Week 6 Notes: Robust Regression

Antonio R. Linero
University of Texas at Austin

Goals

- Learn how to perform *robust* inference within the GLM framework.
  - Quasi-likelihood methods
  - Overdispersed generative models
  - Method of moments
  - Nonparametric bootstraps

Motivation

Box:

> *. . . all models are approximations. Essentially, all models are wrong, but some are useful. However, the approximate nature of the model must always be borne in mind. . .*

Goal: Understand under what situations our inferences break down and how to fix them.

Common Pattern

i. The point estimates obtained via MLE correspond to something of reasonable scientific interest; but

ii. The posterior/likelihood/whatever we use to quantify uncertainty **does not** correspond to something reasonable.

Overdispersion in GLMs is one example (among many) of this phenomenon that we are already familiar with.

Questions

1. How does inference based on the likelihood (including Bayesian inference) behave when the model is misspecified?
2. Are there broader models that we might consider that we can (and, perhaps, should) use instead?

Overdispersion

Recall: Poisson and binomial GLMs necessarily have

$$\phi = 1$$

We say that our count (binomial) data is *overdispersed* relative to the Poisson (binomial) distribution if

$$\text{Var}(Y_i \mid \mu_i) > \frac{V(\mu_i)}{\omega_i}$$

where $V(\mu_i) = \mu_i$ for Poisson data and $V(\mu_i) = \mu_i(1 - \mu_i)$ for count data.

Why Overdispersion Matters

Asymptotic variance:

$$\mathrm{Var}(\widehat{\beta}) \approx \phi(X^\top W X)^{-1}.$$

$\phi = 1$ too small $\rightsquigarrow$ poor coverage/hypothesis testing.

Exercise

Exercise: Ticks

We examine a dataset described by Elston et al. (2001, *Parisitology*) which contains measures of the number of ticks on Red grouse chicks (a ground-nesting species of birds). Chicks were captured, the number of ticks were counted, and then the chicks were released. Interest lies in the relationship between `HEIGHT` - the height above sea level at which the chick was caught - and the number of ticks the chicks had, as well as whether this relationship varies by year. This dataset can be loaded in `R` by running the code

```
ticks <- lme4::grouseticks
```

Note: for the sake of simplicity, we will ignore the variable `BROOD`, which indexes the brood that the chick belongs to (chicks in the same brood come from the same family). A serious analysis of this dataset would control for this, since chicks in the same brood are likely to have similar exposures to ticks.

a. Fit a Poisson loglinear model of the form

$$Y_{ij} \sim \text{Poisson}(\mu_{ij}), \qquad \log(\mu_{ij}) = \alpha_j + \beta_j \times \texttt{HEIGHT}_i$$

where Y_{ij} denotes the i^{th} chick observed in year j.

b. One way to check whether overdispersion is an issue is to look at the statistic $\widehat{\phi} = \frac{1}{N-P}\sum_i (Y_i - \widehat{\mu}_i)^2/\widehat{\mu}_i$. Since this is an estimate of $\phi = 1$ for the Poisson loglinear model, we should be concerned if this quantity is large.

Compute $\widehat{\phi}$; does this seem large enough to cause concern?

c. We can formally test the hypothesis $\widehat{\phi} = 1$ by comparing to its sampling distribution. Use the **parametric** bootstrap (keeping `YEAR` and `HEIGHT` fixed but resampling Y_i for each i) to sample many realizations of $\widehat{\phi}$ from the fitted model. Use this to approximate a p-value which gives the (approximate) probability of observing a value at least as large as the realized value of $\widehat{\phi}$ on a replicated dataset.

Generative Models: Negative Binomial Regression

Overdispersed count model data:

$$f(y \mid \mu, k) = \frac{\Gamma(y+k)}{y!\Gamma(k)} \left(\frac{\mu}{\mu+k}\right)^{y} \left(1 - \frac{\mu}{\mu+k}\right)^{k}.$$

Then, usually set $\log \mu_i = X_i^\top \beta$.

Exercise

Exercise: Negative Binomial

Show that the negative binomial model (with k fixed) is an exponential dispersion family with $\phi = 1$. Argue also that, while k controls the amount of overdispersion in the model, it is not quite the same as a dispersion parameter ϕ.

Negative Binomial in STAN

```
library(rstan)
library(rstanarm)

ships <- MASS::ships

## Fit the negative binomial model

## Using the MASS package
## OPTIONAL HOMEWORK: Why does MASS vomit when it runs this?
# ships_nb <- MASS::glm.nb(
#    incidents ~ type + factor(year) + factor(period) +
#                  offset(log(service)),
#    data = dplyr::filter(ships, service > 0))

## Equivalent code in STAN
ships_nb_stan <-
  rstanarm::stan_glm.nb(incidents ~ type + factor(year) + factor(period),
                        offset = log(service),
                        data = dplyr::filter(ships, service > 0))
```

```
summary(ships_nb_stan)
```

```
##
## Model Info:
##  function:     stan_glm.nb
##  family:       neg_binomial_2 [log]
##  formula:      incidents ~ type + factor(year) + factor(period)
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 34
##  predictors:   9
##
## Estimates:
##                          mean   sd   10%   50%   90%
## (Intercept)           -6.5    0.6 -7.2  -6.5  -5.8
## typeB                 -0.4    0.4 -1.0  -0.4   0.1
## typeC                 -0.5    0.5 -1.2  -0.5   0.1
## typeD                 -0.2    0.5 -0.8  -0.2   0.5
## typeE                  0.5    0.5 -0.1   0.5   1.1
## factor(year)65         0.7    0.5  0.2   0.7   1.3
## factor(year)70         1.1    0.5  0.5   1.0   1.6
## factor(year)75         0.5    0.5 -0.2   0.5   1.2
## factor(period)75       0.3    0.3 -0.1   0.3   0.7
## reciprocal_dispersion  3.2    1.3  1.7   3.0   4.9
##
## Fit Diagnostics:
##            mean   sd   10%   50%   90%
## mean_PPD 12.6    4.0  8.6  12.0  17.2
##
## The mean_ppd is the sample average posterior predictive distribution of the outcome variable (for deta
##
## MCMC diagnostics
##                       mcse Rhat n_eff
## (Intercept)           0.0  1.0  1890
## typeB                 0.0  1.0  2236
```

Comparison of Standard Errors

```
ships_nb_stan$ses /
  sqrt(diag(vcov(glm(
    incidents ~ type + factor(year) + factor(period), family = poisson,
    offset = log(service), data = ships,
    subset = service > 0
  ))))
```

```
##      (Intercept)            typeB            typeC            typeD
##         2.571512         2.372261         1.607573         1.683826
##            typeE   factor(year)65   factor(year)70   factor(year)75
##         1.946983         2.988722         2.669808         2.279227
## factor(period)75
##         2.602886
```

Exercise: More Negative Binomial

Repeat Exercise 1 (all parts) with the negative binomial model, but use

$$\widehat{\phi} = \frac{1}{N-P} \sum_i \frac{(Y_i - \widehat{\mu}_i)^2}{\widehat{\mu}_i + \widehat{\mu}_i^2/\widehat{k}}.$$

If the negative binomial model is correct, we should have $\widehat{\phi} \approx 1$. Does this model seem to do better than the Poisson?

Warning

Warning!

One issue with the negative binomial model is that the variance grows quite quickly in μ — specifically, we get overdispersion by jumping from a linear relationship between the mean and variance to a quadratic relationship. Some would argue that this is overkill, and that (for large μ) we may start overshooting the variance.

Other Generative Models: Binomial Data

Binomial-type data Z_i is overdispersed relative to the Binomial(n_i, μ_i) distribution if

$$\text{Var}(Z_i) > n_i \, \mu_i (1 - \mu_i).$$

Occurs when we have binomial-type experiments where trials are not independent!

Exercise

Exercise: Beta-Binomial

Suppose that $Z_i \sim \text{Binomial}(n_i, p_i)$ with $p_i \sim \text{Beta}\{\rho\mu_i, \rho(1-\mu_i)\}$.

(a) Show that, marginally, Z_i has mass function

$$f(z; \mu_i, \rho) = \binom{n_i}{z} \cdot \frac{\Gamma(\rho)}{\Gamma(\rho\mu_i)\Gamma(\rho[1-\mu_i])} \cdot \frac{\Gamma(\rho\mu + z)\Gamma(\rho[1-\mu] + n_i - z)}{\Gamma(\rho + n_i)}.$$

This distribution is known as a *beta-binomial distribution*.

(b) Show that $\mathbb{E}(Z_i) = n_i\mu_i$.

(c) Show that, for $n_i > 1$, $\text{Var}(Z_i) > n_i\mu_i(1-\mu_i)$ so that Z_i is overdispersed. *Hint*: like the Poisson setting, you can show that this holds without making use of the fact that p_i has a beta distribution. This will save you from having to compute moments of the beta distribution unnecessarily.

Exercise

Exercise: Rats

Quoting Alan Agresti (Categorical Data Analysis, 3rd Edition, Section 4.7.4):

> *Teratology is the study of abnormalities of physiological development. Some teratology experiments investigate effects of dietary regimens or chemical agents on the fetal development of rats in a laboratory setting. Table 4.7 shows results from one such study (Moore and Tsiatis 1991). Female rats on iron-deficient diets were assigned to four groups. Rats in group 1 were given placebo injections, and rats in other groups were given injections of an iron supplement; this was done weekly in group 4, only on days 7 and 10 in group 2, and only on days 0 and 7 in group 3. The 58 rats were made pregnant, sacrificed after three weeks, and then the total number of dead fetuses was counted in each litter. Due to unmeasured covariates and genetic variability the probability of death may vary from litter to litter within a particular treatment group.*

The data can be obtained by running the following commands.

```
rats_path <- paste0("https://raw.githubusercontent.com/theodds/",
                    "SDS-383D/main/rats.csv")
rats <- read.table(rats_path, sep = "\t", header = TRUE)
head(rats)
```

```
##   litter group  n  y
## 1      1     1 10  1
## 2      2     1 11  4
## 3      3     1 12  9
## 4      4     1  4  4
## 5      5     1 10 10
## 6      6     1 11  9
```

Our interest is in the relationship between the treatment `group` and the number of dead fetuses. As this is our first treatment of *binomial* (as opposed to *Bernoulli*) data, I will show how to fit the a binomial glm:

Quasi Likelihood

Quasi-likelihood models replace the likelihood with the *quasi-likelihood*

$$q(y \mid \mu, \phi) = \exp\left\{\int_y^\mu \frac{y-t}{\phi V(t)}\, dt\right\}$$

which encodes the moment conditions

$$\mathbb{E}(Y_i \mid \mu_i) = \mu_i \qquad \text{and} \qquad \text{Var}(Y_i \mid \mu_i) = \phi\, V(\mu_i).$$

We don't specify an exponential dispersion family, just a link function $g(\cdot)$ and variance function $V(\cdot)$.

The Quasi Score and Quasi Fisher Information

Score function is

$$s(\beta) = \sum_i \frac{\omega_i (Y_i - \mu_i)\, X_i}{\phi\, V(\mu_i) g'(\mu_i)}.$$

Should look familiar! Similarly, Fisher information is $\frac{X^\top W X}{\phi}$.

Examples of Quasi-Likelihood Methods

- Quasi-Poisson:

$$V(\mu) = \mu.$$

Allows us to use a Poisson-like model without assuming $\phi \equiv 1$.

- Quasi-Binomial:

$$V(\mu) = \mu(1 - \mu).$$

Allows us to use a binomial-like model without assuming $\phi \equiv 1$.

Ships Again

We can fit the quasi-Poisson model to the ships dataset with the following commands.

```
ships <- MASS::ships
quasi_ships <- glm(incidents ~ type + factor(year) + factor(period),
                   family = quasipoisson,
                   data = ships,
                   offset = log(service),
                   subset = (service != 0))
summary(quasi_ships)
```

```
##
## Call:
## glm(formula = incidents ~ type + factor(year) + factor(period),
##     family = quasipoisson, data = ships, subset = (service !=
##         0), offset = log(service))
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -6.40590    0.28276 -22.655  < 2e-16 ***
## typeB            -0.54334    0.23094  -2.353  0.02681 *
## typeC            -0.68740    0.42789  -1.607  0.12072
## typeD            -0.07596    0.37787  -0.201  0.84230
## typeE             0.32558    0.30674   1.061  0.29864
## factor(year)65    0.69714    0.19459   3.583  0.00143 **
## factor(year)70    0.81843    0.22077   3.707  0.00105 **
## factor(year)75    0.45343    0.30321   1.495  0.14733
## factor(period)75  0.38447    0.15380   2.500  0.01935 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 1.691028)
##
##     Null deviance: 146.328  on 33  degrees of freedom
## Residual deviance:  38.695  on 25  degrees of freedom
## AIC: NA
##
```

Ships Again

```
anova(quasi_ships, test = "F")
```

```
## Analysis of Deviance Table
##
## Model: quasipoisson, link: log
##
## Response: incidents
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev      F    Pr(>F)
## NULL                           33    146.328
## type            4   55.439        29     90.889 8.1961 0.0002289 ***
## factor(year)    3   41.534        26     49.355 8.1871 0.0005777 ***
## factor(period)  1   10.660        25     38.695 6.3039 0.0188808 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Ticks Again

Exercise: Ticks Revisited

Apply the quasi-Poisson model to the `ticks` dataset.

(a) How do the standard errors from the quasi-Poisson model compare to the standard errors you get from (i) Poisson log-linear model, (ii) the negative binomial model, and (iii) the nonparametric bootstrap?

(b) How do the regression coefficient estimates from the quasi-Poisson model compare to the estimates from the Poisson log-linear model. Can you explain the relationship you see?

(c) The `robust` function in the `sjstats` package computes *robust* standard errors for a variety of models in `R` based on the sandwich matrix construction of the variance of the M-estimators; this has the advantage of being generally correct, without even requiring the variance assumption to be correct (but does not allow for an extension of analysis of deviance). How do these standard errors compare to the quasi-likelihood standard errors? What about to the nonparametric bootstrap?

The Quasi-Binomial Model

The quasi-Binomial model makes use the assumptions

$$g(\mu_i) = x_i^\top \beta \qquad \text{and} \qquad \text{Var}(Y_i \mid X_i) = \phi\, \mu_i(1-\mu_i)/n_i,$$

so that the same variance function $V(\mu)$ as the binomial model is used. The quasi-binomial model can be fit in the same as the quasi-Poisson (just change the family `quasipoisson` to `quasibinomial`).

Exercise: Rats Revisited

Apply the quasi-binomial model to the `rats` dataset. Are the results consistent with the results you got from the binomial and beta-binomial models? What about if you use `robust` standard errors instead?

Why Quasi-Likelihood Works

Suppose $Z_1, \ldots, Z_N \overset{\text{iid}}{\sim} F_0$ for some F_0 and we want to estimate $\beta_0 = \beta(F_0)$. An *estimating equation* for estimating β_0 is given by the $\widehat{\beta}$ that solves

$$\frac{1}{N} \sum_i m(Z_i; \beta) = 0$$

where $m(z; \beta)$ is such that

$$\mathbb{E}_{F_0}\{m(Z_i; \beta)\} = 0 \iff \beta = \beta_0.$$

$\widehat{\beta}$ is referred to as an M-estimator.

M-Estimators

Exercise: M Estimators

Let $\beta = \beta(F)$ be a parameter of interest and let $\beta_0 = \beta(F_0)$ denote its true value. Let $m(z;\beta)$ be a function taking values in $\mathbb{R}^P$ where $P = \dim(\beta)$ such that $\mathbb{E}\{m(Z;\beta)\} = 0$ only when $\beta = \beta_0$. We define the *M-estimator* of β_0 via the *estimating equation*

$$\frac{1}{N}\sum_{i=1}^{N} m(Z_i;\widehat{\beta}) = 0,$$

solving the "finite-sample" version of the population equation $\mathbb{E}\{m(Z_i;\beta_0)\} = 0$.

Informally, argue that the asymptotic distribution of $\widehat{\beta}$ is

$$\widehat{\beta} \overset{\bullet}{\sim} \text{Normal}(\beta_0, V_N),$$

where the covariance matrix V_N is given by the *sandwich matrix* $B_N^{-1} C_N B_N^{-\top}/N$ with

$$B_N = -\mathbb{E}\{m'(Z_1;\beta_0)\} \quad \text{and} \quad C_N = \mathbb{E}\{m(Z_i;\beta_0)\, m(Z_i;\beta_0)^\top\},$$

and where $m'(z;\beta) = \frac{\partial}{\partial\beta} m(z;\beta)$ is the Jacobian matrix of $m(z;\beta)$ with respect to β. Then, propose estimators for B_N and C_N that can be used in practice.

Hint: Taylor expand $N^{-1}\sum_{i=1}^{N} m(Z_i;\beta_0)$ about $\widehat{\beta}$ and ignore the remainder.

Exercise

Exercise: Misspecified MLE

Suppose that $Z_1, \ldots, Z_N \overset{\text{iid}}{\sim} F_0$ and we base inference on a working parametric family $\{F_\theta : \theta \in \Theta\}$ which happens to be incorrect (i.e., $F_0 \notin \{F_\theta\}$). Using the M-estimation framework, show that the MLE of θ is (under the unstated assumptions that make M-estimation valid) still asymptotically normal, centered at the solution $\theta^\star$ of the score equation

$$\mathbb{E}\left\{ s(\theta^\star; Z_1) \right\} = 0,$$

and derive the form of the asymptotic covariance matrix of $\widehat{\theta}$. How does this differ from the usual asymptotic variance?

Hint: when the model is misspecificed, there is a simplification which *does not occur*.

Sandwich

Exercise: Sandwich Matrix

Show that the components of the sandwich matrix for the quasi-likelihood model are given by

$$B_N = \frac{1}{\phi N} X^\top W X \quad \text{and} \quad C_N = \frac{1}{\phi N} X^\top W^\star X$$

where

$$W = \operatorname{diag}\left\{ \frac{\omega_i}{V(\mu_i)\, g'(\mu_i)^2} \right\} \quad \text{and} \quad W^\star = \operatorname{diag}\left\{ \frac{\operatorname{Var}(Y_i \mid X_i)}{[V(\mu_i)\, g'(\mu_i)/\omega_i]^2} . \right\}$$

Show also that, when our assumption about the variance $\operatorname{Var}(Y_i \mid X_i) = \frac{\phi}{\omega_i} V(\mu_i)$ is correct then this simplifies to $\phi(X^\top W X)^{-1}$.

What Makes Quasi-Likelihood Special

The M-estimator asymptotics above limit us mostly to Wald-based and score-based inference. Quasi-likelihoods also give us likelihood-based methods.

- quasi log-likelihood

$$\ell(\beta) = \sum_{i=1}^{N} \frac{\omega_i}{\phi} \int_{Y_i}^{\mu_i} \frac{Y_i - t}{V(t)}\, dt,$$

- quasi-deviance

$$D = -2\phi\ell(\widehat{\beta}).$$

- Can test nested models using an F-statistic:

$$F = \frac{(D_0^\star - D_1^\star)/(p-q)}{\widehat{\phi}/\phi} = \frac{D_0 - D_1}{(P-R)\widehat{\phi}} \approx F_{D,N-P}$$

where D is difference in model dimensions of nested models $\mathcal{M}_0 \subseteq \mathcal{M}_1$.

Exercise

Exercise: Quasi-Poisson

Show for the Poisson loglinear model that this does indeed recover the correct likelihood, up-to a normalizing constant.

Other Approaches

Possibilities:

1. Drop the variance assumption $\text{Var}(Y_i \mid X_i) = \phi\, V(\mu_i)$ for some known function $V(\cdot)$.
2. Drop the assumption that $g(\mu_i) = X_i^\top \beta$ for some parameter vector β.

First setting: estimator of β will still be consistent, but might not be *efficient*. Can still use sandwich matrix for the variance, or perform score-like inference, but no immediate likelihood equivalent...

Empirical Likelihood

Definition (Empirical Likelihood)

The *profile empirical likelihood* of β is given by

$$\ell_{\text{EL}}(\beta) = \max \left\{ \prod_{i=1}^{N} p_i : \sum_i p_i \frac{\omega_i (Y_i - \mu_i) X_i}{\phi V(\mu_i) g'(\mu_i)} = 0, p_i \geq 0, \sum_i p_i = 1 \right\}.$$

From here it is possible to prove a version of Wilk's theorem that allows us to build likelihood-based intervals, perform hypothesis tests, and so forth, while invoking minimal assumptions.

Assumption Free Methods?

Exercise: Asymptotic Distribution of the MLE Under Total Misspecificaiton

Argue that, when a GLM (with known ϕ) is misspecified, the parameter β we estimate corresponds to

$$\beta \equiv \arg\max_{\beta} \int \log f(y \mid \beta, x)\, f_0(y, x)\; dy\; dx$$

where $f_0(x, y)$ is the true joint density of (X_i, Y_i). This parameter corresponds to the so-called Kullback-Leibler projection of $f_0(y \mid x)$ onto the family $\{f(y \mid x, \beta, \phi) : \beta \in \mathbb{R}\}$.

Next, show that when the GLM is just a linear regression that the above β corresponds to $\min_\beta \mathbb{E}[\{r_0(X_i) - X_i^\top \beta\}^2]$ where $r_0(X_i) = \mathbb{E}(Y_i \mid X_i)$ is the true regression function; that is, $x^\top \beta$ is the *best linear approximation* to $r_0(x)$ (with respect to the distribution of X_i).