

Modeling a Binomial Response

Bruce A Craig

Department of Statistics
Purdue University

Reading: Faraway Ch. 3-4, Agresti Ch. 5-6, KNNL Ch. 14

Binomial GLM

- Each Y_i now the result of multiple Bernoulli trials
 - $Y_i := \sum_{j=1}^{m_i} Y'_{ij}$, where $\{Y'_{ij}\} \stackrel{ind}{\sim} \text{Bernoulli}(p_i)$
 - $\mathbf{x}_i$: predictor values for observation i
 - m_i : # of Bernoulli trials for observation i
- GLM Model:

$$Y_i \stackrel{ind}{\sim} B(m_i, p_i)$$

$$\text{logit}(p_i) = \mathbf{x}_i \boldsymbol{\beta}$$

- Log-Likelihood:

$$\begin{aligned} l(\boldsymbol{\beta}) &= \log \prod_{i=1}^n \binom{m_i}{Y_i} p_i^{Y_i} (1 - p_i)^{m_i - Y_i} \\ &= \sum Y_i (\mathbf{x}_i \boldsymbol{\beta}) - m_i \log(1 + \exp\{\mathbf{x}_i \boldsymbol{\beta}\}) + \log \binom{m_i}{Y_i} \end{aligned}$$

Equivalency of GLMs

- Bernoulli and binomial log-likelihoods differ by a constant

$$\begin{aligned}\log(L)^{Binomial} &= \sum_{i=1}^n \left[Y_i \log \left(\frac{p_i}{1-p_i} \right) + m_i \log(1-p_i) + \log \binom{m_i}{Y_i} \right] \\&= \sum_{i=1}^n \left[\sum_{j=1}^{m_i} \left\{ Y'_{ij} \log \left(\frac{p_i}{1-p_i} \right) + \log(1-p_i) \right\} + C_i \right] \\&= \sum_{i=1}^n \sum_{j=1}^{m_i} \left\{ Y'_{ij} \log \left(\frac{p_i}{1-p_i} \right) + \log(1-p_i) \right\} + \text{constant} \\&= \log(L)^{Bernoulli} + \text{constant}\end{aligned}$$

- Both models lead to same parameter estimates and inferences, but have different deviances.

Example

- Yes, that O-ring study again.
 - In 1986, space shuttle Challenger exploded after launch
 - Investigation launched that focused on rubber O-rings
 - Cold temperatures make rubber more brittle
 - Is O-ring less effective sealant when cold?
-
- 23 past missions, each with 6 O-rings
 - Does the chance of damage decrease with temperature?

Fitting a Simple Logistic Model

```
> modelo = glm(cbind(damage,6-damage)~orings$temp,family=binomial,orings)
> summary(modelo)
```

Call:

```
glm(formula = cbind(damage, 6 - damage) ~ orings$temp, family = binomial,
    data = orings)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	11.66299	3.29626	3.538	0.000403 ***
orings\$temp	-0.21623	0.05318	-4.066	4.78e-05 ***

Signif. codes:

0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 38.898 on 22 degrees of freedom
Residual deviance: 16.912 on 21 degrees of freedom
AIC: 33.675

Number of Fisher Scoring iterations: 6

Results: Goodness of fit

- Note: Model response is in terms of Y_i and $m_i - Y_i$
- Can use goodness-of-fit measures to assess overall fit
 - Binomial deviance

```
> pchisq(deviance(model0),df.residual(model0),lower=FALSE)
[1] 0.7164099
```
 - Pearson X^2

```
> pchisq(sum(residuals(model0,type="pearson")^2),21,lower=FALSE)
[1] 0.1382507
```
- Both convey same conclusion, but P -values very different
- Should we trust one more than the other?

Simulation Study

- What is the distribution under the null hypothesis?
- Can use simulation to approximate sampling dist

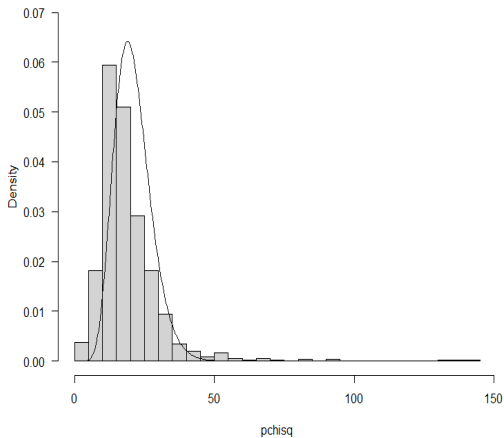
```
nsamp=23
x1 = orings$temp
for(i in 1:1000){

  logitp = 11.663 -0.2162*x1    # the fitted equation
  y = rbinom(nsamp,6,(1+exp(-logitp))^(1))
  model1 = glm(cbind(y,6-y)~x1,family=binomial)
  pchisq[i] = sum(residuals(model1,type = "pearson")^2)
  dev[i] = summary(model1)$deviance
}

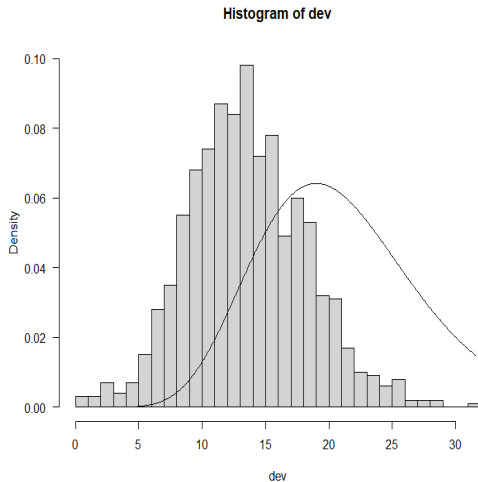
hist(pchisq,prob=T,nclass=25,ylim=c(0,.07),las=1)
lines(seq(min(pchisq),max(pchisq),length=500),dchisq(
  seq(min(pchisq),max(pchisq),length=500),nsamp-2))
```

Pearson χ^2

Histogram of pchisq



Deviance



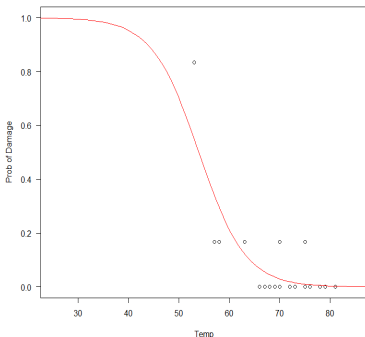
Summarizing the Results

- Note: Model response is in terms of Y_i and $m_i - Y_i$
- χ^2 approximation not reasonable for both measures
- Can use estimated sampling distribution
 - Deviance: $P = 0.237$ (Still same conclusion)
 - Pearson: $P = 0.133$
- Test of β_1 suggests negative relationship with temp
- For each °F increase, the OR decreases 19.5%

```
> exp(coefficients(modelo)[2])
orings$temp
0.8055471
> exp(confint(modelo)[2,])
Waiting for profiling to be done... *** Uses Profile likelihood CI
      2.5 %      97.5 %
0.7170161 0.8867617
```

Visualizing the Fit

```
> library(ggplot2)
> plot(damage/6~orings$temp,orings,ylim=c(0,1),xlim=c(25,85),
      xlab="Temp",ylab="Prob of Damage",las=1)
> xt =seq(25,85,.5)
> lines(xt,ilogit(11.663-0.2162*xt), col="red")
```



Analyze Using Binary Form

```
> borings = with(orings,data.frame(temp=rep(orings$temp,each=6),
      damage=as.vector(sapply(orings$damage,function(x) rep(c(0,1),
        times=c(6-x,x))))))
> modelob = glm(damage~temp, family=binomial, borings)
> summary(modelob)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	11.66299	3.29616	3.538	0.000403	***
temp	-0.21623	0.05318	-4.066	4.77e-05	***

Signif. codes:

0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 76.745 on 137 degrees of freedom
Residual deviance: 54.759 on 136 degrees of freedom
AIC: 58.759

Number of Fisher Scoring iterations: 6

Prediction Interval $Y(\mathbf{x})$

- Topic 4 covered general approach to predict mean $\mu(\mathbf{x})$
- For the binomial distribution $\mu(\mathbf{x}) = mp(\mathbf{x})$:
 - CI for $p(\mathbf{x})$: $g^{-1} \left(\mathbf{x}\hat{\beta} \pm z_{\alpha/2} \sqrt{\mathbf{x}\text{Cov}(\hat{\beta})\mathbf{x}'} \right)$
 - For logit link, $g^{-1} = \exp\{\eta\} / (1 + \exp\{\eta\})$
 - CI for $\mu(\mathbf{x})$ is then $m \times$ the CI for $p(\mathbf{x})$
- When predicting new Y , we are assuming $Y \sim B(m, \hat{p})$
- Can use the $(\alpha/2, 1 - \alpha/2)$ quantiles of distribution for interval but using $\hat{p}$ ignores its uncertainty
- Using upper and lower endpoints of CI for p in quantiles results in interval that is often too wide

Model Misspecification

- If binomial model correct, residual deviance $\sim \chi^2$
- If deviance too large, must determine why incorrect
- Most commonly due to the wrong structural form or the inclusion of outliers. Both can be investigated using common diagnostic tools
- Another possibility is that the binomial dist is not correct
- Binomial distribution assumes a set of Bernoulli trials that
 - are independent
 - have a common probability p_i
- Violations of these distributional assumptions can lead to **overdispersion**

Overdispersion - Correlated Trials

- If $Y \sim B(m, p)$ then $Y = \sum Y'_i$ where $Y'_i \stackrel{ind}{\sim} \text{Bernoulli}(p)$
- Suppose $Y'_1, Y'_2, \dots, Y'_m$ are not independent
- Instead, for all pairs $\text{Corr}(Y'_i, Y'_j) = \rho > 0$

$$\begin{aligned}\text{Var}(Y) &= \sum_{i=1}^m \text{Var}(Y'_i) + \sum_{i \neq j} \text{Corr}(Y'_i, Y'_j) \sqrt{\text{Var}(Y'_i) \text{Var}(Y'_j)} \\ &= mp(1-p) + m(m-1)\rho p(1-p) \\ &= mp(1-p)(1 + (m-1)\rho)\end{aligned}$$

- Variance exceeds binomial variance by factor $(1 + (m-1)\rho)$

Overdispersion - Nonconstant p

- Consider there to be I indep clusters of size $k \rightarrow m = Ik$
- For each cluster, $Z_I \sim B(k, p_I)$ and $p_I \overset{\text{ind}}{\sim} \text{Beta}(p, \tau^2)$
- The mean and variance of $Y = \sum Z_I$ are

$$\begin{aligned} E(Y) &= E\left(\sum Z_I\right) = \sum E(E(Z_I|p_I)) = \sum kE(p_I) = mp \\ \text{Var}(Y) &= \sum \text{Var}(Z_I) \\ &= \sum \{E(\text{Var}(Z_I|p_I)) + \text{Var}(E(Z_I|p_I))\} \\ &= \sum \{E(kp_I(1 - p_I)) + \text{Var}(kp_I)\} \\ &= \sum \{k(E(p_I) - (\text{Var}(p) + E^2(p_I))) + k^2\text{Var}(p_I)\} \\ &= mp(1 - p) [1 + (k - 1)\tau^2] \end{aligned}$$

- The variance again exceeds the variance of the binomial

Overdispersion - Nonconstant p

- Clustering idea leads to beta-binomial distribution
- Each $Y|p \sim B(m, p)$ and $p \stackrel{ind}{\sim} \text{Beta}(p, \tau^2)$
- Integrate out p , the mean and variance of Y are

$$\begin{aligned} E(Y) &= mp \\ \text{Var}(Y) &= mp(1-p)((1 + (m-1)\tau^2)) \end{aligned}$$

- Variability increases with m rather than a constant

Modeling Overdispersion in GLM

- Could consider introducing a dispersion parameter so that

$$E(Y_i) = m_i p_i, \text{ Var}(Y_i) = \phi m_i p_i (1 - p_i)$$

- Can estimate ϕ using $\hat{\phi} = \chi^2 / (n - p)$
- Because ϕ does not affect $E(Y_i)$, modeling overdispersion does not change $\hat{\beta}$
- $\text{Cov}(\hat{\beta})$ is scaled up by $\hat{\phi}$
- Beta-binomial distribution has the property that $\text{Var}(Y)/mp(1 - p)$ is a linear function of m , not a constant as it is here.

Comparing Nested Models

- Regular likelihood-based approaches are not applicable because of the unknown scaling factor
- F test approximates deviance-based LR test

$$F = \frac{(D_R - D_F)/(df_R - df_F)}{\hat{\phi}} \underset{asympt, H_0}{\sim} F_{df_R - df_F, df_F}$$

- Modeling strategy:
 - fit the full model (with all predictors)
 - estimate $\hat{\phi}$
 - compare nested models with F test to reduce the number of predictors

Example

- Experiment studied the survival of trout eggs
- Treatment factors
 - Stream location (5 levels)
 - Weeks at location (4 levels)
- For each of the 20 factor combinations, a box of eggs was buried and then later retrieved
- The number of surviving eggs and total eggs were recorded

The Data

```
> ftable(xtabs(cbind(survive,total)~location+period,troutegg))
```

		survive	total	
location	period			
1	4	89	94	
	7	94	98	
	8	77	86	
	11	141	155	
2	4	106	108	
	7	91	106	
	8	87	96	
	11	104	122	
3	4	119	123	
	7	100	130	
	8	88	119	
	11	91	125	
4	4	104	104	****All survived
	7	80	97	
	8	67	99	
	11	111	132	
5	4	49	93	
	7	11	113	
	8	18	88	
	11	0	138	****None survived

Fitting Binomial GLM

```
> bmod = glm(cbind(survive,total-survive)~location+period,  
             family=binomial, troutegg)  
> summary(bmod)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	4.6358	0.2813	16.479	< 2e-16 ***
location2	-0.4168	0.2461	-1.694	0.0903 .
location3	-1.2421	0.2194	-5.660	1.51e-08 ***
location4	-0.9509	0.2288	-4.157	3.23e-05 ***
location5	-4.6138	0.2502	-18.439	< 2e-16 ***
period7	-2.1702	0.2384	-9.103	< 2e-16 ***
period8	-2.3256	0.2429	-9.573	< 2e-16 ***
period11	-2.4500	0.2341	-10.466	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

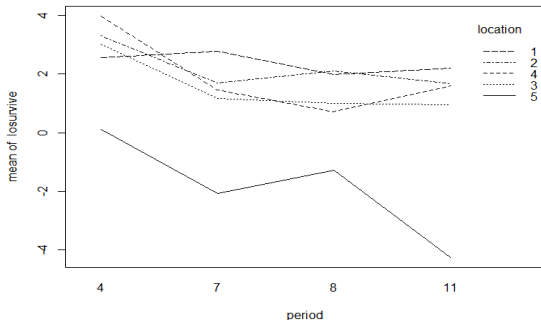
Null deviance: 1021.469 on 19 degrees of freedom
Residual deviance: 64.495 on 12 degrees of freedom
AIC: 157.03

Summarizing the Results

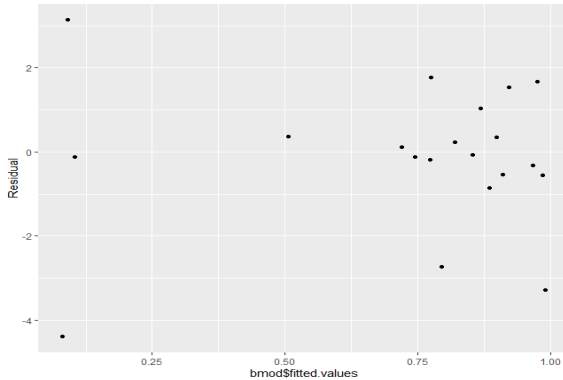
- Residual deviance is 64.5 on 12 df
- This is too big and suggests poor model fit
- It is due to overdispersion?
- First need to rule out the other options
 - We have large number of eggs per box, so the χ^2 approximation should be reasonable
 - Need to check for outliers
 - Could there be an interaction between factors?
- Can use randomized quantile residuals to investigate

Visualizing the Fit

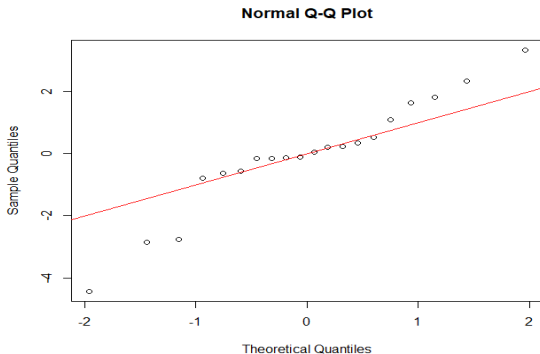
```
> losurvive = with(troutegg,logit((survive+2)/(total+4)))  
> with(troutegg,interaction.plot(period,location,losurvive))
```



Residual Plot



Normal QQplot



Summarizing the Results

- Interaction plot suggests possible interaction (will ignore)
- Residual plot does not show any obvious outliers
- Spread in residuals much larger than ± 2 as expected
- QQplot suggests approx Normality but larger variance
- Using Pearson Goodness of Fit, $\hat{\phi} = 5.33$

```
> phihat = sum(residuals(bmod,type="pearson")^2)/df.residual(bmod)
> phihat
[1] 5.330322
```

- Need to test significance of factors using F tests

F Tests

```
> drop1(bmod,scale=phihat,test="F")
```

Single term deletions

Model:

```
cbind(survive, total - survive) ~ location + period
```

scale: 5.330322

	Df	Deviance	AIC	F value	Pr(>F)
<none>		64.50	157.03		
location	4	913.56	308.32	39.494	8.142e-07 ***
period	3	228.57	181.81	10.176	0.001288 **

Signif. codes:

0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Warning message:

```
In drop1.glm(bmod, scale = phihat, test = "F") :
```

F test assumes 'quasibinomial' family

t Tests

```
> summary(bmod,dispersion=phihat)
```

Call:

```
glm(formula = cbind(survive, total - survive) ~ location + period,  
     family = binomial, data = troutegg)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	4.6358	0.6495	7.138	9.49e-13	***
location2	-0.4168	0.5682	-0.734	0.4632	
location3	-1.2421	0.5066	-2.452	0.0142	*
location4	-0.9509	0.5281	-1.800	0.0718	.
location5	-4.6138	0.5777	-7.987	1.39e-15	***
period7	-2.1702	0.5504	-3.943	8.05e-05	***
period8	-2.3256	0.5609	-4.146	3.38e-05	***
period11	-2.4500	0.5405	-4.533	5.82e-06	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 5.330322)

Quasi-binomial GLM

- When $\phi \neq 1$, the Y_i follow a *quasi-binomial distribution*
 - $E(Y_i) = \mu_i$ and $\text{Var}(Y_i) = \phi V(\mu_i)$
 - But the pdf of Y_i is unspecified (cannot simulate)
 - GLM estimation requires only mean and variance information so this is possible
- For inference, replacing likelihood with **quasi-likelihood**
- Using the scores

$$U_i = \frac{Y_i - \mu_i}{\phi V(\mu_i)}$$

in place of l' so log quasi-likelihood defined as

$$Q_i = \int_{Y_i}^{\mu_i} \frac{Y_i - t}{\phi V(t)} dt$$

- This idea of specifying only the mean and variance can be used with a wide range of response types

Zero-Inflated Binomial

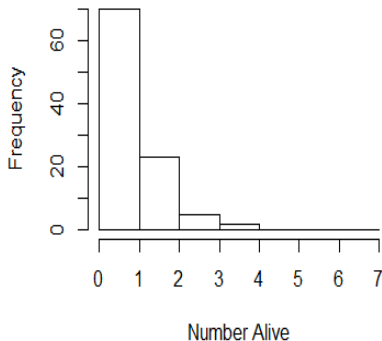
- Refers to there being an excess of zero counts
- Alternative cause for overdispersion
- Assuming there to be a second process generating 0's
- When p small, sometimes not easy to detect
 - Example: Number of live insects after pesticide application
- Can refer to any over-observed value but 0 most common

Example: Comparing Two Application Methods

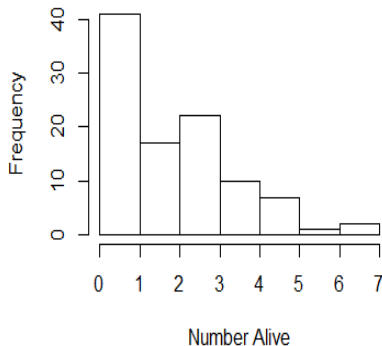
- Consider that there is a pesticide that can be applied:
 - 1 Directly to the plant in powder form
 - 2 Mixed in water and sprayed on the plant
- Experiment run in a greenhouse using 200 plants
- Each plant infested with 30 bugs
- Response: Number alive after 24 hours

Visualizing the Data

Method 1



Method 2



Initial Analyses

- Can we just compare means using a t test?
- Is $n = 100$ per group enough for CLT to apply here?
- Is the mean the right summary to compare?

```
> t.test(live, trt)
```

Welch Two Sample t-test

```
data: live and trt
```

```
t = 0.39309, df = 240.73, p-value = 0.6946
```

```
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
```

```
-0.1805081  0.2705081
```

```
sample estimates:
```

```
mean of x mean of y
```

```
1.545
```

```
1.500
```

```
***Estimated p's 0.0515 versus 0.0500
```

Initial Analyses

- Fit GLM to account for non-negative, integer response

```
> mod1 = glm(cbind(live,total-live)~as.factor(trt),family=binomial,pest)
> summary(mod1)
```

Call:

```
glm(formula = cbind(live, total - live) ~ as.factor(trt),
     family = binomial, data = pest)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.3367	0.1003	-33.277	< 2e-16 ***
as.factor(trt)2	0.7293	0.1236	5.903	3.57e-09 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 358.67 on 199 degrees of freedom
Residual deviance: 321.80 on 198 degrees of freedom
AIC: 665.14

Results of GLM

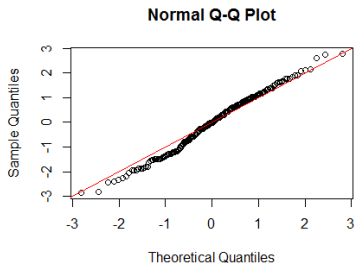
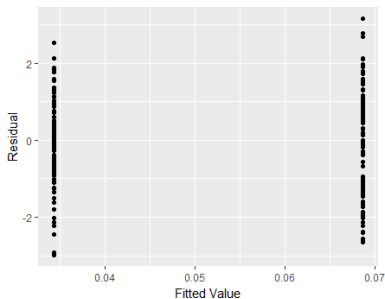
- Now a highly significant treatment difference
- Estimated proportions alive

$$\begin{aligned}\hat{p}_1 &= \frac{\exp\{-3.3367\}}{1 + \exp\{-3.3367\}} \\ &= 0.034 \\ \hat{p}_2 &= \frac{\exp\{-3.3367 + 0.7293\}}{1 + \exp\{-3.3367 + 0.7293\}} \\ &= 0.067\end{aligned}$$

- Residual deviance, however, suggests lack of fit

```
> pchisq(321.8,198,lower=FALSE)
[1] 6.139839e-08
```

Diagnostics



- Spread in residuals a little larger than expected

Quasi binomial?

- Just assume quasi-binomial and be done with it?

```
> phihat = sum(residuals(mod1,type="pearson")^2)/df.residual(mod1)
> phihat
[1] 1.338789
> summary(mod1,dispersion=phihat)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-3.3367	0.1160	-28.760	< 2e-16 ***
as.factor(trt)2	0.7293	0.1430	5.102	3.37e-07 ***

Signif. codes:

0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

(Dispersion parameter for binomial family taken to be 1.338789)

- Let's look more closely at binomial fit

Examine the Count Data

```
> bc <- table(trt, live)
> bc
      live
trt  0  1  2  3  4  5  6  7
  1 36 34 23  5  2  0  0  0
  2 30 11 17 22 10  7  1  2

> grp = c(1:4, rep(5, length=max(live)-3))
> bcc = t(rowsum(t(bc), grp))
> bcc
      0  1  2  3  4
1 36 34 23  5  2
2 30 11 17 22 20
```

###The expected counts are

	0	1	2	3	4
dist1	35.06043	37.39618	19.27897	6.397487	1.866361
dist2	11.83462	26.17680	27.98500	19.257633	14.672619

###Struggling to fit the low count values in Method 2

Zero-Inflated Binomial Model

- Additional process represented as a point mass

$$P(Y_i = y_i) = \begin{cases} p_i^* + (1 - p_i^*)B(0; m_i, p_i) & \text{for } y_i = 0 \\ (1 - p_i^*)B(y_i; m_i, p_i) & \text{for } y_i > 0 \end{cases}$$

with

$$\text{logit}(p_i) = \mathbf{x}_i\boldsymbol{\beta}$$

$$\text{logit}(p_i^*) = \mathbf{x}_i\boldsymbol{\gamma}$$

- Estimates obtained via MLE or Bayesian inference
- In R: `zibinomialff` or `zibinomial` from VGAM package

Fitting the Model

```
mod2 = vglm(cbind(live,total-live)~as.factor(trt),lprob="logitlink",
            lonempstro0="logitlink",
            zibinomialff,data=pest,trace=TRUE)
summary(mod2)
```

Call:

```
vglm(formula = cbind(live, total - live) ~ as.factor(trt),
     family = zibinomialff, data = pest, lprob = "logitlink",
     lonempstro0 = "logitlink",trace = TRUE)
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept):1	-3.1338	0.1216	-25.769	< 2e-16	***
(Intercept):2	1.3087	0.2496	5.244	1.57e-07	***
as.factor(trt)2	0.8362	0.1402	5.965	2.44e-09	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-likelihood: -313.6055 on 397 degrees of freedom

Warning: Hauck-Donner effect detected in the following estimate(s):
'(Intercept):1'

Assessing Fit

```
zbccoef = coef(mod2)
p0 = 1/(1+exp(zbccoef[2]))
p1 = 1/(1+exp(-zbccoef[1]))
bc1cz[1,] = c(p0+(1-p0)*dbinom(0,30,p1), (1-p0)*dbinom(1,30,p1),
              (1-p0)*dbinom(2,30,p1), (1-p0)*dbinom(3,30,p1),
              (1-p0)*(1-pbinom(3,30,p1)))
```

```
p1 = 1/(1+exp(-(zbccoef[1]+zbccoef[3])))
bc1cz[2,] = c(p0+(1-p0)*dbinom(0,30,p1), (1-p0)*dbinom(1,30,p1),
              (1-p0)*dbinom(2,30,p1), (1-p0)*dbinom(3,30,p1),
              (1-p0)*(1-pbinom(3,30,p1)))
```

```
#### p0 = 0.213           p1 = 0.042 or 0.091
```

```
###The expected counts are
```

	1	2	3	4	5
dist1	43.18459	28.63159	18.08069	7.349425	2.753709
dist2	25.72122	13.41891	19.55486	18.342605	22.962407

Prospective Study

- In a prospective (or cohort) study: predictors are fixed and the outcomes are observed
 - Example 1: Recruit subjects of 2 genotypes and observe the occurrence of a rare disease
 - Example 2: Recruit infants breast or bottle fed and observe occurrence of food allergies
 - Must wait before collecting outcomes
 - Can require large samples for rare diseases
 - Can study more than one outcome
 - Less susceptible to bias

Retrospective Study

- In a retrospective (or case-control) study: outcome is fixed, observe the predictors
 - Example 1: Recruit patients with and without disease, observe their genotypes
 - Example 2: Recruit infants with and without food allergies, observe how they were fed
 - Can answer immediately; Better handles rare diseases
 - Can investigate many predictors but only one outcome
 - Potential bias in the recruitment of “controls”

Analysis via Logistic Regression

- With logit link, log odds are the same for both studies
- Consider a prospective model where

$$p(\mathbf{x}_i) = P(Y_i = 1|\mathbf{x}_i) = \exp\{\mathbf{x}_i\beta\} / (1 + \exp\{\mathbf{x}_i\beta\})$$

but the subjects are selected retrospectively.

- Let $Z_i = 1_{\{\text{study includes subject } i\}}$ be conditional on Y with $\theta_0 = P(Z_i = 1|Y_i = 0)$ and $\theta_1 = P(Z_i = 1|Y_i = 1)$
- Note: θ_0 and θ_1 are independent of X_i , otherwise introduce selection bias
- Of interest in retrospective study is $P(Y_i = 1|X_i, Z_i = 1)$

Analysis via Logistic Regression

- With retrospective sampling,

$$\begin{aligned}P(Y_i = 1|\mathbf{X}_i, Z_i = 1) &= \frac{P(Y_i = 1, Z_i = 1|\mathbf{X}_i)}{P(Z_i = 1|\mathbf{X}_i)} \\&= \frac{P(Y_i = 1, Z_i = 1|\mathbf{X}_i)}{P(Y_i = 0, Z_i = 1|\mathbf{X}_i) + P(Y_i = 1, Z_i = 1|\mathbf{X}_i)} \\&= \frac{\theta_1 \times p(\mathbf{X}_i)}{\theta_0(1 - p(\mathbf{X}_i)) + \theta_1 p(\mathbf{X}_i)}\end{aligned}$$

- Uncover the same β as in the prospective study

$$\Rightarrow \log \left\{ \frac{P(Y_i = 1|\mathbf{X}_i, Z_i)}{P(Y_i = 0|\mathbf{X}_i, Z_i)} \right\} = \log \left(\frac{\theta_1 p(\mathbf{X}_i)}{\theta_0(1 - p(\mathbf{X}_i))} \right) = \log(\theta_1/\theta_0) + \mathbf{X}_i\beta$$

Matched Case-Control Study

- In a case-control study, often want to focus on a small number of risk factors but expect known nuisance factors to also affect outcome
- One approach to address this is to **match** a case and control based on these nuisance factors
- Similar to blocking in experimental design
- Can be 1 to 1 or 1 to M matching
- Will use matching algorithm or propensity scoring to determine the matched groups

Conditional Logistic Regression

- For each group j , consider the logistic model

$$\text{logit}(p_j(\mathbf{x}_{ij})) = \alpha_j + \mathbf{x}_{ij}\beta$$

$i = 0$ for the case and $i = 1, 2, \dots, M$ for the controls

- Consider the conditional probability $P(Y_{0j} = 1, Y_{ij} = 0 | \mathbf{x}_{0j}, \mathbf{x}_{ij})$

$$\begin{aligned} &= \frac{P(Y_{0j} = 1 | \mathbf{x}_{0j})P(Y_{ij} = 0 | \mathbf{x}_{ij})}{P(Y_{0j} = 1 | \mathbf{x}_{0j})P(Y_{ij} = 0 | \mathbf{x}_{ij}) + P(Y_{0j} = 0 | \mathbf{x}_{0j})P(Y_{ij} = 1 | \mathbf{x}_{ij})} \\ &= \frac{\exp\{\mathbf{x}_{0j}\beta\}}{\exp\{\mathbf{x}_{0j}\beta\} + \exp\{\mathbf{x}_{ij}\beta\}} \end{aligned}$$

- The different α_j 's cancel out
- In R: clogit function within survival package

Beta Regression

- Focus, so far, has been on proportions that result from a set of m_i trials
- Can be situations of proportional outcomes that are not the result of a set of trials
 - Percent of a leaf consumed by a larvae
 - Yield of output (as percent of input)
- Beta: $E(Y_i) = \mu_i$ and $\text{Var}(Y_i) = \mu_i(1 - \mu_i)/(1 + \phi)$
- Most common to use logit link
- Similar to quasi-binomial but a full distributional model
- Can be used as approximate model for proportion of trials
 - Best when m_i reasonably large and relatively constant
- In R: gam function in mgcv package