Slide 1

Binomial Distribution

For $Y_i \sim \text{Binomial}(m_i, p_i)$, one has

$$l_i(\theta_i; y_i) = y_i\theta_i - m_i \log(1 + e^{\theta_i}) + \log C_{y_i}^{m_i},$$

where $\theta_i = \log \frac{p_i}{1-p_i}$; $\mu_i = m_i p_i$ and $v_i = m_i p_i(1 - p_i)$. The commonly used links for the binomial family are listed below,

$$\begin{aligned} \text{Logit:} \quad & \eta = \log p/(1-p) \\ \text{Probit:} \quad & \eta = \Phi^{-1}(p) \\ \text{Complementary log-log:} \quad & \eta = \log(-\log(1-p)) \end{aligned}$$

where Φ^{-1} is the inverse cdf of $N(0, 1)$. The logit is the canonical link, which yields the logistic linear model.

The logit and probit links, both symmetric with respect to $p = .5$, are very similar to each other. The complimentary log-log is asymmetric. All these links are mapping $p \in (0, 1)$ to $\eta \in (-\infty, \infty)$.

Slide 2

Dose Response Models

Suppose the "success" probability of a substance depends on the dose x through a cdf $F((x - \mu)/\sigma) = F(z)$ from a location-scale family. The link is $\eta = F^{-1}(p)$ and the model is $\eta = \beta_0 + \beta_1 x$, where $\beta_0 = -\mu/\sigma$ and $\beta_1 = 1/\sigma$. For the three links, one has

Link	$F(z)$	Family
Logit	$e^z/(1 + e^z)$	Logistic
Probit	$\Phi(z)$	Normal
C log-log	$1 - \exp(-e^z)$	Extreme value

and here are some plots.

```
x <- (-150:150)/25; plot(x,1-1/(1+exp(x)),type="l",ylab="p")
lines(x,pnorm(x),col=2); lines(x,1-exp(-exp(x)),col=3)
lines(c(-6,6),c(.5,.5),lty=2,col=1)
lines(x,pnorm(x,sd=1.7),col=6)
```

Slide 3

Deviance and Residuals

The log likelihood of the data is given by

$l(\hat{\boldsymbol{\theta}}; \mathbf{Y}) = \sum_{i=1}^n l_i(\theta_i; y_i) = \sum_{i=1}^n \{y_i \log p_i + (m_i - y_i)\log(1-p_i)\}$,

and $\tilde{p}_i = y_i/m_i$, so the deviance of a fitted model is given by

$$D(\mathbf{y}; \hat{\boldsymbol{\theta}}) = 2\sum_{i=1}^n \{y_i \log \frac{y_i/m_i}{\hat{p}_i} + (m_i - y_i)\log\frac{1-y_i/m_i}{1-\hat{p}_i}\}.$$

Compare with the Pearson X^2 statistic,

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

The deviance and Pearson residuals are the square roots of the summands in above expressions with signs. With the logit link, $d\eta/d\mu = 1/mp(1-p)$, so the working residuals are given by

$$(y_i - m_i\hat{p}_i)/m_i\hat{p}_i(1-\hat{p}_i).$$

Slide 4

Binomial Family in R

Consider the budworm data in the MASS book, page 190.

```
budwm.lgt0 <- glm(SF~ldose*sex,family=binomial,data=budworm)
budwm.lgt1 <- update(budwm.lgt0,numdead/20~.,weight=rep(20,12))
summary(budwm.lgt0); summary(budwm.lgt1)
anova(budwm.lgt0); drop1(budwm.lgt0)
plot(budwm.lgt0); step(budwm.lgt0)
budwm.lgt <- update(budwm.lgt0,.~.-ldose:sex)
budwm.probit <- update(budwm.lgt,family=binomial(probit))
```

Internally, the fitting were done with y_i/m_i as the responses and m_i as weights, and the predictions and residuals of `type="response"` are calculated on the p scale, not mp.

```
predict(budwm.lgt,new=data.frame(ldose=2,sex="M"))
predict(budwm.lgt,type="response")
residuals(budwm.lgt); residuals(budwm.lgt,"pearson")
```

Slide 5

Accuracy of Asymptotic Inference

The asymptotic inferential tools are valid only under conditions. In general, $\hat{p}_i$'s too close to 1 or 0 are unfavorable.

For $I^{-1}(\hat{\boldsymbol{\beta}})$ to be reasonable estimates of $\text{Cov}(\hat{\boldsymbol{\beta}})$, $l(\boldsymbol{\beta})$ nees to be "sufficiently quadratic" around $\hat{\boldsymbol{\beta}}$. Too large a $|\hat{\beta}_j|$ leads to too small a t-statistic in logistic regression.

For the log likelihood ratio statistic (the deviance difference) to be approximately χ^2, one needs a large "sample size" compared to the dimension of the *full* model.

For Bernoulli data, the deviance is *not* χ^2, but for binomial data with m_i large, the deviance is approximately χ^2 for a proper model. Remember that $Y \sim \text{Bin}(m, p)$ is the sum of m *i.i.d.* Bernoulli responses.

Slide 6

Retrospective Logistic Model

Consider a logistic model with covariates $\mathbf{x}$,

$$P(D|\mathbf{x}) = \exp\{\alpha + \mathbf{x}^T\boldsymbol{\beta}\}/(1 + \exp\{\alpha + \mathbf{x}^T\boldsymbol{\beta}\}).$$

In rare disease studies, data are often sampled retrospectively, with $\pi_0 = P(Z = 1|D)$ and $\pi_1 = P(Z = 1|\bar{D})$, where Z is the selection indicator. Applying Bayes's Theorem, one has

$$\begin{aligned} P(D|Z = 1, \mathbf{x}) &= \frac{P(Z = 1|D, \mathbf{x})P(D|\mathbf{x})}{P(Z = 1|D, \mathbf{x})P(D|\mathbf{x}) + P(Z = 1|\bar{D}, \mathbf{x})P(\bar{D}|\mathbf{x})} \\ &= \frac{\pi_0 \exp\{\alpha + \mathbf{x}^T\boldsymbol{\beta}\}}{\pi_0 \exp\{\alpha + \mathbf{x}^T\boldsymbol{\beta}\} + \pi_1} = \frac{\exp\{\alpha^* + \mathbf{x}^T\boldsymbol{\beta}\}}{1 + \exp\{\alpha^* + \mathbf{x}^T\boldsymbol{\beta}\}}, \end{aligned}$$

where $\alpha^* = \alpha + \log(\pi_0/\pi_1)$. Hence, the covariate effects in a logistic model can be estimated from retrospective samples so long as the covariates $\mathbf{x}$ play no role in the selection. Note that other links are not as "friendly" to retrospective sampling.

Slide 7

Sampling Efficiency: 2×2 Table

Consider the following 2×2 table.

Exposure	$\bar{D}$	D
$\bar{X}$	p_{00}	p_{01}
X	p_{10}	p_{11}

The log odds ratio,

$$\delta = \log(p_{01}/p_{00}) - \log(p_{11}/p_{10}),$$

measures the association between exposure and disease.

To estimate the parameter δ, one may use prospective sampling with fixed row totals, or retrospective sampling with fixed column totals.

With fixed row totals $n_{i\cdot}$, one has $n_{i1} \sim \text{Bin}(n_{i\cdot}, p_{i1}/(p_{i0} + p_{i1}))$.

$$\hat{\delta} = \log(n_{01}/n_{00}) - \log(n_{11}/n_{10})$$

has an approximate variance

$$\text{var}[\hat{\delta}] \approx \frac{n_{0\cdot}}{n_{00}n_{01}} + \frac{n_{1\cdot}}{n_{10}n_{11}} = \frac{1}{n_{00}} + \frac{1}{n_{01}} + \frac{1}{n_{10}} + \frac{1}{n_{11}};$$

the δ-method is used along with

$$\frac{d}{dp} \log \frac{p}{1-p} = \frac{1}{p(1-p)}.$$

The same variance formula results with fixed columnn counts.

For the estimate to be accurate, one needs to avoid low cell counts.

Slide 8

2×2 Table: Examples

Consider artificially constructed data of not so rare disease.

Prospective sampling.

Exposure	$\bar{D}$	D	
$\bar{X}$	49	1	50
X	46	4	50
	95	5	100

The log odds ratio is estimated by

$$\hat{\delta} = \log \tfrac{1/49}{4/46} = -1.45,$$

with standard error

$$\sqrt{\tfrac{1}{49} + 1 + \tfrac{1}{46} + \tfrac{1}{4}} = 1.14.$$

Retrospective sampling.

Exposure	$\bar{D}$	D	
$\bar{X}$	26	10	36
X	24	40	64
	50	50	100

The log odds ratio is estimated by

$$\hat{\delta} = \log \tfrac{10/26}{40/24} = -1.47,$$

with standard error

$$\sqrt{\tfrac{1}{26} + \tfrac{1}{24} + \tfrac{1}{10} + \tfrac{1}{40}} = 0.45.$$

Slide 9

Retrospective Logistic Model: Example

Consider a simulation example with $\text{logit}(D|x) = -10 + 4x$ for $x \in [0, 1]$. In prospective sampling, one gets "around" 60 cases in every 100,000 individuals.

```
x <- runif(100000); y <- rbinom(x,1,plogis(-10+4*x)); sum(y)
ind <- c(sample(100000,60),(1:100000)[y==1])
smpl <- data.frame(x=x[ind],y=y[ind])
summary(glm(y~x,family=binomial,data=smpl))
```

With $\text{logit}(D|x) = -2 + 4x$ and $\text{logit}(D|x) = -6 + 4x$, one has

```
x <- runif(120); y <- rbinom(x,1,plogis(-2+4*x))
summary(glm(y~x,family=binomial))
x <- runif(2000); y <- rbinom(x,1,plogis(-6+4*x)); sum(y)
ind <- (1:2000)[y==1]; ind <- c(ind,sample((1:2000)[-ind],60))
smpl <- data.frame(x=x[ind],y=y[ind])
summary(glm(y~x,family=binomial,data=smpl))
```

Slide 10

Empirical Logistic Transform, Over-Dispersion

Consider the *empirical logistic transform*,

$$Z = \log(Y + \tfrac{1}{2})/(m - Y + \tfrac{1}{2}).$$

For m large, it can be shown that $E[Z] = \log p/(1-p) + O(m^{-2})$, with $\text{var}[Z] \approx (y + \frac{1}{2})^{-1} + (m - y + \frac{1}{2})^{-1}$. The transform can be used to calculate starting values for maximum likelihood iteration.

A binomial r.v. is a sum of *i.i.d.* Bernoulli r.v.'s. For m large, the components may not be *i.i.d.*, and *over-dispersion* may occur. Suppose $Y = \sum_{i=1}^{c} X_i$, where $X_i \sim \text{Binomial}(k_i, p_i)$, independent, $\sum_{i=1}^{c} k_i = m$, $E[p_i] = p$, and $\text{var}[p_i] = \tau^2 p(1-p)$. It follows that

$$E[Y] = mp, \quad \text{var}[Y] = mp(1-p)\{1 + \tau^2 \textstyle\sum_{i=1}^{c}(k_i^2 - k_i)/m\}.$$

Clustering is seen to be one possible cause for over-dispersion.

Slide 11

Over-Dispersion: Example

A simple "model" for over-dispersion is to assume $\text{var}[Y] = v\sigma^2$ and estimate σ^2 by $\tilde{\sigma}^2 = X^2/(n-p)$.

Below is the sex-ratio data concerning 72069 six-child families.

No. Boy	0	1	2	3	4	5	6
No. Family	1096	6233	15700	22221	17332	7908	1579

Fitting a constant, one has $\hat{p} = .5148723$ and $\tilde{\sigma}^2 = 1.047134$.

```
boys.fit <- glm(cbind(boy,girl)~1,binomial,wei=fr,data=boys)
predict(boys.fit,type="res")
p <- sum(boys$fr*(0:6))/sum(boys$fr)/6
sum(resid(boys.fit,type="pear")^2)/(sum(boys$fr)-1)
```

One can test for the binomial assumption via the standard χ^2 test.

```
obs <- boys$fr; xpec <- dbinom(0:6,6,p)*72069
chisq <- sum((obs-xpec)^2/xpec); 1-pchisq(chisq,5)
```

Slide 12

Residual Analysis

Deviance and Pearson residuals are often very similar, and can be checked by `qqnorm()` for outliers. An exception to this is when the data are "sparse", such as with Bernoulli counts.

```
budworm; budwm.lgt; res.dev <- resid(budwm.lgt)
res.pear <- resid(budwm.lgt,type="pear")
plot(res.dev,pch=as.character(budworm$sex))
points(res.pear,pch=as.character(budworm$sex),col=2)
qqnorm(res.dev); qqnorm(res.pear)
```

To possibly check for nonlinearity, one may plot the working or partial residuals against x variables.

```
res.work <- resid(budwm.lgt,type="work")
plot(budworm$l,res.work,pch=as.character(budworm$s))
abline(h=0,lty=2)
```

Slide 13

Terms Predictions and Partial Residuals

Recall the *working responses* (slide 8 of the previous set)

$$\tilde{y}_i = \tilde{\eta}_i + (d\eta_i/d\mu_i)(y_i - \tilde{\mu}_i),$$

where $(d\eta_i/d\mu_i)(y_i - \tilde{\mu}_i)$ are the *working residuals* and $\tilde{\eta}_i$ are the *link predictions*. Writing

$$\tilde{\eta}_i = \hat{\beta}_0 + x_{i,1}\hat{\beta}_1 + \cdots = \tilde{\beta}_0 + (x_{i,1} - \bar{x}_1)\hat{\beta}_1 + \cdots,$$

one has the *terms predictions* in the right-hand side expression. Adding the working residuals to the terms predictions, one gets the *partial residuals*.

```
res.work+predict(budwm.lgt,type="term")
resid(budwm.lgt,type="part")
```

Slide 14

Regression Diagnostics

Regression diagnostics for GLM can be calculated based on the weighted LS problem at the MLE (slide 8 of the previous set),

$$\textstyle\sum_{i=1}^n \tilde{w}_i(\tilde{y} - \mathbf{x}_i^T\boldsymbol{\beta})^2 = (\tilde{\mathbf{Y}}_w - X_w\boldsymbol{\beta})^T(\tilde{\mathbf{Y}}_w - X_w\boldsymbol{\beta}),$$

where $\tilde{\mathbf{Y}}_w = W^{1/2}\mathbf{Y}$ and $X_w = W^{1/2}X$. For the delete-one quantities such as $\hat{\boldsymbol{\beta}}_{(i)}$ used in dfbetas and Cook's D, such calculation yields the one-step update from the full-sample MLE.

Consider again the budworm example.

```
plot(budwm.lgt)
influence.measures(budwm.lgt)
```

Remember that the fit `budwm.lgt` is also a class `lm` object.