

- Let $\widehat{\pi}_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})/(1+\exp(\boldsymbol{x}_i'\boldsymbol{\beta}))$ denote the MLE for π_i under the logistic model; and $\widehat{\pi}_i = y_i/m_i$ is the MLE for the saturated model
- Then the likelihood ratio statistic for testing the logistic regression model as the *reduced model* versus *saturated model as the full model* is

$$D^2 := 2\sum_{i=1}^{n}[y_i \log(y/(m_i\widehat{\pi}_i)) + (m_i - y_i)\log((1-y_i/m_i)/(1-\widehat{\pi}_i))]$$

- This statistics is called the *Deviance Statistic*, the *Residual Deviance*, or just the *Deviance*.
- The statistics can be compared to χ^2_{n-p} distribution as discussed before 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

$$d_i := \text{sign}(y_i - m_i\widehat{\pi}_i)\sqrt{2[y_i \log(y/(m_i\widehat{\pi}_i)) + (m_i - y_i)\log((1-y_i/m_i)/(1-\widehat{\pi}_i))]}$$

    is called a *deviance residual*
  - So the residual deviance statistic satisfies $D^2 = \sum_{i=1}^{n} d_i^2$

Pearson χ^2 GOF Test

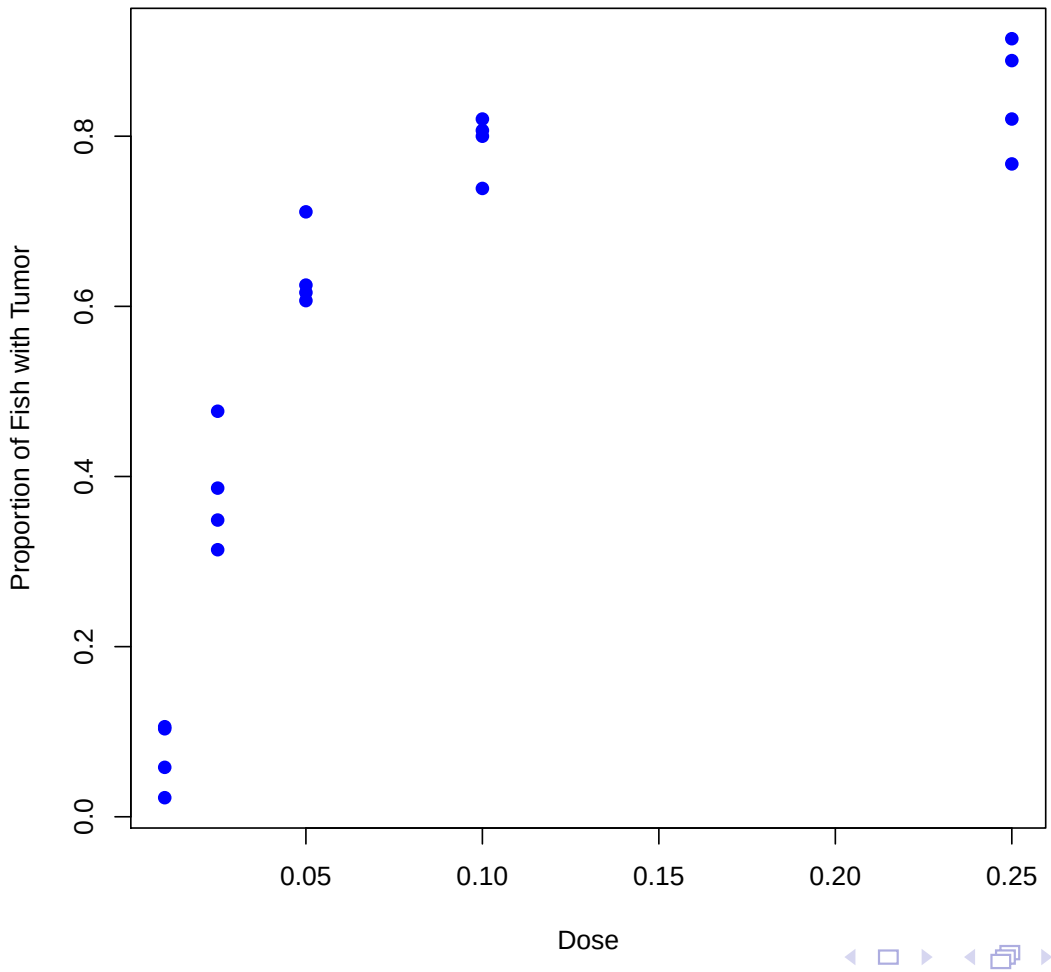
- Another goodness of fit testing statistic that is approximately χ^2_{n-p} under the null is the *Pearson chi-square Statistic*

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{y_i - m_i\widehat{\pi}_i}{\sqrt{m_i\widehat{\pi}_i(1-wh\pi_i)}} \right]^2 = \sum_{i=1}^{n} \left[\frac{y_i - \widehat{\mathbb{E}}(y_i)}{\sqrt{\widehat{\mathsf{Var}}(y_i)}} \right]^2$$

- $r_i = [y_i - m_i\widehat{\pi}_i]/[\sqrt{m_i\widehat{\pi}_i(1-wh\pi_i)}]$ is known as a *Pearson residual* and $\chi^2 = \sum_{i=1}^{n} r_i^2$
  - 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

R Example

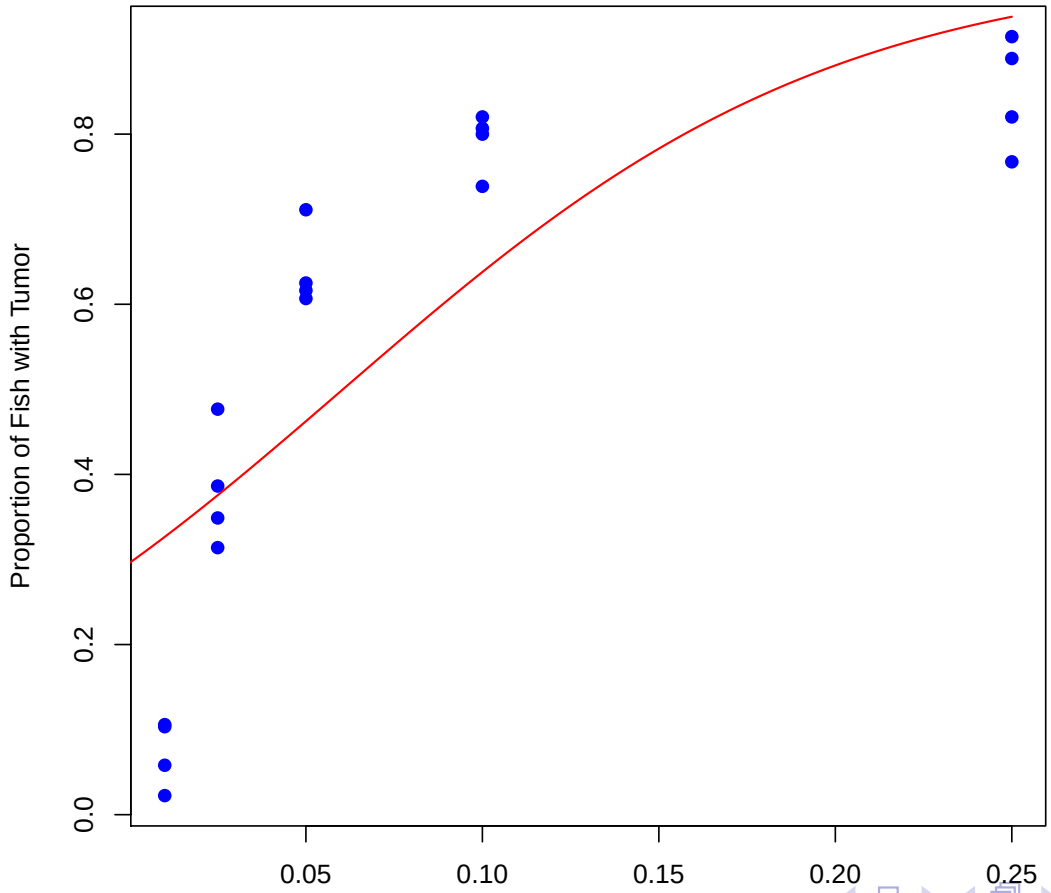
```
> #Let's plot observed tumor proportions for each tank.
> plot(d$dose,d$tumor/d$total,col=4,pch=19,xlab="Dose", ylab="Proportion of Fish with Tumor")
```



```
> #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   -11.3   <2e-16 ***
dose        14.33377    0.93695    15.3   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(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 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   -11.3   <2e-16 ***
dose        14.33377    0.93695    15.3   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(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)
```

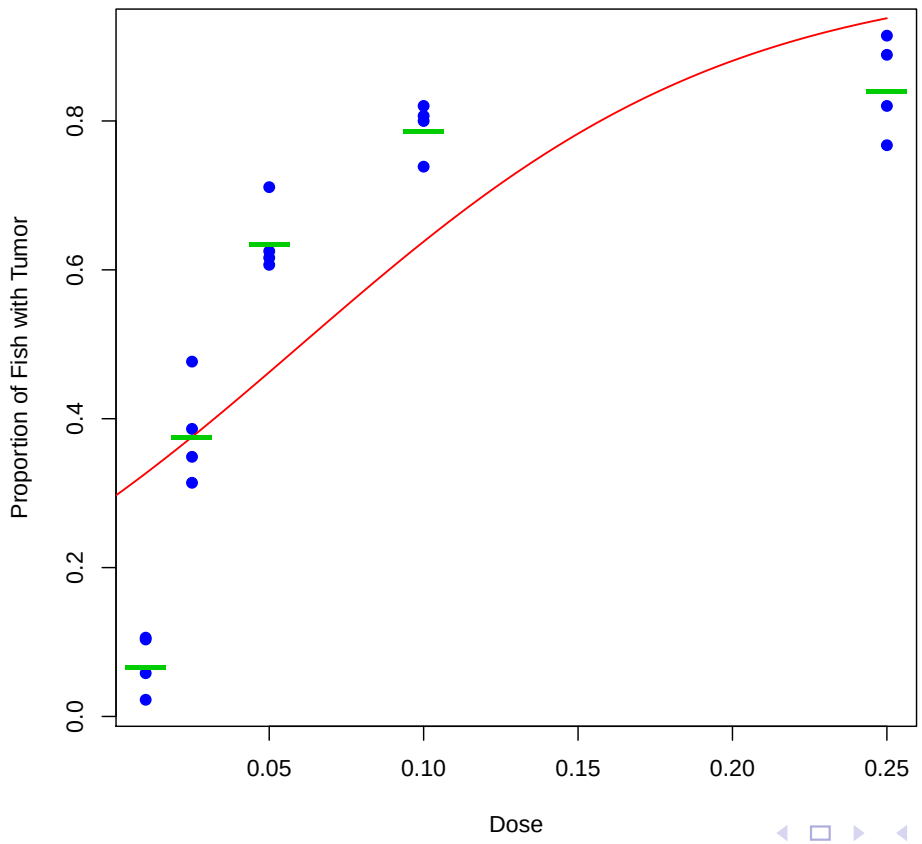


```
> #Let's use a reduced versus full model likelihood ratio test to te
> #lack of fit relative to the saturated model.
> 1-pchisq(deviance(o),df.residual(o))
[1] 0
>
> #We could try adding higher-order polynomial terms, but let's just
> #skip right to the model with dose as a categorical variable.
```

```
> 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 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(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
0.0720461 0.0720461 0.0720461 0.0720461 0.3815028 0.3815028 0.3815028
        8         9        10        11        12        13        14
0.3815028 0.6402265 0.6402265 0.6402265 0.6402265 0.7915492 0.7915492
       15        16        17        18        19        20
0.7915492 0.7915492 0.8461538 0.8461538 0.8461538 0.8461538
> points(d$dose,fitted(o),pch="_",cex=3,col=3)
```



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

Model Selection Criteria

- For GLM, AIC and BIC are commonly-used criteria

$$\begin{aligned} AIC_p &= -2\log L(\widehat{\boldsymbol{\beta}}) + 2p \\ BIC_P &= -2\log L(\widehat{\boldsymbol{\beta}}) + p\log(n) \end{aligned}$$

- Promising models have relatively small values.
- The penalty terms are $2p$ for AIC and $p\log(n)$ for BIC.
- Most software packages also report $-2\log L(\hat{\boldsymbol{\beta}})$, which always decreases as more predictor variables are added to the model.

Model Selection

- The idea of best subsets in multiple linear regression applies here.
- A best subsets procedure identifies a group of subset models that give the best values of a given criterion.
- When the number of predictor variables is large, however, all-possible best subsets may not be feasible.
- In this case, a stepwise selection procedure offers a feasible approach.
- The ideas of forward selection, backward elimination, and stepwise selection continue to apply.
- The rule for adding or deleting a predictor variable often involves a p-value from the Wald test of individual regression parameters.

Overdispersion

- Experiment 1:
  - Bucket of fair coins.
  - Each person selects a random coin.
  - Flips it 20 times. Counts H =Number of Heads.
  - Histogram the H values.
- Experiment 2:
  - Bucket is full of fair and unfair coins, with a variety of different $\mathbb{P}$[Heads].
  - Then proceed as above.
- Which histogram is wider?

- Experiment 2 has more dispersed histogram.
- Stochastic process generating the data is *not* binomial.
- "Extra-binomal variation" (general GLM term: "overdispersion")
- In the GLM framework, it is often the case that var(y_i) is a function of $\mathbb{E}(y_i)$
  - For example, in logistic regression,
    var$(y_i) = m_i\pi_i(1-\pi_i) = \mathbb{E}(y_i) - [\mathbb{E}(y_i)]^2/m_i$
  - Thus, when one fit a logistic regression model and obtain estimates of the mean of the response, we get the estimates of variance of the response (data) as well
  - If, however, the variability of our response is greater than than we should expect based on the estimated mean, we say that there is *overdispersion*.
- *If either the likelihood ratio-based or the Pearson χ^2 squared-based test of GOF (or LOF), suggests a LOF that cannot be explained by other reasons (e.g. poor model for the mean or outliers), overdispersion may be the problem.*

Quasi-likelihood

- If there is overdispersion, a quasi-likelihood approach maybe used.
- In the binomial case we make all the same assumption 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 $(n-p)^{-1} \sum_{i=1}^{n} d_i^2$ (residual deviance statistic) or $(n-p)^{-1} \sum_{i=1}^{n} r_i^2$ (Pearson chi-squared statistic)
- All analysis are as before except that
  1. The estimated variance of $\widehat{\boldsymbol{\beta}}$ is multiplied by $\widehat{\phi}$
  2. For Wald type inference, the standard normal null distribution is replaced by t_{n-p}
  3. A test statistic T that was assumed χ_q^2 under H_0 is now replaced by $T/(q\widehat{\phi})$ and compared to an F distribution with q and $n-p$ degree of freedom

- The above changes to the inference in the presence of overdispersion are analogous to the change that would take place in normal theory for Gauss-Markov linear model if we switched from assuming σ^2 was know to be 1 to assuming σ^2 was unknown and estimating it with MSE
  - Here ϕ is like σ^2 and $\widehat{phi}$ is like MSE, loosely.
- Whether there is overdispersion or not, all the usual ways of conducting GLM inference are approximate except for the special case of normal theory linear models.

R Example

```
> #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 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(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 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> #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
          [,1]
[1,] 0.1714103
```

Poisson Regression

- We have discussed the case of Bernoulli or binomial response, where logistic regression modeling is a natural GLM
- Another commonly encountered special case of generalized linear modeling involve Poisson response or Negative-Binomial response
- If $y \sim Pois(\mu)$ then $f(y) = \mu^y e^{-\mu}/y!\mathbb{I}(y \in \mathbb{Z}^+ \cup \{0\})$ and $\mathbb{E}(y) = \text{Var}(y) = \mu$
- The usual GLM for Poisson response is then, for all $i = 1, \ldots, n$

$$y_i \sim Pois(\mu_i), \text{ where } \mu_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta})$$

and $y_1, \ldots, y_n$ are independent

  - Note that $\mu_i = \exp(\boldsymbol{x}_i'\boldsymbol{\beta}) \Leftrightarrow \log(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$, hence, $g(\cdot) = \log(\cdot)$ is the link function in this case
  - All the subsequent details for the Poisson case are analogous to those we discussed for the binomial case