Chapter 11: Binomial and Poisson Regression

1 Distribution for Counted Data

1.1 Bernoulli Distribution

Suppose the random variable Y has two possible values, perhaps called "success" or "failure", with probability of success equal to θ, where $0 \leq \theta \leq 1$. We label $Y = 1$ if success occurs, and $Y = 0$ otherwise. We will say that Y has a Bernoulli distribution with probability of success θ:

$$E(Y) = \theta, \quad \text{Var}(Y) = \theta(1 - \theta).$$

An important feature of the Bernoulli distribution is that the variance depends on the mean.

1.2 Binomial Distribution

The binomial distribution generalizes the Bernoulli. Suppose we have m random variables $B_1, B_2, \ldots, B_m$, such that (1) each B_j has a Bernoulli distribution with the same probability θ of success, and (2) all the B_j's are independent. Then if Y is the number of successes in the m trials, $Y = \sum B_j$, we say that Y has a binomial distribution with m trials and probability of success θ. The probability mass

function is

$$P(Y = j) = \binom{m}{j} \theta^j (1-\theta)^{m-j},$$

for $j \in \{0, 1, \ldots, m\}$. The mean and variance of the distribution are

$$E(Y) = m\theta, \quad \text{Var}(Y) = m\theta(1-\theta).$$

1.3 Poisson Distribution

The Poisson distribution is the number of events of a specific type that occur in a fixed time of space. A Poisson variable Y can take the value of any nonnegative integer $\{0, 1, 2, \ldots\}$. The probability mass function of the Poisson distribution is given by

$$P(Y = y) = \exp(-\lambda)\lambda^y / y!, \quad y = 0, 1, 2, \ldots,$$

The mean and variance are given by

$$E(Y) = \lambda, \quad \text{Var}(Y) = \lambda.$$

2 Regression Model for Counts

The big idea is that the parameter for the counted distribution, θ for the binomial or λ for the Poisson, can depend on the values of predictors.

2.1 Binomial Regression

We assume that $\theta(\boldsymbol{x})$ depends on the values $\boldsymbol{x}$ of the regressors only through a linear combination

$\boldsymbol{\beta}'\boldsymbol{x}$ for some unknown $\boldsymbol{\beta}$:

$$\theta(\boldsymbol{x}) = m(\boldsymbol{\beta}'\boldsymbol{x}),$$

where $\boldsymbol{\beta}'\boldsymbol{x}$ is called the linear predictor. For logistic regression, we have

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

Most presentation of logistic regression work with the link function, which is the inverse of the kernel

mean function; that is,

$$\log\left(\frac{\theta(\boldsymbol{x})}{1-\theta(\boldsymbol{x})}\right) = \boldsymbol{\beta}'\boldsymbol{x},$$

where the left side is called a logit or log-odds, and the ratio $\frac{\theta(\boldsymbol{x})}{1-\theta(\boldsymbol{x})}$ is called the odds of success.

Logistic regression models are not fit with OLS. Rather, maximum likelihood estimation is used, based on the binomial distribution.

```
> g1 <- glm(cbind(died, m-died) ~ log(d), family=binomial, data=
> summary(g1)
glm(formula = cbind(died, m - died) ~ log(d), family = binomial,
    data = BlowBS)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -7.8925     0.6325  -12.48   <2e-16 ***
log(d)        3.2643     0.2761   11.82   <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: 250.856  on 34  degrees of freedom
Residual deviance:  49.891  on 33  degrees of freedom
AIC: 117.52
Number of Fisher Scoring iterations: 4
```

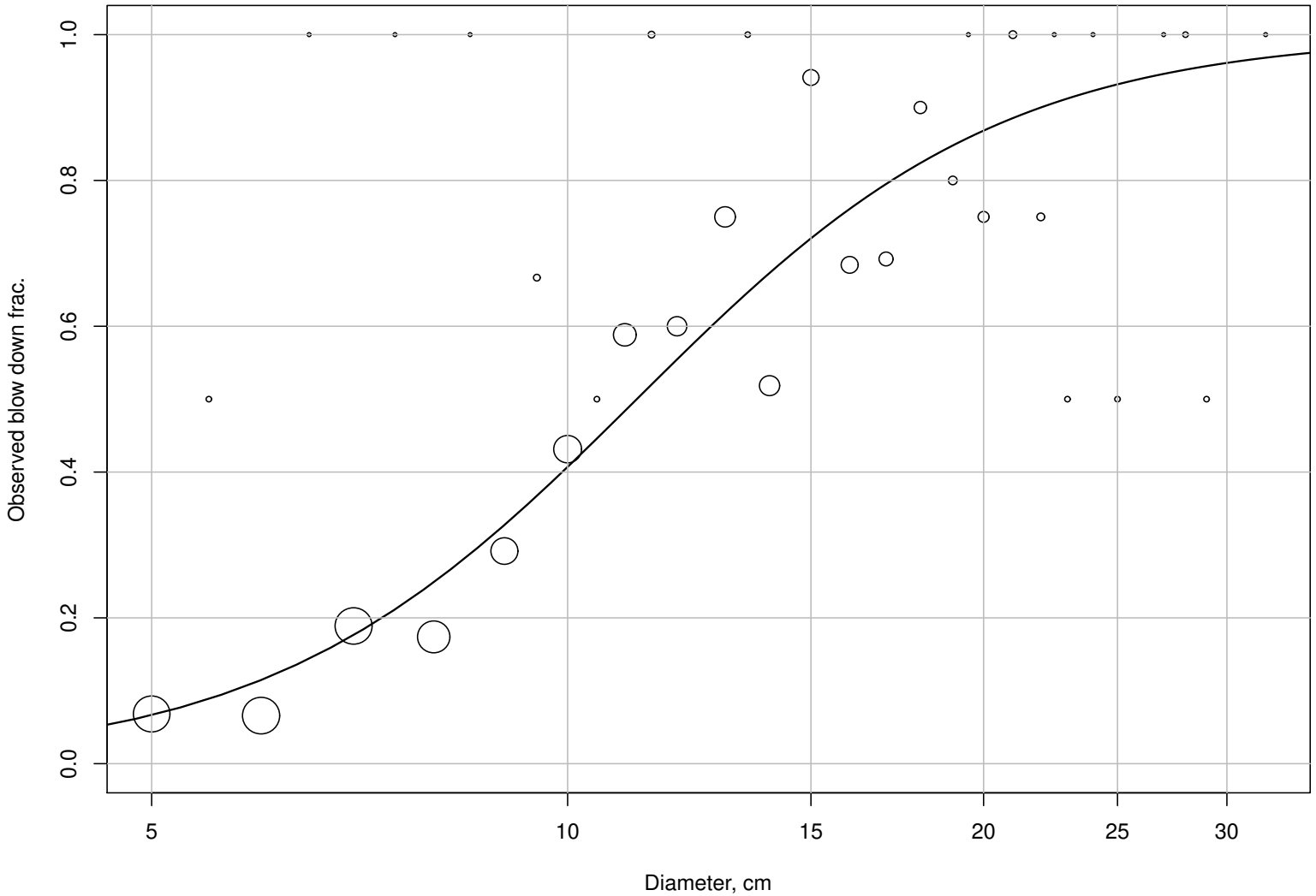

Figure 1: Plot of the blowdown fraction versus d, with the horizontal axis in log scale.

2.2 Deviance

In multiple linear regression, the residual sum of squares provides the basis for tests for comparing mean functions. In logistic and Poisson regression, the residual sum of squares is replaced by the deviance, which is often called G^2. The deviance is defined for logistic regression to be

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

where $\hat{y}_i = m_i\hat{\theta}_i(\boldsymbol{x}_i)$ are the fitted number of successes in m_i trials. The df associated with the deviance is equal to the number of cases n used in the calculation minus the number of elements of $\boldsymbol{\beta}$.

Methodology for comparing models parallels the results in multiple linear regression. Write

$$\boldsymbol{\beta}'\boldsymbol{x} = \boldsymbol{\beta}_1'\boldsymbol{x}_1 + \boldsymbol{\beta}_2'\boldsymbol{x}_2,$$

and consider testing

$$NH: \theta(\boldsymbol{x}) = m(\boldsymbol{\beta}'\boldsymbol{x}),$$
$$AH: \theta(\boldsymbol{x}) = m(\boldsymbol{\beta}_1'\boldsymbol{x}_1 + \boldsymbol{\beta}_2'\boldsymbol{x}_2).$$

Obtain the deviance G^2_{NH} and degrees of freedom df_{NH} under the null hypothesis, and then obtain G^2_{AH} and df_{AH} under the alternative hypothesis. As with linear models, we will have evidence against the null hypothesis if $G^2_{NH} - G^2_{AH}$ is too large. To get a p-value, we compare the differ-

ence $G^2_{NH} - G^2_{AH}$ with the χ^2 distribution with $df = df_{NH} - df_{AH}$, not with an f-distribution as was done for linear models.

3 Poisson Regression

When the data are to be modeled as if they are Poisson counts, the rate parameter is assumed to depend on the regression with linear predictors $\boldsymbol{\beta}'\boldsymbol{x}$ through the link function

$$\log[\lambda(\boldsymbol{\beta}'\boldsymbol{x})] = \boldsymbol{\beta}'\boldsymbol{x}.$$

Poisson regression models are often called log-linear models.

Maximum likelihood estimation is the usual method used to fit Poisson regression models. The deviance for Poisson regression is given by

$$G^2 = 2\sum_{i=1}^{n}[y_i \log(y_i/\hat{y}_i) - (y_i - \hat{y}_i)],$$

where $\hat{y}_i$ us the fitted value $\exp(\hat{\boldsymbol{\beta}}'\boldsymbol{x}_i)$.

```
> p1 <- glm(count ~ (type + sex + citizen)^3, poisson, AMSsurvey
> summary(p1)
(Intercept)                 3.21888    0.20000  16.094  < 2e-16
typeI(Pu)                   0.14842    0.27292   0.544 0.586557
typeII                      0.69315    0.24495   2.830 0.004658
typeIII                     0.44469    0.25621   1.736 0.082623
typeIV                      1.43508    0.22254   6.449 1.13e-10
typeVa                     -0.73397    0.35119  -2.090 0.036622
sexMale                     1.15057    0.22947   5.014 5.33e-07
citizenUS                  -0.22314    0.30000  -0.744 0.456990
typeI(Pu):sexMale           0.34967    0.30795   1.135 0.256181
typeII:sexMale             -0.57396    0.28964  -1.982 0.047525
typeIII:sexMale            -0.84384    0.31172  -2.707 0.006788
typeIV:sexMale             -1.00051    0.26529  -3.771 0.000162
typeVa:sexMale             -0.30327    0.41437  -0.732 0.464239
typeI(Pu):citizenUS         0.41120    0.39122   1.051 0.293233
typeII:citizenUS            0.16127    0.36232   0.445 0.656249
```

```
typeIII:citizenUS                0.02532    0.38326   0.066 0.947331
typeIV:citizenUS                -0.44183    0.34357  -1.286 0.198445
typeVa:citizenUS                 0.37729    0.49473   0.763 0.445690
sexMale:citizenUS                0.31960    0.33786   0.946 0.344173
typeI(Pu):sexMale:citizenUS     -0.49239    0.43872  -1.122 0.261722
typeII:sexMale:citizenUS        -0.18202    0.42081  -0.433 0.665351
typeIII:sexMale:citizenUS       -0.24192    0.45955  -0.526 0.598589
typeIV:sexMale:citizenUS        -0.19597    0.40556  -0.483 0.628947
typeVa:sexMale:citizenUS        -0.27960    0.57796  -0.484 0.628552
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

3.1 Goodness of Fit Tests

If a Poisson mean function is correctly specified, the residual deviance G^2 will be distributed as a $\chi^2(n-p')$ random variable, where n is the number of cells and p' is the number of regressor fit. If the mean function is not correctly specified, or if the Poisson assumption is wrong, then G^2 will generally be too large, and so a lack of fit test can be obtained by comparing the value of G^2 to the rel-

evant χ^2 distribution. The same idea can be used for binomial regression when the sample sizes are larger than 1.

An alternative to using G^2 for lack of fit testing is to use Pearson's χ^2 for testing, given by the familiar formula

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

Like G^2, X^2 is compared with $\chi^2(n - p')$ to get

significance levels. In large samples, the two tests will give the same inference, but in smaller samples χ^2 is generally more powerful.

In binomial regression with all or nearly all the $m_i = 1$, neither G^2 nor X^2 provides a lack of fit tests.

4 Transferring what you know about linear models

Most of the methodology developed in this book transfers to problems with binomial or Poisson responses. In this section, important connections are briefly summarized.

4.1 Scatterplots and Regression

Graphing data is just as important in binomial and Poisson regression as it is in linear regression. In

problems with a binary response, plots of the response versus predictors or regressors are generally not very helpful because the response only has two values. Smoothers, however, can help look a these plots as well. Plots of predictors with color used to indicate the level of the response can also be helpful.

The general ideas in Chapters 2 and 3 apply to binomial and Poisson models, even if the details differ. With the counted data models, estimates $\hat{\boldsymbol{\beta}}$ and Var$(\hat{\boldsymbol{\beta}}|X)$ are computed using the appropriate maximum likelihood methods, not with the formulas in these chapters. Once these are found, they can be used in these formulas and methods given in the text. For example, a point estimate

and standard error for a linear combination of the elements of $\boldsymbol{\beta}$ is given by

$$\hat{l} = \boldsymbol{a}'\boldsymbol{\beta}, \quad se(\hat{l}|X) = \hat{\sigma}\sqrt{\boldsymbol{a}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{a}},$$

for linear regression. For the binomial and Poisson fit, we can replace $\hat{\sigma}$ by 1 and replace $(\boldsymbol{X}'\boldsymbol{X})^{-}$ by the covariance matrix of $\hat{\boldsymbol{\beta}}$. Confidence intervals and tests use the standard normal rather than a t-distribution.

The t-tests discussed in Chapters 2, 3, and 6 are replaced by z-tests for binomial and Poisson models. The F-tests in Chapter 6 are replaced by χ^2 tests based on changes in deviance. The marginality principle, Section 6.2, is the guiding principle for testing with counted responses.

In linear models, the t-tests and F-tests for the same hypothesis have the same value, and so they

are identical. With binomial and Poisson responses, the tests are identical only for very large samples, and in small samples they can give conflicting summaries. The G^2 tests are generally preferred.

4.4 Variances

Failure of the assumptions needed for binomial or Poisson fitting may be reflected in overdispersion, meaning that the variation between observations

given the predictors is larger than the value required by the model. One general approach to overdispersion is to fit models that allow for it, such as the binomial or Poisson mixed models similar to those in Section 7.4. Other models, for example, using negative binomial distributions rather than binomial, can account for overdispersion.

4.5 Transformations

Transformation of the response is not relevant with binomial and Poisson models. Transformation of predictors is relevant, however, and all the methodology in Chapter 8 can be used.

4.6 Regression Diagnostics

Many diagnostic methods depend on residuals. In binomial and Poisson models, the variance depends

on the mean, and any useful residuals must be scaled to account for variance. A generalization of the Pearson residuals defined in Section 9.1.3, is appropriate for most purposes.

4.7 Variable Selection

All the ideas discussed in Chapter 10 carry over to binomial and Poisson models.

5 Generalized Linear models

The multiple linear regression, logistic, and Poisson log-linear models are particular instances of generalized linear models. They share three basic characteristics:

1. The conditional distribution of the response $Y|X$ is distributed according to an exponential family distribution. The important members of this class include the normal, binomial, Poisson,

and gamma distributions.

2. The response Y depends on the regressors only through the linear combinations of terms $\boldsymbol{\beta}'\boldsymbol{x}$.

3. The mean $E(Y|X = \boldsymbol{x}) = m(\boldsymbol{\beta}'\boldsymbol{x})$ for some kernel mean function m. For the multiple linear regression model, m is the identity function, and for logistic regression it is the logistic function. The Poisson was specified us-

ing the log link, so its m is the inverse of the log, or the exponential function. Other choices of the kernel mean function are possible but are used less often in practice.

These three components are enough to specify completely a regression problem along with methods for computing estimates and making inferences. The methodology for these models generally builds on these methods in this book, usually with only

minor modification.