Book 2, Chapter 2: Binomial Data

Stats 500, Fall 2015
Brian Thelen, University of Michigan
443 West Hall, bjthelen@umich.edu

Review: The Binomial Distribution

- n independent trials $Z_1, \ldots, Z_n$
- $P(Z_i = 1) = p$ (**"success"**)
  $P(Z_i = 0) = 1 - p$ (**"failure"**)
- The binomial variable $Y = \sum_{i=1}^{n} Z_i$
- Probability distribution function is given by

$$P(Y = k) = \binom{n}{k} p^k (1-p)^{n-k} \text{ for } k = 0, 1, \ldots, n$$

- $E(Y) = np$
- $Var(Y) = np(1-p)$
- As $n \to \infty$, Binomial $\to$ Normal:

$$\frac{Y - np}{\sqrt{np(1-p)}} \to N(0,1)$$

- Sample proportion (estimate of p)

$$\hat{p} = \frac{Y}{n}$$

Binomial Data

- **Response** y_i: number of successes out of n_i independent trials with probability of success p_i
- $x = (x_1, x_2, \ldots, x_p)$: **predictors** (quantitative, factors, or both)
- For all trials contributing to one response y_i, the predictors x_i have the same value (*covariate class*)
- Goal: model the relationship between y and $x_1, \ldots, x_p$ via modeling **the relationship between p_i and $x_1, \ldots, x_p$** .

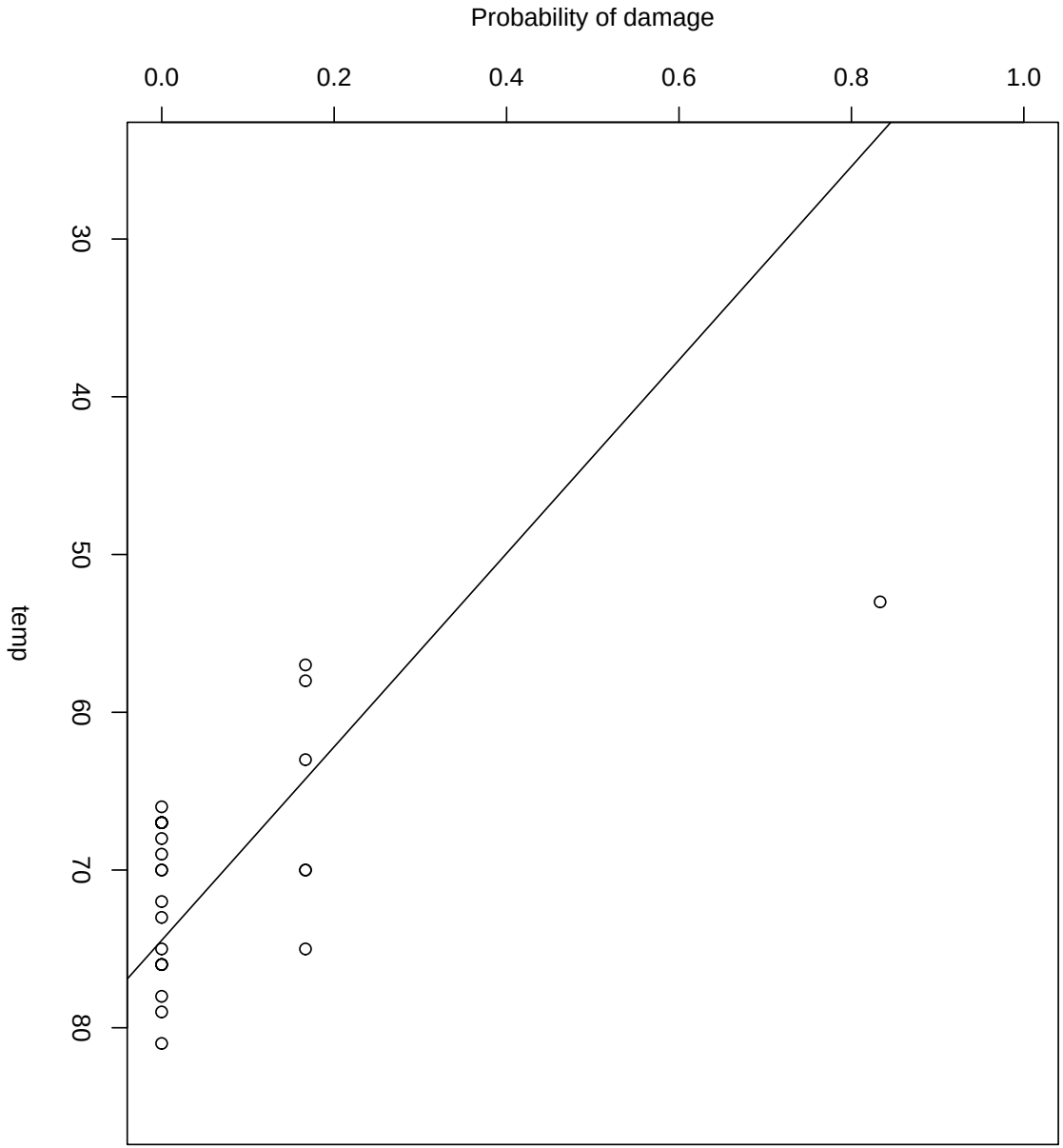
Challenger Disaster Example

- The space shuttle Challenger exploded after launch in 1986
- One explanation: rubber seals called O-rings
- Rubber gets brittle at cold temperatures and becomes less effective as a sealant, and it was an unusually cold day (31F)
- Have data on damage to O-rings (how many showed evidence of damage out of 6 total) and temperature from previous launches

```
## Load the data
> library(faraway)
> data(orings)
## Fit a linear model to observed proportions
> plot(damage/6 ~ temp, orings, xlim=c(25,85),
+   ylim = c(0,1),ylab="Probability of damage")
> abline(lm(damage/6 ~ temp, orings))
```

The linear model is clearly inappropriate here.

Challenger Disaster Data



Binomial Regression

- Assume that y_i is Binomial(n_i, p_i)
- Assume all y_i's are independent
- Linear predictor:

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

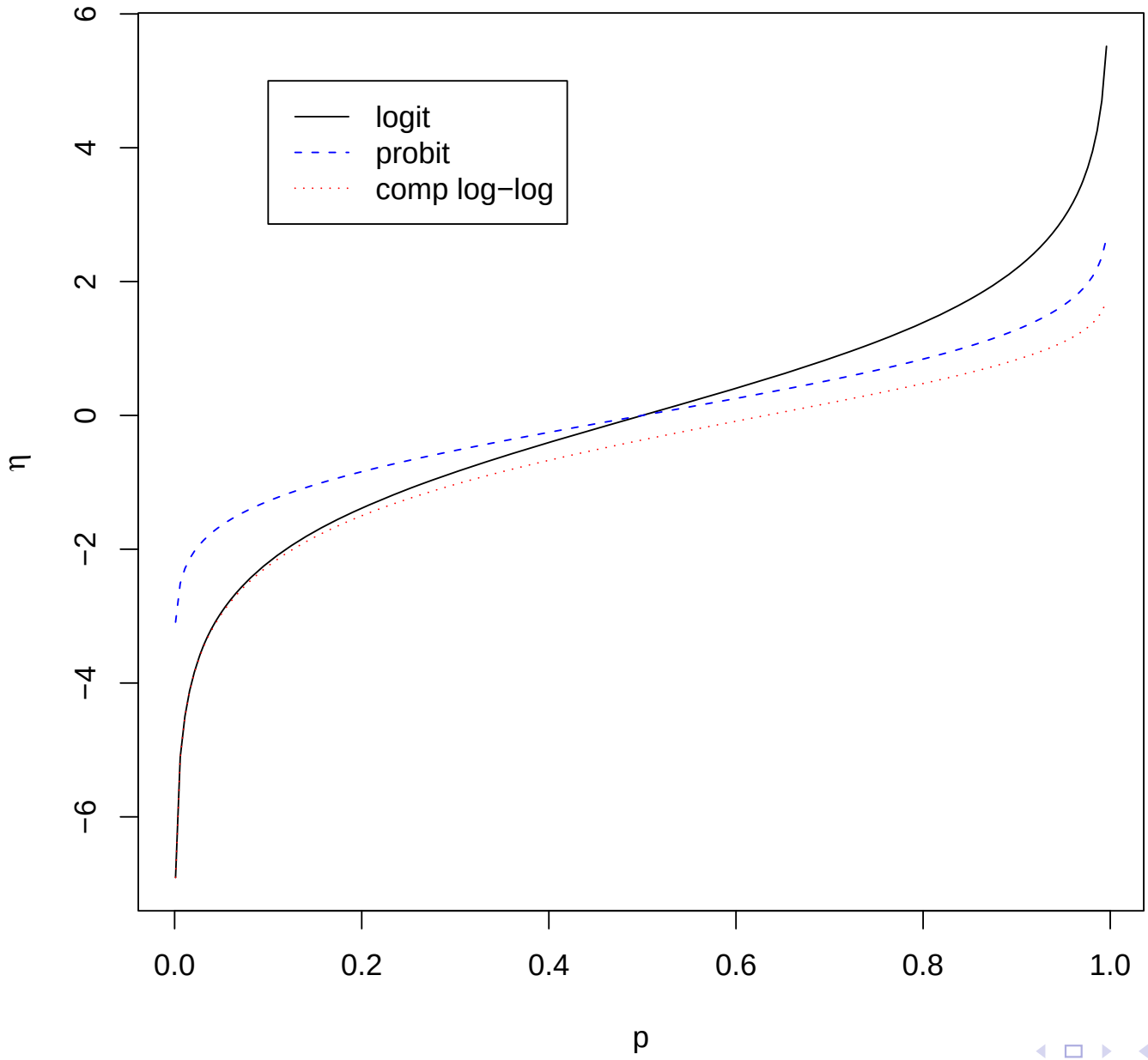
- Cannot use $\eta_i = p_i$ (need $0 \leq p \leq 1$)
- Main idea: use a **link function**

$$\eta_i = g(p_i)$$

Binomial link functions

- **Logit** : $\eta = \log(p/(1-p))$
- **Probit** : $\eta = \Phi^{-1}(p)$, where Φ is the cumulative distribution function of $N(0,1)$
- **Complementary log-log** : $\eta = \log(-\log(1-p))$
- All transform $p \in (0,1)$ to $\eta \in (-\infty, \infty)$

Binomial link functions



Estimating parameters

- **Maximum likelihood** approach: find parameters (in this case p_i) that maximize the likelihood of the data,

$$\prod_{i=1}^{n} P(Y_i = y_i) \quad \text{or equivalently}$$

  where Y_i is Binomial(n_i, p_i).
- Log-likelihood is given by

$$\ell(p_1, \ldots, p_n; y) = \sum_{i=1}^{n} \left[\log \binom{n_i}{y_i} + y_i \log p_i + (n_i - y_i) \log(1 - p_i) \right]$$

- For the logit link, need to maximize with respect to β

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i (x_i^T \beta) - n_i \log(1 + \exp(x_i^T \beta)) \right]$$

- Optimization algorithm is complicated (Ch. 6)

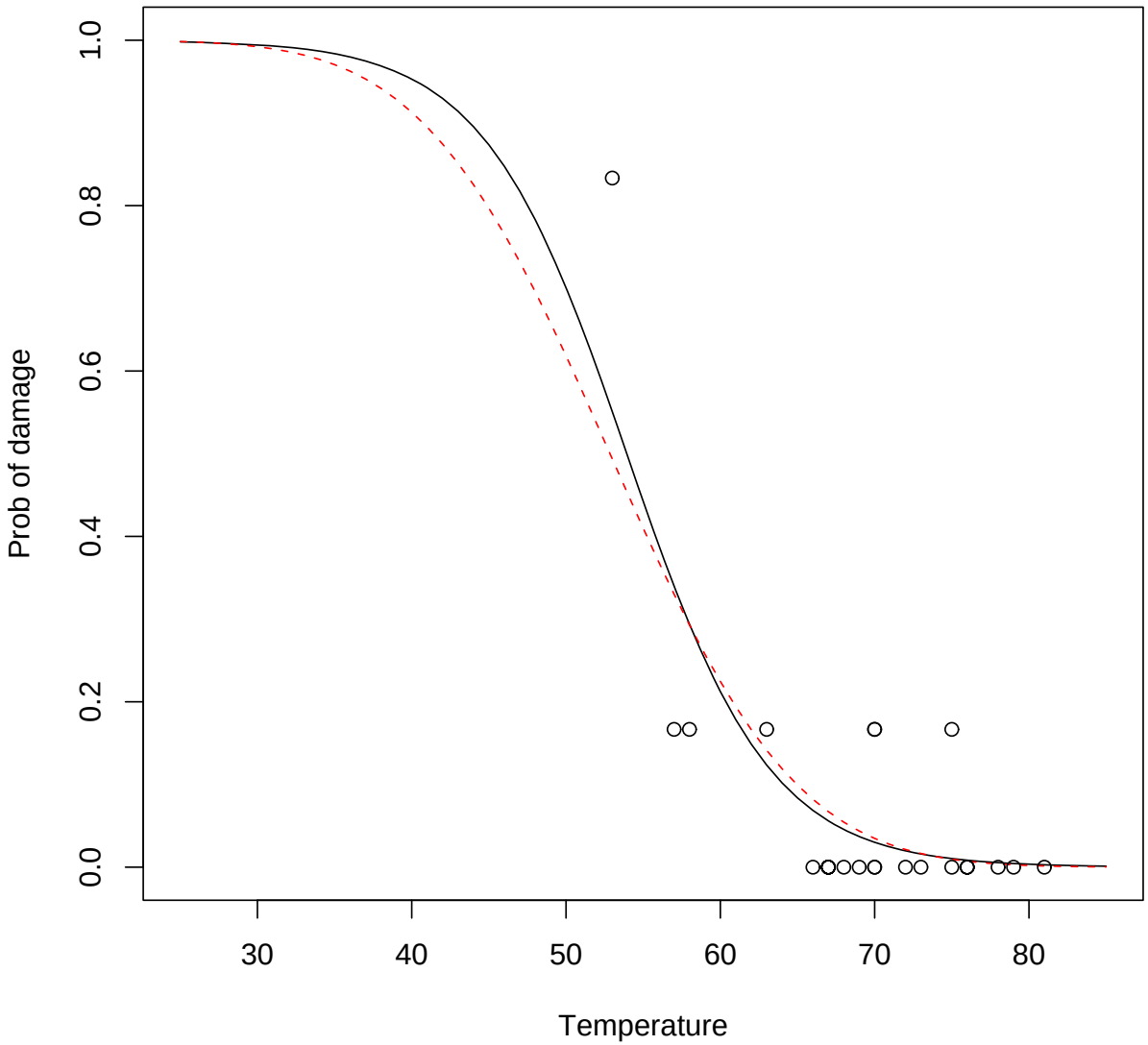
Challenger Example

```
> logitm = glm(cbind(damage,6-damage) ~ temp,
+              family=binomial(link=logit), data=orings)
> summary(logitm)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.9529  -0.7345  -0.4393  -0.2079   1.9565
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 11.66299    3.29626   3.538 0.000403 ***
temp        -0.21623    0.05318  -4.066 4.78e-05 ***
---
(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
```

```
## estimate probability of failure at temp = 31F
> test = data.frame(temp=31)
> ilogit(predict(logitm,test))
[1] 0.9930342

## fit a probit model to compare
> probitm = glm(cbind(damage,6-damage) ~ temp,
+               family=binomial(link=probit), data=orings)

> summary(probitm)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0134  -0.7760  -0.4467  -0.1581   1.9982
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.59145    1.71055   3.269  0.00108 **
temp        -0.10580    0.02656  -3.984 6.79e-05 ***
---
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 38.898  on 22  degrees of freedom
Residual deviance: 18.131  on 21  degrees of freedom
AIC: 34.893
```

```
## Probit prediction at temp = 31F
> pnorm(predict(probitm,test))
[1] 0.9895983

# Make predictions for the whole range and plot
> range = data.frame(temp=seq(25,85,by=1))
> pred.l = ilogit(predict(logitm, range))
> pred.p = pnorm(predict(probitm, range))
> matplot(range, cbind(pred.l,pred.p), xlim=c(25,85),
+  ylim=c(0,1), xlab="Temperature", ylab="Prob of damage",
+  type =’ll’,lty=c(’solid’,’dashed’))
```

Logit and probit fits for Challenger data



Inference

How do we test the goodness-of-fit?

Likelihood ratio test:

- two nested models
- L is the larger model with l parameters and likelihood L_L
- S is the smaller model with $s < l$ parameters and likelihood L_S
- The likelihood ratio statistic is

Deviance

- Take L to be the **saturated** model: n parameters to fit each data point perfectly, with fitted values $\hat{p}_i = y_i/n_i$.
- In this case, the test statistic is called **the deviance of** S and is given by

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

  where $\hat{y}_i = n_i\hat{p}_i$, $\hat{p}_i$ are the fitted probabilities from S.
- If Y_i's are truly binomial, independent, n_i are large

$$D \approx \chi^2_{n-s}$$

Deviance Ctd

- Can use to test the **goodness-of-fit** :
  p-value $= P(\chi^2_{n-s} > D)$
- Can also use to **compare two nested models** , e.g. null (no predictors) and current model. In this case, use

$$D_S - D_L \approx \chi^2_{(n-s)-(n-l)}$$

  and the p-value $= P(\chi^2_{l-s} > D_S - D_L)$
- Note: if $n_i = 1$, deviance cannot be used.

Other measures of goodness of fit

- The χ^2 goodness-of-fit statistic (**Pearson's X^2**):

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

- For binomial data, add successes & failures to get

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - n_i\hat{p}_i)^2}{n_i\hat{p}_i(1-\hat{p}_i)}$$

- **Pearson residuals** :

$$r_i^P = \frac{y_i - n_i\hat{p}_i}{\sqrt{var(\hat{y}_i)}}$$

  Then $X^2 = \sum_{i=1}^{n}(r_i^P)^2$.
- Typically X^2 is close to deviance and is used in the same way.

An analogue of R^2

- Proportion of deviance explained can be computed
- A better statistic based on likelihood is

$$R^2 = \frac{1 - (\hat{L}_0/\hat{L})^{2/N}}{1 - \hat{L}_0^{2/N}} = \frac{1 - \exp((D - D_0)/N)}{1 - \exp(-D_0/N)}$$

- N is the total number of binary observations,
  $\hat{L}_0$ and D_0 are the maximized likelihood and deviance under the null model (intercept only),
  $\hat{L}$ and D are the maximized likelihood and deviance under the full model.
- $0 \leq R^2 \leq 1$.

```
## Goodness of fit for the Challenger data
## Deviance test
> pchisq(logitm$dev, df=logitm$df.resid,
+        lower.tail=F)
[1] 0.7164099

## Compare null to model with temperature
> pchisq(logitm$null.dev - logitm$dev,
+  df=logitm$df.null - logitm$df.resid, lower.tail=F)
[1] 2.747351e-06
```

```
## Pearson's chi-squared
> ( X2 = sum(residuals(logitm,type="pearson")^2))
[1] 28.06738
> pchisq(X2, df=logitm$df.resid, lower=F)
[1] 0.1382507

##  R-squared
> dim(orings)
[1] 23  2
> N = 23*6
> (1 - exp((logitm$dev - logitm$null.dev)/N))/
+       (1 - exp(-logitm$null.dev/N))
[1] 0.599577
```

Confidence Intervals for Parameters

- Asymptotically $\hat{\beta}$ is normal – can use z-intervals
- **Profile likelihood confidence intervals** are more accurate (based on considering the likelihood of one parameter with all others fixed)

```
> library(MASS)
> confint(logitm)
Waiting for profiling to be done...
                2.5 %    97.5 %
(Intercept)  5.575195 18.737598
temp        -0.332657 -0.120179
```

Confidence Intervals for Predictions

- No distinction here between future observation and mean response
- Based on asymptotic normality of $\hat{\beta}$ and $x_0\hat{\beta}$

```
> predict(logitm, test, se=T)
$fit
       1
4.959746
$se.fit
[1] 1.66731

> ilogit(c(4.96-1.96*1.67,4.96+1.96*1.67))
[1] 0.8438029 0.9997344
```

Interpreting Odds

- **Odds** : $\frac{p}{1-p}$
- **Logistic regression** (logit link) models **log odds** :

$$\log(\text{odds}) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- **Interpretation** : a unit increase in x_1 with all other predictors held fixed leads to an increase of β_1 in log-odds, or equivalently, odds being multiplied by $\exp(\beta_1)$.
- No such interpretation available for other link functions

Example: breastfeeding and respiratory disease

- Factors: gender (boy/girl), feeding (breast/bottle/supplement)
- Response: incidence of respiratory disease in the 1st year

```
> data(babyfood)
> babyfood
  disease nondisease  sex   food
1      77         381  Boy Bottle
2      19         128  Boy  Suppl
3      47         447  Boy Breast
4      48         336 Girl Bottle
5      16         111 Girl  Suppl
6      31         433 Girl Breast
```

```
# look at rates of disease in each group
> round(xtabs(disease/(disease+nondisease) ~ sex+food,
+       data=babyfood), 4)
      food
sex     Bottle Breast  Suppl
  Boy   0.1681 0.0951 0.1293
  Girl  0.1250 0.0668 0.1260

> babyglm = glm(cbind(disease,nondisease) ~ sex+food,
+           family=binomial, babyfood)
> summary(babyglm)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.6127     0.1124 -14.347  < 2e-16 ***
sexGirl      -0.3126     0.1410  -2.216   0.0267 *
foodBreast   -0.6693     0.1530  -4.374 1.22e-05 ***
foodSuppl    -0.1725     0.2056  -0.839   0.4013
---
    Null deviance: 26.37529  on 5  degrees of freedom
Residual deviance:  0.72192  on 2  degrees of freedom
# Residual deviance shows the interaction term
# is not significant
```

```
# The effect on odds for respiratory disease:
> exp(babyglm$coef)
(Intercept)     sexGirl  foodBreast   foodSuppl
  0.1993479   0.7315770   0.5120696   0.8415226

# Confidence intervals
> exp(confint(babyglm))
Waiting for profiling to be done...
                2.5 %    97.5 %
(Intercept) 0.1591988 0.2474333
sexGirl     0.5536209 0.9629225
foodBreast  0.3781905 0.6895181
foodSuppl   0.5555372 1.2464312
```

Prospective vs. Retrospective Sampling

- **Prospective sampling:** select a sample of newborns in each group and follow them for 1 year to see which ones get disease (cohort study)
- **Retrospective sampling:** record sex and feeding method for infants who report with respiratory disease, and obtain an independent sample of no disease (case-control study)
- Log odds ratio is the same for prospective and retrospective designs
- Suppose D is whether you have the disease. We want to estimate

$$P_x(D) = p(x) = \frac{\exp(\beta_0 + x^T\beta)}{1 + \exp(\beta_0 + x^T\beta)}$$

- Suppose $Z = 1$ means you're included in the study, and

$$\pi_1 = P(Z = 1|D), \quad \pi_0 = P(Z = 1|not\ D).$$

- We can only estimate

$$P_x(D|Z = 1) = p^*(x) = \frac{\pi_1 p(x)}{\pi_1 p(x) + \pi_0(1 - p(x))}$$

- In prospective studies, $\pi_0 = \pi_1$, so $p^*(x) = p(x)$.
- In retrospective studies, typically $\pi_1 \gg \pi_0$ and unknown, but

$$\text{logit}(p^*(x)) = \log \frac{\pi_1}{\pi_0} + \text{logit}(p(x))$$

- Thus the only difference is the intercept

Overdispersion

- What does a large deviance indicate?
- The usual reasons: outliers, non-linearity, model structure (will look at diagnostics in Ch 6)
- Sparse data (small n_i)
- **Overdispersion** : the model implies $var(y_i) = n_i\hat{p}_i(1-\hat{p}_i)$ but in reality $var(y_i)$ is greater
- Some common causes of overdispersion:
  - the trials are **not independent**
  - the probability of success is not constant; **clustering**
- Underdispersion is also possible but rare in practice

Estimating Overdispersion

- Introduce an additional **dispersion parameter** $\phi = \sigma^2$, so that $var(y_i) = \sigma^2 n_i p_i (1 - p_i)$
- Can estimate σ^2 (as in linear regression) as

$$\hat{\sigma}^2 = \frac{X^2}{n - p}$$

- This does not affect $\hat{\beta}$
- All **standard errors** must be multiplied by $\hat{\sigma}$

- Deviance can no longer be used to compare models
- An approximate F-test can be used:

$$F = \frac{(D_S - D_L)/(df_S - df_L)}{\hat{\sigma}^2}$$

  has the F distribution with $df_S - df_L$ and $n - p$ degrees of freedom
- Goodness of fit cannot be tested
- Estimating overdispersion is only reasonable when n_i's are roughly equal

Overdispersion example: trout data

- Boxes of trout eggs buried in a stream and retrieved after some time
- Five different locations (`location`), four lag times in weeks (`period`)
- Number of surviving eggs (`survive`), total in box (`total`)

```
> tmod = glm(cbind(survive, total-survive) ~ location +
+    period, family = binomial, data = troutegg)
> summary(tmod)
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 ***
(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
##  estimate sigma2
> sigma2 = sum(residuals(tmod,type="pearson")^2)/12
> sigma2
[1] 5.330322
```

```
> drop1(tmod, scale=sigma2, 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 **
---
Warning message:
In drop1.glm(tmod, scale = sigma2, test = "F") :
  F test assumes 'quasibinomial' family
```

```
## use estimated dispersion to recompute p-values
> summary(tmod, dispersion=sigma2)
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 ***
---
(Dispersion parameter for binomial family taken to be 5.330322)
```

Summary

- With suitable link functions, binomial data can be modeled easily
- Approximate inference available for testing models and parameter values
- Logit has advantages in interpretation

Warnings:

- The estimation algorithm may not converge
- With small n_i, the χ^2 approximation is poor
- Overdispersion can be accounted for, but binomial assumption is sacrificed