Binomial Regression

- Consider data where the response is discrete $Y \in \{0,1,\dots,n\}$ and represents the number of successes out of a fixed number of n independent trials.
- Here a natural model for the response is binomial $Y_i \overset{ind}{\sim} \text{Bin}(n_i, \pi_i)$ where n_i is the number of trials associated with Y_i, and $\pi_i \in (0,1)$ is the probability of success on each trial.
- We consider regression for a response $y_1, \dots, y_n$ where $y_i \overset{ind}{\sim} \text{Bin}(n_i, \pi_i)$ with probabilities π_i linked to covariates $x_i = (1, x_{i1}, \dots, x_{i,p-1})'$ and interest lies with the regression coefficients $\beta_0, \dots, \beta_{p-1}$.
- An important special case is the setting of binary data where $y_i \overset{ind}{\sim} \text{Bernoulli}(\pi_i)$ where it is of interest to relate a binary response to a set of covariates.
- We could estimate π_i as $\tilde{\pi}_i = y_i/m_i$ in which case we could estimate π_i as $\tilde{\pi}_i = y_i/m_i$ in which case we have estimated as many parameters as we have observations.
- Rather than doing this we reduce the number of unknowns $(\pi_1, \dots, \pi_n)$ by introducing a regression on the $\pi_i's$ so that they are determined by a smaller set of $p < n$ regression parameters and the covariates. This is achieved by formulating a regression model

$$\log\{\frac{\pi_i}{1-\pi_i}\} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1} x_{i,p-1}$$

  where we have gone from n unknowns $(\pi_1, \dots, \pi_n)$ to p unknowns $\beta_0, \dots, \beta_{p-1}$.
- Note that the regression model has been formulated on the scale of the log-odds $\log\{\frac{\pi_i}{1-\pi_i}\}$ which serves as the link connecting the unknown π_i's to the explanatory variables.
- One objective is then to examine how much worse the reduced model (p parameters) fits the data compared to the full model, the so called saturated model, and to determine if the model with p parameters is adequate relative to the saturated model.

$$L(\pi_1, \dots, \pi_n) = \prod_{i=1}^{n} p\,(y_i|\pi_i) \propto \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{m_i - y_i}$$

the loglikelihood takes the form

$$l(\pi_1, \dots, \pi_n) = \log L(\pi_1, \dots, \pi_n)$$
$$= \sum_i \{ y_i \log\{\frac{\pi_i}{1-\pi_i}\} + m_i \log(1-\pi_i)\}$$

- Consider a regression model of the form $\log\{\frac{\pi_i}{1-\pi_i}\} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1} x_{i,p-1}$ so that the log-likelihood

$$l(\beta_0, \dots, \beta_{p-1}) = \sum_i \{ y_i x_i{}'\beta - m_i \log\{1 + \exp\{x_i{}'\beta\}\}\}$$

we will maximize this log-likelihood function to obtain $\hat{\beta}$ and compute

$$\hat{\pi}_i = \frac{\exp\{x_i{}'\hat{\beta}\}}{1 + \exp\{x_i{}'\hat{\beta}\}}$$

the MLE's for the π_i's under the regression model.

- We want to estimate $E[Y_i] = m_i \pi_i$ under the saturated and reduced models. Under the saturated model we fit the data perfectly in the sense that $\tilde{E}[Y_i] = \tilde{\mu}_i = m_i \tilde{\pi}_i = m_i \left(\frac{y_i}{m_i}\right) = y_i$. and with the reduced (regression) model

$$\hat{E}[Y_i] = \hat{\mu}_i = m_i \hat{\pi}_i.$$

- Aside: Let $L(\theta)$ be a likelihood of an n-dimensional parameter. Let $\tilde{\theta}$ be the MLE with no constraints and $\hat{\theta}$ the MLE with regression constraints. The model with no constraints has n parameters and the reduced model has p parameters, then under the reduce model $D = -2\log\{\frac{L(\hat{\theta})}{L(\tilde{\theta})}\} \sim \chi^2_{n-p}$.

- A significance level $S.L. = Pr\left(\chi^2_{n-p} > D^{(obs)}\right)$ indicates whether the constrained model is reasonable, with low values of the p-value providing evidence against the smaller model.

- $D^{(obs)}$ is what R computes as "residual deviance". As a rule of thumb, if the model fit is adequate, then the residual deviance should be less than or on the order of the degrees of freedom $n - p$.

- We can write

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

where

$$\hat{\pi}_i = \frac{\exp\{x_i'\hat{\beta}\}}{1+\exp\{x_i'\hat{\beta}\}}$$

and for testing fit,

$$H_0: \log\{\frac{\pi_i}{1-\pi_i}\} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1}x_{i,p-1}$$

$$H_A: \text{saturated model}$$

with p-value $= Pr\left(\chi^2_{n-p} > D^{(obs)}\right)$ for testing the fit of the regression model.

- We note that we can express $D = \sum_{i=1}^{n} d_i$ where

$$d_i = 2\left[y_i \log\left(\frac{y_i}{m_i\hat{\pi}_i}\right)\right] + (m_i - y_i)\log\{\frac{m_i - y_i}{m_i(1-\hat{\pi}_i)}\}$$

now take $r_{D_i} = \text{sign}(y_i - m_i\hat{\pi}_i)\sqrt{d_i}$ which we call the deviance residual.

- Under the models assumptions we have $r_{D_i} \overset{iid}{\sim} N(0,1)$ approximately. These residuals can be used for model checking.
- As a more general form for the regression model we can write $g(\pi) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_{p-1}x_{i,p-1}$ where $g(\cdot)$ is a map $[0,1] \to \mathbf{R}$.
- Aside from the logit link $g(\pi) = \log\{\frac{\pi}{1-\pi}\}$, alternatives are the probit link $\Phi^{-1}(\pi)$ where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution, and the log-log link $g(\pi) = \log\{-\log\pi\}$.
- For comparing two groups a logistic regression would take the form

$$y_i \overset{ind}{\sim} \text{Binomial}(m_i, \pi_i)$$

$$\log\{\frac{\pi_i}{1-\pi_i}\} = \beta_0 + \beta_1 x_{1i}$$

with one covariate where

$$x_{i1} = \begin{cases} 1 & \text{subject } i \text{ in group 1} \\ 0 & \text{subject } i \text{ in group 2} \end{cases}$$

when $x_{i1} = 0$, $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 = \log\left(\frac{\pi_2'}{1-\pi_2'}\right)$ (group 2) when $x_{i1} = 1$, $\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 = \log\left(\frac{\pi_1'}{1-\pi_1'}\right)$ (group 1) where $\pi_k' = P(\text{event in group k}),\ k = 1,2$.

- We then have that $\log\left(\frac{\pi_1'}{1-\pi_1'}\right) - \log\left(\frac{\pi_2'}{1-\pi_2'}\right) = \beta_0 + \beta_1 - \beta_0 = \beta_1$ so that

$$\log\{\frac{\pi_1'}{1-\pi_1'}/\frac{\pi_2'}{1-\pi_2'}\} = \beta_1$$

a log-odds ratio comparing the likelihood of an event in group 1 relative to an event in group 2.

- As an additional example to help interpret logistic regression suppose $Y_i \overset{ind}{\sim} \text{Bin}(n_i = 1, \pi_i)$ where

$$Y_{i1} = \begin{cases} 1 & \text{subject } i \text{ had a heart attack} \\ 0 & \text{subject } i \text{ did not have a heart attack} \end{cases}$$

and

$$x_{i1} = \begin{cases} 1 & \text{smoker} \\ 0 & \text{non-smoker} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{heavy drinking} \\ 0 & \text{light drinking} \end{cases}$$

and an interaction term $x_{i3} = x_{i1}x_{i2}$.

- Consider the logistic regression

$$\log\{\frac{\pi_i}{1-\pi_i}\} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

so that $x_i' = (1, x_{i1}, x_{i2})$ and $\beta = (\beta_0, \beta_1, \beta_2)'$.

- To interpret the model consider the following:

1. Smoker (no); heavy drinking (no); $x_i = (1,0,0)' \rightarrow \frac{\pi_i}{1-\pi_i} = \exp\{\beta_0\}$
2. Smoker (no); heavy drinking (yes); $x_i = (1,0,1)' \rightarrow \frac{\pi_i}{1-\pi_i} = \exp\{\beta_0 + \beta_2\}$
3. Smoker (yes); heavy drinking (no); $x_i = (1,1,0)' \rightarrow \frac{\pi_i}{1-\pi_i} = \exp\{\beta_0 + \beta_1\}$
4. Smoker (yes); heavy drinking (yes); $x_i = (1,1,1)' \rightarrow \frac{\pi_i}{1-\pi_i} = \exp\{\beta_0 + \beta_1 + \beta_2\}$ Where we use the fact that $\frac{\pi_i}{1-\pi_i} = \exp\{x_i'\beta\}$.

- From the expressions above, among non-smokers the relative odds of a heart attack for heavy consumers versus light is

$$\frac{\exp\{\beta_0 + \beta_2\}}{\exp\{\beta_0\}} = \exp\{\beta_2\}.$$

- Among smokers the relative odds of a heart attack for heavy consumers versus light is

$$\frac{\exp\{\beta_0+\beta_1+\beta_2\}}{\exp\{\beta_0+\beta_1\}}=\exp\{\beta_2\}$$

which is the same expression we obtained for non-smokers.

- By similar methods we can see that the relative odds of heart attack for smokers versus non-smokers is $\exp\{\beta_1\}$ regardless of drinking status.
- If we consider instead the model with the interaction

$$\log\{\frac{\pi_i}{1-\pi_i}\}=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\beta_3x_{i3}$$

where now $x_i=(1,x_{i1},x_{i2},x_{i3})'$ and $\beta=(\beta_0,\beta_1,\beta_2,\beta_3)'$.

- In this case the odds ratio of a heart attack for heavy consumers is versus light consumers is

$$\frac{\exp\{\beta_0+\beta_2\}}{\exp\{\beta_0\}}=\exp\{\beta_2\}$$

whereas the corresponding odds ratio among smokers is

$$\frac{\exp\{\beta_0+\beta_1+\beta_2+\beta_3\}}{\exp\{\beta_0+\beta_1\}}=\exp\{\beta_2+\beta_3\}.$$

- Thus if $\beta_3=0$ then the effect of drinking does not depend on smoking status and vice versa. If in addition $\beta_3=0$ and $\beta_2=0$ then not only does the effect of drinking not depend on smoking, but there is no such effect.
- Example: Estimation of Prognosis for Children with Neuroblastoma

```
neuro.dat<-
read.table(file='~/Desktop/stat359/data/neuro.txt',header=TRUE,sep="")
nrow(neuro.dat)

## [1] 15

library(knitr)
kable(neuro.dat, caption = 'Neuroblastoma Data')
```

Neuroblastoma Data

age	stage	y	m
1	1	11	12
1	2	15	16
1	3	2	4
1	4	5	18

age	stage	y	m
1	5	18	19
2	1	3	4
2	2	3	7
2	3	5	8
2	4	0	25
2	5	1	3
3	1	4	5
3	2	4	12
3	3	3	15
3	4	3	93
3	5	2	5

- Purpose of study: To investigate the relationship between the probability of surviving 2 years free of disease following diagnosis and treatment for neuroblastoma, age at diagnosis and stage of disease at diagnosis.
- The data are summarized as y/m where y represents the number of patients surviving 2 years and m representing the total number of patients.

```
# age and stage are factors
neuro.dat$age<-as.factor(neuro.dat$age)
neuro.dat$stage<-as.factor(neuro.dat$stage)
# The reponse for logistic regression consists of a y/m pair
# We construct this here:
neuro.dat$resp<-cbind(neuro.dat$y,neuro.dat$m)
neuro.dat
```

```
##    age stage  y  m resp.1 resp.2
## 1    1     1 11 12     11     12
## 2    1     2 15 16     15     16
## 3    1     3  2  4      2      4
## 4    1     4  5 18      5     18
## 5    1     5 18 19     18     19
## 6    2     1  3  4      3      4
## 7    2     2  3  7      3      7
## 8    2     3  5  8      5      8
## 9    2     4  0 25      0     25
## 10   2     5  1  3      1      3
## 11   3     1  4  5      4      5
## 12   3     2  4 12      4     12
## 13   3     3  3 15      3     15
## 14   3     4  3 93      3     93
## 15   3     5  2  5      2      5
```

```r
# fit the logistic model with age and stage and print out summary statistics
model1<-glm(resp ~ age +stage,family=binomial(link=logit), data=neuro.dat)
summary(model1) # note that the first level of each factor is represented by
the intercept and is thus baseline level
```

```
##
## Call:
## glm(formula = resp ~ age + stage, family = binomial(link = logit),
##     data = neuro.dat)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.72314  -0.49391  -0.01124   0.27481   1.43268
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.2132     0.3518   0.606  0.54456
## age2         -0.7299     0.4147  -1.760  0.07844 .
## age3         -1.0799     0.3679  -2.935  0.00333 **
## stage2       -0.2759     0.4322  -0.638  0.52326
## stage3       -0.4547     0.5173  -0.879  0.37943
## stage4       -2.2772     0.5055  -4.505 6.65e-06 ***
## stage5       -0.2636     0.4476  -0.589  0.55602
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 77.4617  on 14  degrees of freedom
## Residual deviance:  8.1785  on  8  degrees of freedom
## AIC: 62.775
##
## Number of Fisher Scoring iterations: 4
```

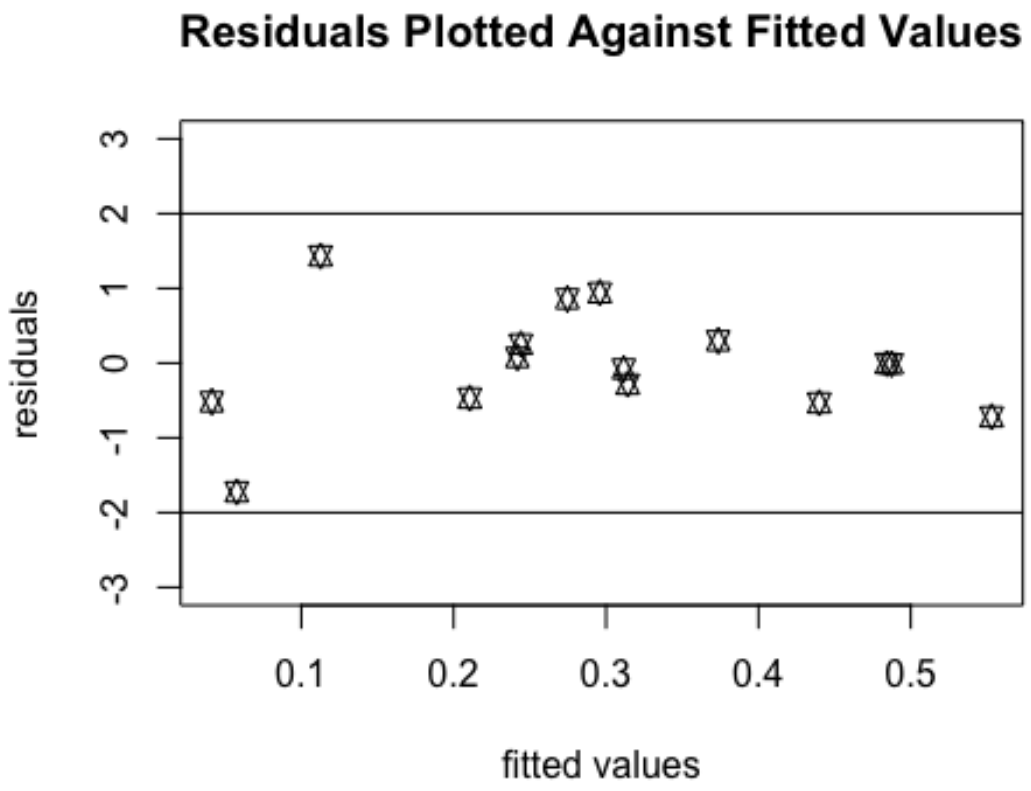
1. age (x_1,x_2); stage (IV); $x' = (1, x_1, x_2, 0,0,1,0)$; $\frac{\pi}{1-\pi} = \exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_5\}$
2. age (x_1,x_2); stage (I); $x' = (1, x_1, x_2, 0,0,0,0)$; $\frac{\pi}{1-\pi} = \exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2\} \rightarrow$ odds ratio comparing stage (IV) to stage (I) is

$$\frac{\exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_5\}}{\exp\{\beta_0 + \beta_1 x_1 + \beta_2 x_2\}} = \exp\{\beta_5\}.$$

- The estimated odds ratio is $\exp\{-2.277\} = 0.104 \rightarrow$ odds of surviving in stage IV are much lower, only about 10% of the odds of survival in stage I.
- 95% confidence interval (CI) for $\exp\{\beta_5\}$, first find 95% for β_5: $\hat{\beta}_5 \pm 1.96 * 0.5055 = (-3.27, -1.286)$ 95% CI for the odds ratio comparing survival in stage IV relative to stage I.

- Exponentiating this we have (0.038,0.276) 95% CI for the odds ratio comparing survival in stage IV relative to stage I.

```
#here we record the deviance residuals, linear predictor, and fitted values
rd1<-residuals.glm(model1,"deviance")
lp1<-model1$linear.predictors
fv1<-model1$fitted.values
plot(fv1,rd1,ylim=c(-3,3),xlab='fitted values',ylab='residuals',pch=11)
abline(h=-2)
abline(h=2)
title('Residuals Plotted Against Fitted Values')
```




```
# Here we fit two reduced models to enable us to test the importance of age
and stage
model2<- glm(resp ~ age,family=binomial(link=logit), data=neuro.dat)
model3 <- glm(resp ~stage,family=binomial(link=logit), data=neuro.dat)
```

```
#Test the significance of stage using a likelihood ratio test
# The test statistic
D.obs<-model2$deviance - model1$deviance
# Under the null hypothesis this came from a chi-squared distn with 4 DOF
1-pchisq(D.obs,4)
```

```
## [1] 1.213232e-06
```

```
anova(model2,model1,test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: resp ~ age
## Model 2: resp ~ age + stage
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        12     41.145
## 2         8      8.179  4   32.967 1.213e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#Test the significance of age
D.obs<-model3$deviance - model1$deviance
# Under the null hypothesis this came from a chi-squared distn with 2 DOF
1-pchisq(D.obs,2)
```

```
## [1] 0.009009017
```

```
# or equivalently
anova(model3,model1,test='Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: resp ~ stage
## Model 2: resp ~ age + stage
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        10    17.5976
## 2         8     8.1785  2   9.4191 0.009009 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Conclude that both age and stage play significant roles in determining prognosis, even after controlling for the other variable.