Week 6 Notes: Robust GLMs

1 Motivation

The GLMs discussed up to this point, as well as the approaches to inference we saw, form a nice starting point for an analysis. Inevitably, however, the models we use in practice will be misspecified.

On the point of model misspecification, George Box famously said "all models are wrong, but some are useful" For example, Box states in one of his books

> . . . 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. . .

While the first part of this quote gets the most play, I think it is important to also bear in mind the second part of the quote: **once we understand that our model is an approximation, we need to understand where the approximation breaks down.**

This is particularly important for likelihood-based and Bayesian methods, because it is very frequently the case that

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. A particular case of this occurs when the outcome is overdispersed.

Towards this end, in this batch of notes we will consider the following 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?

We begin with the topic of *quasi likelihood* methods, which are typically used to deal with *overdispersion* in a GLM. This will serve as motivation for other techniques based on robust and semiparametric inference.

2 Overdispersion

Overdispersion is a common phenomenon for count/binomial type data. By “binomial type” data, I just mean the usual setting of repeated success/failure experiments with each experiment having success probability p, but possibly without the assumption of independence between trials and with only the number of success reported.

For GLMs we know that $\text{Var}(Y_i \mid X_i) = \frac{\phi}{\omega_i} V(\mu_i)$, so that ϕ is linked directly to the variability in Y_i. For the Poisson and Binomial GLMs, we know exactly that $\phi \equiv 1$; overdispersion relative to these models occurs when $\text{Var}(Y_i \mid X_i) > \omega_i^{-1} V(\mu_i)$.

The problem introduced by overdispersion, as such, is primarily that it screws up our uncertainty quantification for the regression coefficients. The asymptotic variance of the regression coefficients in a GLM is given by

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

where W depends on the means μ_i and the weights ω_i, but notably not on ϕ. If the self-imposed value $\phi = 1$ is too small, however, we will end up underestimating $\text{Var}(\widehat{\beta})$. This implies that failure to account for overdispersion has a material consequence: we will end up *underestimating our uncertainty in* β! This will cause (for example) our confidence intervals to not have nominal coverage levels and our hypothesis tests to have higher than nominal error rates.

Exercise 1: 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.

d. Do a **nonparametric bootstrap** to approximate the standard error of the regression coefficients (i.e., sample the rows of the `data.frame` with replacement and compute the MLE of the β's for each resample). How do these compare with the variance estimates produced by the Poisson loglinear model?

e. The test above is predicated on the assumption that the structure of the mean model is correctly specified; if the structure of $\log(\mu_{ij})$ is incorrect, this can manifest in large values of $\widehat{\phi}$, even if there is no overdispersion. To assess this, fit the model

$$\log(\mu_{ij}) = \alpha_j + \beta_{1j} \times \texttt{HEIGHT}_i + \beta_{2j} \times \texttt{HEIGHT}_i^2 + \beta_{3j} \times \texttt{HEIGHT}_i^3.$$

   Does $\widehat{\phi}$ still seem to be too large for this bigger model?

Other Generative Models: Count Data

The most obvious strategy for dealing with overdispersion is to use a bigger (but still parametric) probabilistic model that can accommodate for overdispersion. The simplest models to use are the *negative binomial* model for count data and the *beta-binomial* model for binomial-type data.

The negative binomial model sets $[Y_i \mid \mu_i, k] \sim f(y \mid \mu_i, k)$ where $f(y \mid \mu, k)$ is a negative binomial mass function

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

Exercise 2: 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 regression in STAN can be fit as follows:

```r
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))
```

```r
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   0.4
typeE                  0.5    0.5 -0.1   0.5   1.1
factor(year)65         0.7    0.5  0.1   0.7   1.3
factor(year)70         1.0    0.5  0.5   1.0   1.6
factor(year)75         0.5    0.6 -0.2   0.5   1.2
factor(period)75       0.3    0.3 -0.1   0.3   0.7
reciprocal_dispersion  3.2    1.4  1.7   3.0   5.1

Fit Diagnostics:
           mean   sd   10%   50%   90%
mean_PPD 12.5    4.0  8.5  11.7  17.1

The mean_ppd is the sample average posterior predictive distribution of the outcome variable

MCMC diagnostics
                      mcse Rhat n_eff
(Intercept)           0.0  1.0  1984
typeB                 0.0  1.0  2061
typeC                 0.0  1.0  2741
typeD                 0.0  1.0  2999
typeE                 0.0  1.0  2853
factor(year)65        0.0  1.0  2345
factor(year)70        0.0  1.0  2250
factor(year)75        0.0  1.0  2536
factor(period)75      0.0  1.0  3566
reciprocal_dispersion 0.0  1.0  2581
mean_PPD              0.1  1.0  3093
log-posterior         0.1  1.0  1290

For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effectiv
```

The default prior that `rstanarm` claims to use is $k \sim \text{Gam}(1, s_Y)$ where s_Y is an estimate of the standard deviation of Y, which does not seem to be based on any kind of careful reasoning about the problem. The catch with the Bayesian analysis for this problem is that, because $k \to \infty$ corresponds to a Poisson model, our choice of prior is unfortunately rather informative (i.e., we shouldn't take a value of $k \ll \infty$ to be evidence that the model is not Poisson). A better thing to do here would be to reparameterize the model using something like $\nu = k/(k+1)$ so that $\nu = 1$ corresponds to $k = \infty$, the Poisson model (**Note:** I haven't tried this).

We do see that the negative binomial model produces somewhat higher standard errors than the Poisson loglinear model due to it accounting for overdispersion:

```
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 
        2.513202         2.345921         1.558028         1.739285 
           typeE   factor(year)65   factor(year)70   factor(year)75 
        1.931022         3.019812         2.607449         2.316915 
factor(period)75 
        2.575268 
```

Exercise 3: 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!

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.

Curiously (and possibly related to the above point, although this is just speculation), negative binomial regression is not built into base R; instead a competing method for handling overdispersion (quasi-Poisson regression) is.

Other Generative Models: Binomial Data

We call binomial-type data Z with mean np is overdispersed if the variance is larger than would be suggested by the binomial sampling model, i.e., $\text{Var}(Z) > np(1-p)$. There is a model analogous to the negative binomial response model for binomial data. Binomial regression is based on the modeling assumption $Z_i \sim \text{Binomial}(n_i, p_i)$, which models an experiment which is *independently* replicated n_i times with success probability p_i. In many cases where binomial-type data occurs, we won't have strong evidence for the *independence* assumption underlying the binomial distribution. In addition to introducing the beta-binomial model, the following exercise illustrates that overdispersion relative to the binomial model holds very generally.

Exercise 4: 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 5: 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.

```r
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:

```r
rats_binomial <- glm(cbind(y, n - y) ~ factor(group),
                     family = binomial,
                     data = rats)
```

The response is given in two columns: the number of successes and number of failures for each observation (we did not need to do this with Bernoulli data since the number of trials is always 1).

(a) Based on the fit of the binomial model, do rats in the placebo group appear to have a fewer proportion of dead fetuses? Justify your conclusions by appropriately accounting for uncertainty.

(b) Using the same strategies you used for the Poisson, assess whether this data is overdispersed relative to the binomial distribution (make the necessary modifications to $\widehat{\phi}$).

(c) Use the `aod` package to fit a beta-binomial model to the data. Do your qualitative conclusions change? How does this choice affect the standard errors of the effects of interest?

3 Semiparametric Modeling with Quasi Likelihood

In some sense, the problem with overdispersion for count and proportion data is that we are restricted to having $\phi = 1$. We might instead be better served by some other value $\phi > 1$.

For the Poisson model this would allow us (say) to take $\mathrm{Var}(Y_i) = 2\mu_i$ rather than forcing $\mathrm{Var}(Y_i) = \mu_i$.

One might hope that we could write down an exponential dispersion family which has $\phi \neq 1$ but is otherwise like the Poisson model in that $b(\theta) = e^\theta$. Unfortunately this is not possible. Oddly, while a density/mass function of the form above may not exist, we can still develop useful methodology as though it did.

As the idea of using a probabilistic model which "doesn't exist" might seem concerning to reasonable people, before proceeding we will present a general tool for constructing estimators from moment equations.

M-Estimators

Consider iid random vectors $Z_1, \ldots, Z_N \stackrel{\text{iid}}{\sim} F_0$ for some distribution F_0. Often we will only be interested in a parameter $\beta = \beta(F)$ rather than the whole distribution F.

Exercise 6: 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

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

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 7: Misspecified MLE

Suppose that $Z_1, \ldots, Z_N \stackrel{\text{iid}}{\sim} F_0$ 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*.

Quasi-Likelihood Estimating Equations

Rather than basing our inferences on a parametric model, we instead directly impose the moment conditions

$$\mathbb{E}(Y_i \mid X_i = x_i) = \mu_i,$$
$$\operatorname{Var}(Y_i \mid X_i = x_i) = \frac{\phi}{\omega_i} V(\mu_i).$$

where $g(\mu_i) = x_i^\top \beta$ and $V(\mu_i)$ are specified by the user. Notice that, rather than specifying a parametric family for Y_i, we are instead specifying a relationship between the mean and the variance directly and avoiding making any assumptions about the distribution of Y_i beyond that.

The jumping off point for quasi-likelihood methods is to treat the score equations of the Poisson or Binomial GLMs as estimating equations. Given the above moment restrictions, and motivated by the likelihood equations of a GLM, we define $\widehat{\beta}$ to be the solution to the estimating equation

$$\frac{1}{N} \sum_{i=1}^N \frac{\omega_i (Y_i - \mu_i) X_i}{\phi V(\mu_i) g'(\mu_i)} \stackrel{\text{set}}{=} \mathbf{0}.$$

The associated population-level equation is

$$\mathbb{E}\left\{\frac{1}{N} \sum_i \frac{\omega_i (Y_i - \mu_i) X_i}{\phi\, V(\mu_i)\, g'(\mu_i)} \mid X_1, \ldots, X_N\right\} = \mathbf{0}$$

which occurs when $\mathbb{E}(Y_i \mid X_i) = \mu_i$ for all i, i.e., at β_0. Note that, interestingly, this estimating equation has mean $\mathbf{0}$ even when our assumption about the form of $\operatorname{Var}(Y_i \mid X_i = x_i)$ is incorrect.

Exercise 8: 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 $\phi(X^\top W X)^{-1}$.

As sketched out above, we have presented quasi-likelihood as just an M-estimation technique. There is an added twist with quasi-likelihood methods, however: analysis of deviance techniques can be used *as though* we were using likelihood-based methods! To do this, we basically just "pretend" that the estimating equation $\sum_i \omega_i (Y_i - \mu_i) X_i / [\phi V(\mu_i) g'(\mu_i)] = \mathbf{0}$ is the score function of a bona-fide exponential dispersion family. We make use of an associated quasi (i.e., "pretend") density/mass function

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

When our estimating equation *does* correspond to a score equation, this is a valid way to reverse-engineer the density/mass function (up-to a normalizing constant).

> **Exercise 9: Quasi-Poisson**
>
> Show for the Poisson loglinear model that this does indeed recover the correct likelihood, up-to a normalizing constant.

We can then define the quasi log-likelihood function

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

which tends to behave as a likelihood function should. The quasi-scaled deviance $D^\star$ is given by $-2\ell(\widehat{\beta})$, and from this we can apply the usual analysis of deviance with the estimate

$$\widehat{\phi} = \frac{1}{N - P} \sum_{i=1}^N \frac{\omega_i (Y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}.$$

We then have the usual fact that if $\mathcal{M}_0$ (of dimension R) is nested in model $\mathcal{M}_1$ (of dimension P) then $D_0^\star - D_1^\star$ is asymptotically χ^2_{P-R}. Of course, ϕ is unknown so we cannot compute $D_0^\star$ and $D_1^\star$; instead we look at

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

which, by analogy with ANOVA, is compared with an an $F_{P-R,N-P}$ distribution.

The Quasi-Poisson Model

The quasi-Poisson model makes the assumptions

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

so that the same variance function $V(\mu)$ is used, but we now allow for ϕ to differ from 1.

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
```

```
Number of Fisher Scoring iterations: 5
```

The usual functions work here; for example, we can do analysis of deviance as follows.

```
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
```

Exercise 10: 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

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

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 11: 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?

4 Other Approaches to Robust Inference

If we are unsatisfied with quasi-likelihood methods or the alternative generative models, there are other weaker sets of assumptions we might use:

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

Dropping the first assumption is not so bad. For example, we might specify a model of the form

$$\mathbb{E}(Y_i \mid X_i = x, \beta) = g^{-1}(X_i^\top \beta),$$

in which case the estimator defined by the solution to

$$\frac{1}{N} \sum_i \frac{\omega_i (Y_i - \mu_i)\, X_i}{\phi\, V(\mu_i)\, g'(\mu_i)} = \mathbf{0}$$

still usually produces a consistent estimator for β, where $\frac{\phi}{\omega} V(\mu)$ is instead thought of as a *working variance* model. This is just the estimators we have been studying for GLMs all along, but now pointing out that the estimate is valid even if we get the variance relationship incorrect entirely; getting $V(\mu)$ correct only improves the *efficiency* of the estimator, so we shouldn't neglect it entirely, but it is reassuring that the coefficient estimates obtained from a GLM are, broadly speaking, still valid as long as we get the mean relationship correct.

5 Optional: Empirical Likelihood

The selling point of quasi-likelihood, relative to this weaker set of assumptions, I think is that one can get something resembling likelihood-based inference from the quasi-likelihood. This is not possible with the robust standard errors, where we are mostly limited to something resembling Wald-type methods.

A natural follow up question is then "is it possible to get something that behaves like a likelihood function even if we don't make any assumptions about the variance?" It turns out that the answer is "yes, by using the *empirical likelihood.*"

Definition 1: 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.

Exercise 12: Empirical Likelihood

Use the `melt` package (see here) to fit the `ships` dataset using empirical likelihood. How do the confidence intervals compare to those obtained from quasi-likelihood?

6 Assumption Free Methods?

A final approach we can take is to drop the assumption that *either* the mean or variance is correctly specified. In this case, it is not clear what we are even estimating from a scientific perspective. We can still ask, however, what the behavior of the MLE is when the model is totally incorrect.

Exercise 13: 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).