Overdispersion and Quasilikelihood

Objectives:

- Mechanisms that generate overdispersion.
- Analysis with overdispersion
  - Model the distribution
  - Model the variance
  - Correct standard errors
- Quasilikelihood and estimating equations.

Overdispersion and Quasilikelihood

● Recall that when we used Poisson regression to analyze the seizure data that we found the var$(Y_i) \approx 2.5 \times \mu_i$.

Define: **Overdispersion** describes the situation above. That is, data are overdispersed when the actual var(Y_i) exceeds the GLM variance $\phi V(\mu)$.

● For Binomial and Poisson models we often find overdispersion

1. Binomial: $Y = s/m$,

$$E(Y) = \mu,$$
$$\text{var}(Y) > \mu(1-\mu)/m.$$

2. Poisson: $E(Y) = \mu$,

$$\text{var}(Y) > \mu.$$

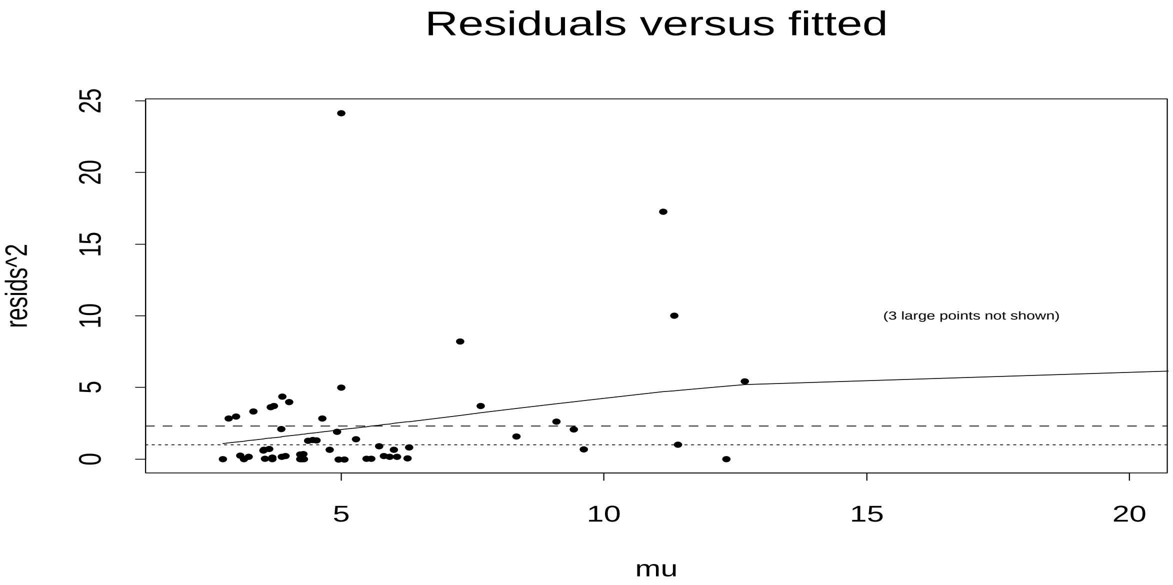
Overdispersion and Quasilikelihood

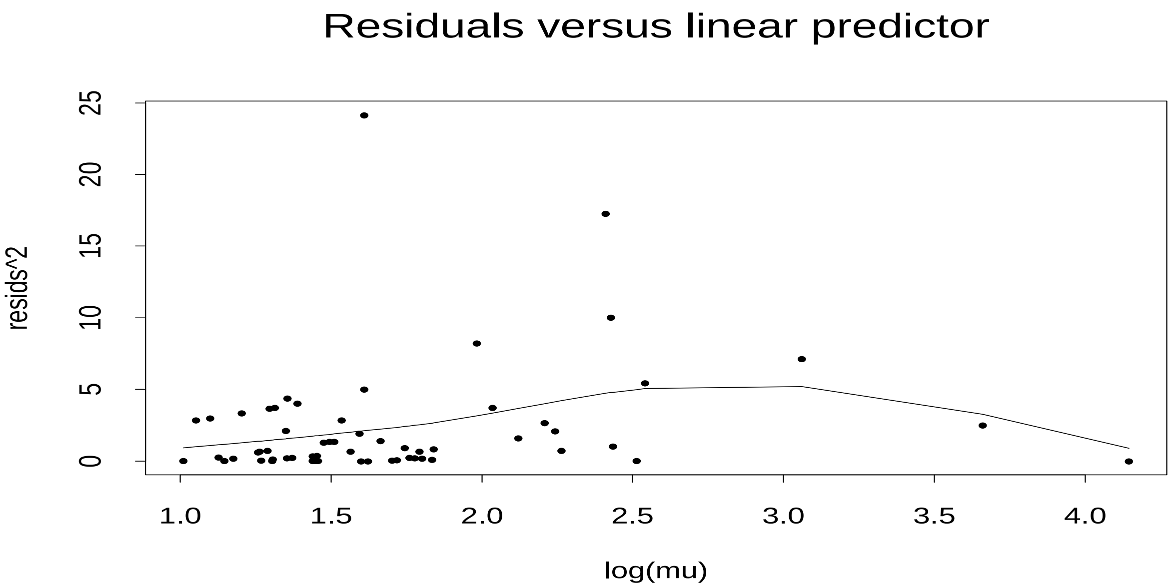
Q: How does overdispersion arise?

Q: What is the impact on the GLM regression estimate $\widehat{\beta}$ and the associated inference?

Q: What expanded methods exist that model, or correct for, overdispersion?

Seizure Data:





Example: Teratology Data

Low-Iron Teratology Data

Shepard, Mackler & Finch (1980)
(analyzed by Moore & Tsiatis, 1991)

- Female rats were put on iron deficient diets and divided into four groups. The groups received different numbers of injections to maintain iron levels:

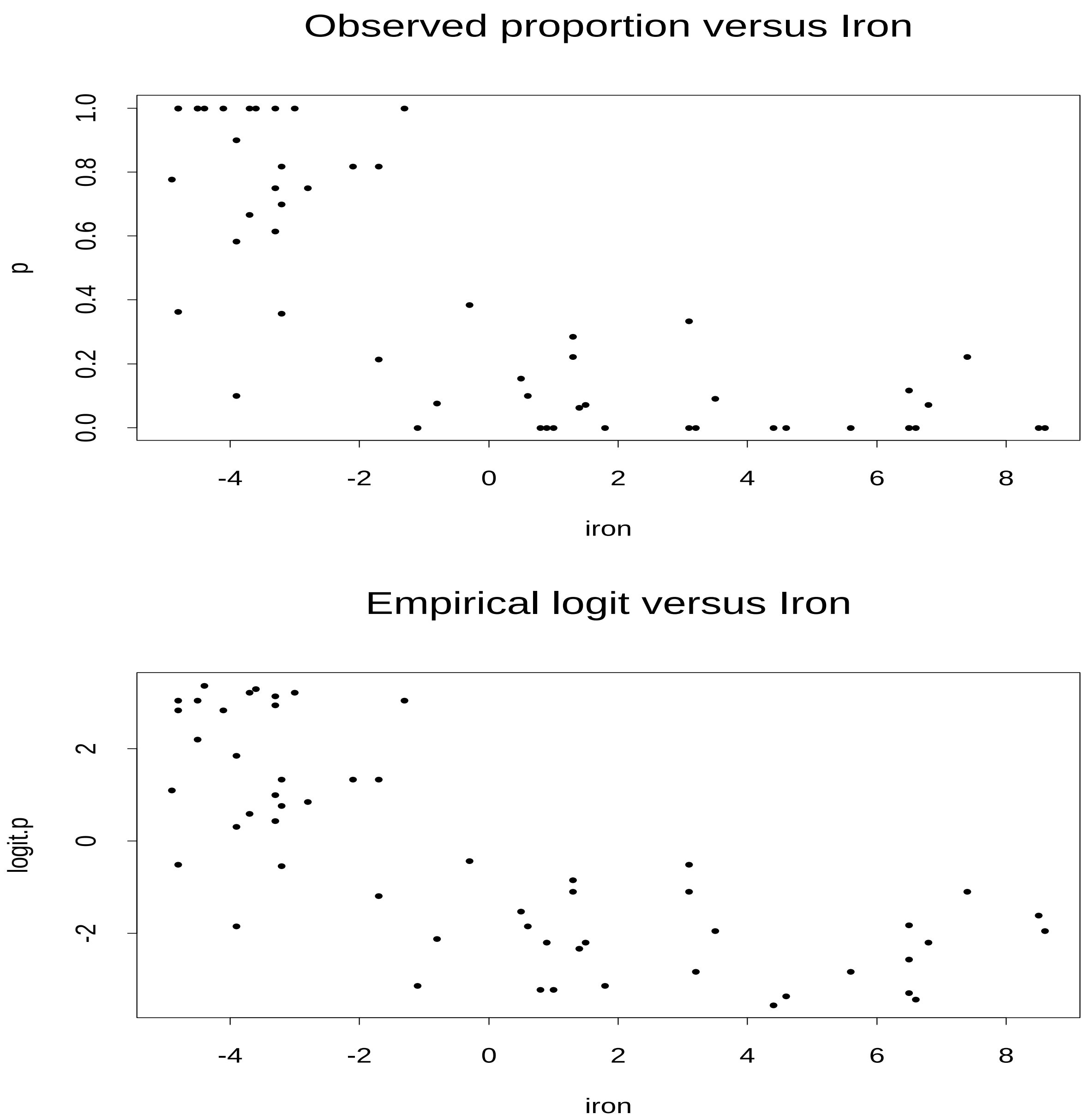
group	description
group 4 (normal)	3 injections (weekly for 3 weeks)
group 3	2 injections (days 0 and 7)
group 2	1 injection (day 7 or 10)
group 1	placebo injections

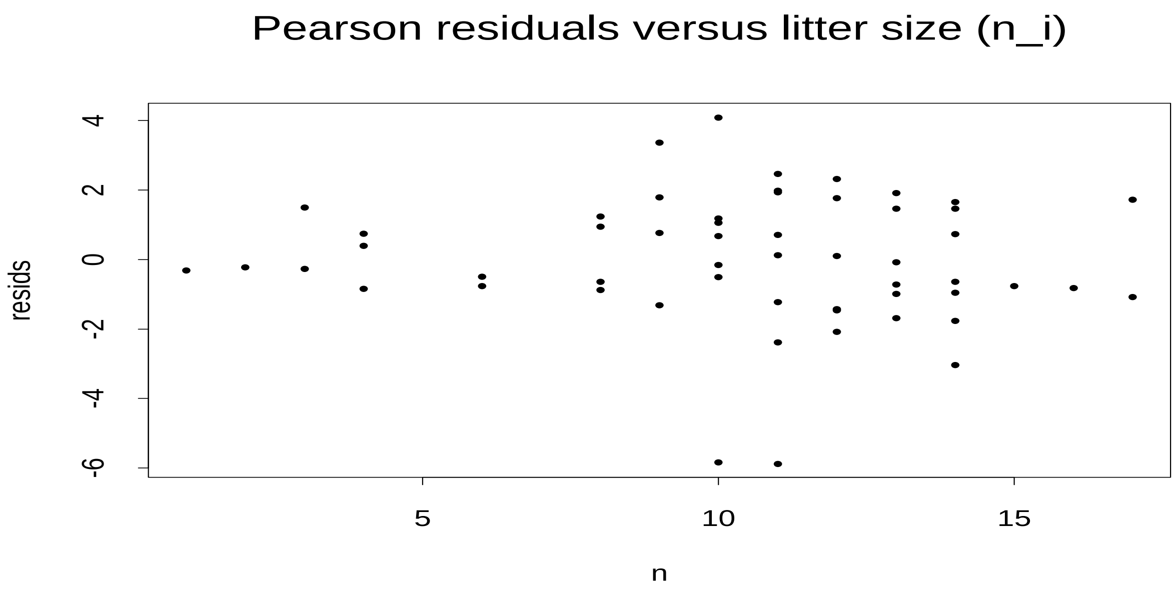
Example: Teratology Data

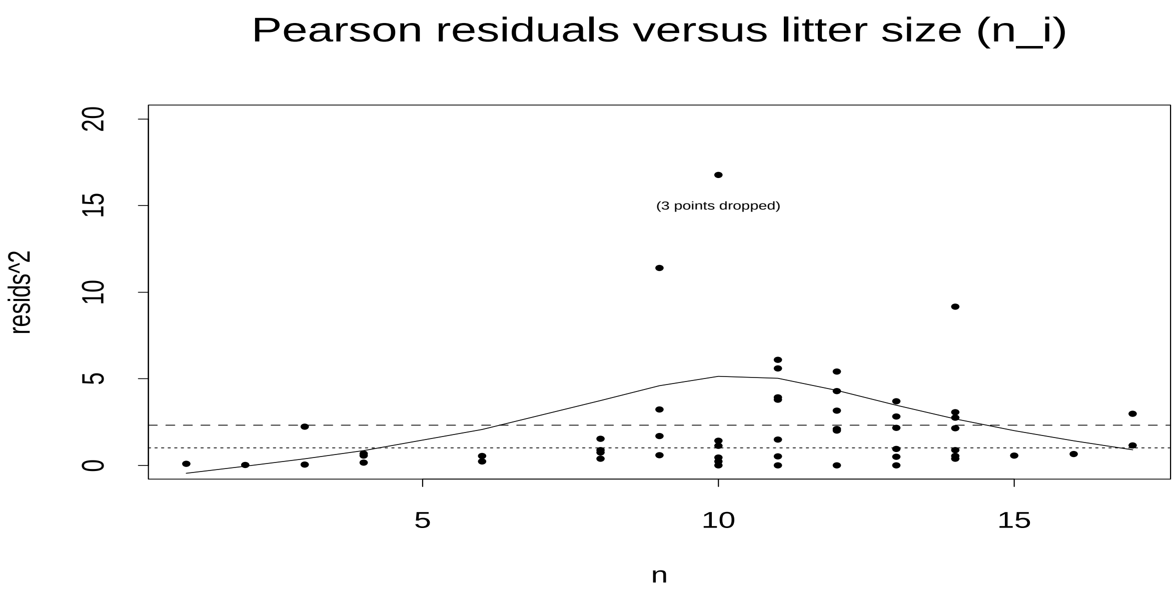
- Rats were pregnant and sacrificed at 3 weeks. For each animal the total number of fetuses (N_i) and the number of dead fetuses (Y_i) were recorded.

- The hemoglobin (blood iron) levels were recorded.

- These data are typical of studies of the of chemical agents or dietary conditions on fetal development.


Observed proportion versus Iron
p
1.0
0.8
0.6
0.4
0.2
0.0
-4
-2
0
2
4
6
8
iron
Empirical logit versus Iron
logit.p
2
0
-2
-4
-2
0
2
4
6
8
iron


Pearson residuals versus litter size (n_i)
resids
4
2
0
-2
-4
-6
5
10
15
n


Pearson residuals versus litter size (n_i)
resids^2
(3 points dropped)
20
15
10
5
0
5
10
15
n

Overdispersion?

⋆ If there is population **heterogeneity** then this can introduce overdispersion.

Example: Suppose there exists a binary covariate, Z_i, and that

$$
\begin{aligned}
Y_i \mid Z_i = 0 &\sim \text{Poisson}(\lambda_0) \\
Y_i \mid Z_i = 1 &\sim \text{Poisson}(\lambda_1)
\end{aligned}
$$

$$
P(Z_i = 1) = \pi
$$

$$
E(Y_i) = \pi\lambda_1 + (1 - \pi)\lambda_0 \quad = \mu
$$

Overdispersion?

$$\begin{aligned}\text{var}(Y_i) &= E(\ \lambda_1 Z_i + \lambda_0(1 - Z_i)\ )\ + \\ &\quad \text{var}(\ \lambda_1 Z_i + \lambda_0(1 - Z_i)\ ) \\ &= \mu\ +\ (\lambda_1 - \lambda_0)^2 \pi(1 - \pi)\end{aligned}$$

Therefore, if we do not observe Z_i then this omitted factor leads to increased variation.

Impact of Model Violation

⋆ Huber (1967) and White (1982) studied the properties of MLEs when the model is misspecified.

Setup:

1. Let $F_{\boldsymbol{\theta}}$ be the <u>assumed</u> distributional family for independent data $Y_i, \quad i = 1, 2, \ldots, n$.
2. Let $\widehat{\boldsymbol{\theta}}_n$ be the MLE (based on n observations). That is, $\widehat{\boldsymbol{\theta}}_n$ solves the score equations that arise from the assumption $F_{\boldsymbol{\theta}}$:

$$\sum_{i=1}^{n} \boldsymbol{U}_i^F(\widehat{\boldsymbol{\theta}}_n) = 0$$

3. However, the true distribution of Y_i is given by $Y_i \sim G$.

Result:

1. $\widehat{\boldsymbol{\theta}}_n \rightarrow \boldsymbol{\theta}^*$ such that

$$\lim_n E_G \left[\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{U}_i^F(\boldsymbol{\theta}^*) \right] = 0 \quad (\star\star\star)$$

2. The estimator $\widehat{\boldsymbol{\theta}}_n$ is asymptotically normal:

$$\sqrt{n} \left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^* \right) \rightarrow \mathcal{N}(\ 0,\ A^{-1} B\ A^{-1})$$

$$A = -\lim \frac{1}{n} \sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \boldsymbol{U}_i^F(\boldsymbol{\theta})|_{\boldsymbol{\theta}^*}$$

$$B = \lim \frac{1}{n} \sum_{i=1}^{n} \text{var} \left[\boldsymbol{U}_i^F(\boldsymbol{\theta})|_{\boldsymbol{\theta}^*} \right]$$

Note:

1. A is just the observed information (times $1/n$).
2. B is just the true variance of $\boldsymbol{U}_i^F(\boldsymbol{\theta})$ which may no longer be equal to minus the expected derivative of $\boldsymbol{U}_i^F(\boldsymbol{\theta})$ if $Y_i \sim F_{\boldsymbol{\theta}}$ doesn't hold.
3. Sometimes we get lucky (or we've selected $\widehat{\boldsymbol{\theta}}_n$ wisely) and $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$ – the model misspecification doesn't hurt the consistency of $\widehat{\boldsymbol{\theta}}_n$.
4. Sometimes we get lucky and (3) holds, and $A = B$ and the model misspecification doesn't hurt our standard error estimates either.

Huber, White, & GLMs

Q: What does this misspecification story mean for GLMs?

A: It says that if we are modelling the mean $E(Y_i) = \mu_i$ via a regression model, $\boldsymbol{\beta}$, then our estimator, $\widehat{\boldsymbol{\beta}}$, will converge to whatever value solves

$$E_G\left[\boldsymbol{U}(\boldsymbol{\beta})\right] \;=\; 0$$

Recall that we have

$$\boldsymbol{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right)^T \left[a_i(\phi) \cdot V(\mu_i) \right]^{-1} (Y_i - \mu_i)$$

$\Rightarrow$ As long as $Y_i \sim G$ such that $E_G(Y_i) = \mu_i$ then our regression estimator will be consistent! We don't need Poisson, or Binomial for the GLM point estimate $\widehat{\boldsymbol{\beta}}$ to be valid.

Huber, White & GLMs

Comment:

⋆ We obtain properties of $\widehat{\boldsymbol{\beta}}$ by considering properties (expectation, variance) of $\boldsymbol{U}(\boldsymbol{\beta})$...

Q: Yes, but what about the standard errors for $\widehat{\boldsymbol{\beta}}$?

A: Depends... Let's consider the matrices A and B:

Teratology and Binomial Overdispersion

⋆ Binomial model information matrix

$$A = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_i^T \boldsymbol{V}_i \boldsymbol{X}_i$$

$$\boldsymbol{V}_i = \text{diag}[N_i \mu_i (1 - \mu_i)]$$

Teratology and Binomial Overdispersion

- Two common variance models:

1. **Scale model**:

$$\begin{aligned} \text{var}(Y_i) &= \phi N_i \mu_i (1-\mu_i) \\ B &= \phi \cdot A \end{aligned}$$

2. **beta-binomial**:

$$\begin{aligned} \text{var}(Y_i) &= N_i \mu_i (1-\mu_i)[1+\rho(N_i-1)] \\ B &= \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i^T \mathsf{V}_i^* \boldsymbol{X}_i \\ \boldsymbol{V}_i^* &= \text{diag}\{N_i \mu_i (1-\mu_i)[1+\rho(N_i-1)]\} \end{aligned}$$

Teratology and Binomial Overdispersion

Scale model

var$(\widehat{\boldsymbol{\beta}}) =$

beta-binomial model

var$(\widehat{\boldsymbol{\beta}}) =$

Teratology and Binomial Overdispersion

Unspecified variance model

$\text{var}(\widehat{\boldsymbol{\beta}}) =$

Teratology Analysis

```
# rats.q
#
# ----------------------------------------------------------------------
#
# PURPOSE:  Analysis of overdispersed binomial data
#
# AUTHOR:  P. Heagerty
#
# DATE:  00/01/12
#
# ----------------------------------------------------------------------
#
data<-matrix( scan("rat_teratology.data"), ncol=4, byrow=T )
#
rats <- data.frame(
                y=data[,2],
                n=data[,1],
                iron = data[,3]-8,
```

```
                iron0 = (data[,3]-8)*(data[,3]>8),
                id=c(1:nrow(data)) )
#
#####
##### EDA
#####
#
p <- rats$y/rats$n
logit.p <- log( rats$y+0.5 ) - log( rats$n - rats$y + 0.5 )
#
postscript( file="rats-EDA-1.ps", horiz=F )
par( mfrow=c(2,1) )
attach( rats )
plot( iron, p )
title("Observed proportion versus Iron")
plot( iron, logit.p )
title("Empirical logit versus Iron")
detach()
graphics.off()
#
logit<-function(x){ log(x)-log(1-x) }
```

```
antilogit<-function(x){ exp(x)/(exp(x)+1) }
#
#####  GLM fit
#
glmfit <-glm( cbind(y,n-y) ~ iron + iron0, family=binomial, data=rats )
#
resids <- residuals( glmfit, type="pearson" )
#
postscript( file="rats-EDA-2.ps", horiz=F )
par( mfrow=c(2,1) )
attach( rats )
plot( n, resids )
title("Pearson residuals versus litter size (n_i)")
plot( n, resids^2, ylim=c(0,20) )
title("Pearson residuals versus litter size (n_i)")
lines( smooth.spline( n, resids^2, df=4 ) )
text( 10, 15, "(3 points dropped)", cex=0.5 )
abline( h=1, lty=2 )
abline( h=mean( resids^2, trim=0.05 ), lty=3 )
detach()
graphics.off()
```

```
#
#####
##### Correction of standard errors
#####
#
attach( rats )
nobs <- nrow( rats )
#
##### (1) estimate of scale parameter
#
phi <- sum( resids^2 )/(nobs-3)
cat( paste("phi =", round(phi,3), "\n\n") )
#
##### (2) estimate of correlation parameter
#
rho <- mean( ((resids^2 - 1)/(rats$n-1))[rats$n>1] )
cat( paste("rho =", round(rho,3), "\n\n") )
detach()
#
##### Calculate standard errors...
#
```

```
X <- cbind( 1, rats$iron, rats$iron0 )
#
mu <- fitted( glmfit )
V.mu <- rats$n * mu * (1-mu)
#
est.fnx <- X*(rats$y - rats$n*mu)
#
##### Fisher Information
#
XtWX <- t(X)%*%( V.mu*X )
I0.inv <- solve( XtWX )
se.model <- sqrt( diag( I0.inv ) )
#
##### with Scale parameter
#
se.scale <- sqrt(phi)*se.model
#
##### with beta-binomial variance
#
cheese <- t(X)%*%( rats$n*mu*(1-mu)*(1+rho*(rats$n-1))*X )
se.sandwich <- sqrt( diag( I0.inv %*% ( cheese ) %*% I0.inv ) )
```

```
#
##### with empirical variance of estimating function
#
UUt <- t(est.fnx)%*%est.fnx
se.empirical <- sqrt( diag( IO.inv %*% ( UUt ) %*% IO.inv ) )
#
beta <- glmfit$coef
#
out <- cbind( beta, se.model, se.scale, se.sandwich, se.empirical )
print( round( out, 4 )  )
cat("\n\n -------------------------------------------------- \n\n")
#
summary( glmfit, cor=F )
cat("\n\n -------------------------------------------------- \n\n")
#
quasifit <- glm(  cbind(y,n-y) ~ iron + iron0,
                  family=quasi(link=logit, variance="mu(1-mu)" ),
                  data=rats )
summary( quasifit, cor=F )
cat("\n\n -------------------------------------------------- \n\n")
#
```

```
phi = 3.596

rho = 0.23

                  beta se.model se.scale se.sandwich se.empirical
(Intercept) -1.5196   0.2405   0.4561      0.4511       0.4388
       iron -0.7708   0.0830   0.1575      0.1532       0.1700
      iron0  0.5056   0.1517   0.2876      0.2842       0.2836
```

```
Call: glm(formula = cbind(y, n - y) ~ iron + iron0,
            family = binomial, data = rats)
Deviance Residuals:
       Min         1Q     Median        3Q      Max
 -4.932802 -1.130846 -0.1433674 1.431611 4.438619

Coefficients:
                 Value Std. Error   t value
(Intercept) -1.5196252  0.2404516 -6.319880
       iron -0.7707906  0.0830197 -9.284430
      iron0  0.5056392  0.1514921  3.337726

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 482.3326 on 54 degrees of freedom
Residual Deviance: 188.1113 on 52 degrees of freedom
Number of Fisher Scoring Iterations: 4
```

```
Call: glm(formula = cbind(y, n - y) ~ iron + iron0,
            family = quasi(link = logit, variance = "mu(1-mu)"),
            data = rats)
Deviance Residuals:
       Min        1Q     Median       3Q      Max
 -4.932802 -1.130846 -0.1433674 1.431611 4.438619

Coefficients:
                 Value Std. Error   t value
(Intercept) -1.5196252  0.4559841 -3.332627
       iron -0.7707906  0.1574357 -4.895907
      iron0  0.5056392  0.2872844  1.760065

(Dispersion Parameter for Quasi-likelihood family taken to be 3.596202 )

    Null Deviance: 482.3326 on 54 degrees of freedom
Residual Deviance: 188.1113 on 52 degrees of freedom
Number of Fisher Scoring Iterations: 4
```

Quasilikelihood

McCullagh and Nelder (1989) - Chapter 9

Model:

$$\begin{aligned} E(Y_i) &= \mu_i(\boldsymbol{\beta}) \\ \text{var}(Y_i) &= \sigma^2 V(\mu_i) \qquad \sigma^2 \text{ unknown} \end{aligned}$$

Quasilikelihood Function:

$$\begin{aligned} U(\boldsymbol{\mu}, \boldsymbol{y}) &= \sum_i \left[\sigma^2 V(\mu_i)\right]^{-1} (Y_i - \mu_i) \\ Q(\boldsymbol{\mu}, \boldsymbol{y}) &= \sum_i \int_{y_i}^{\mu_i} \left[\sigma^2 V(t)\right]^{-1} (Y_i - t)dt \end{aligned}$$

Quasilikelihood

Properties:

1. $E(U_i) = 0$.
2. $\text{var}(U_i) = \left[\sigma^2 V(\mu_i)\right]^{-1}$.
3. $-E(\partial U_i / \partial \mu_i) = \left[\sigma^2 V(\mu_i)\right]^{-1}$.

Quasilikelihood Estimating Equations

The quasilikelihood regression estimator, $\widehat{\boldsymbol{\beta}}$, for $Y_i,\ 1 = 1, 2, \ldots, n$ is obtained as the solution to the "quasi-score equations":

$$\mathbf{0} \;=\; \boldsymbol{U}(\boldsymbol{\beta}) \;=\; \boldsymbol{D}^T \boldsymbol{V}^{-1} (\boldsymbol{Y} - \boldsymbol{\mu})$$

$$D(i,j) = \frac{\partial \mu_i}{\partial \beta_j}$$

$$\boldsymbol{V} = \text{diag}(\ \sigma^2 V(\mu_i)\ )$$

Properties:

$$\mathcal{I}_n = \frac{1}{n}\boldsymbol{D}^T\boldsymbol{V}^{-1}\boldsymbol{D}$$

$$\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \rightarrow \mathcal{N}(\,\mathbf{0},\, \mathcal{I}_n^{-1}\,)$$

Note:

($\star$) These properties are based only on the correct specification of the **mean** and **variance** of Y_i.

Quasilikelihood Example – Teratology

```
Call: glm(formula = cbind(y, n - y) ~ iron + iron0,
                  family = quasi(link = logit,
                  variance = "mu(1-mu)"), data = rats)
Coefficients:
                 Value Std. Error   t value
(Intercept) -1.5196252  0.4559841 -3.332627
       iron -0.7707906  0.1574357 -4.895907
      iron0  0.5056392  0.2872844  1.760065

(Dispersion Parameter for Quasi-likelihood family taken
 to be 3.596202 )

    Null Deviance: 482.3326 on 54 degrees of freedom
Residual Deviance: 188.1113 on 52 degrees of freedom
```

Note:

Estimating Functions

Definition: A function of data and parameters, $g(Y, \theta)$, is an **estimating function**.

Definition: The function $g(Y, \theta)$ is **unbiased** if $\forall \theta \in \Theta$,

$$E_\theta[g(Y, \theta)] = 0 \quad .$$

Definition: For an unbiased estimating function, the **estimating equation**

$$g(Y, \hat{\theta}) = 0$$

defines an estimator, $\hat{\theta}$.

Estimating Functions

Estimating functions form the basis for almost all of frequentist statistical estimation.

Pearson (1894)

$$0 \;=\; \frac{1}{n}\sum_{i=1}^{n} Y_i^p \;-\; E[Y_i^p] \quad \text{for } p = 1, 2, \ldots, m$$

Fisher (1922)

$$0 \;=\; \sum_{i} \frac{\partial}{\partial\theta} \log f(Y_i;\theta)$$

- Method of Moments:

- Maximum Likelihood:

Estimating Functions - Optimality ($\star\star\star$)

Godambe (1960):

$$\mathcal{G} = \{ \, g \; : \; E_\theta[g(\boldsymbol{Y}, \theta)] = 0 \, \}$$

then, $g^* \in \mathcal{G}$, the score function, minimizes

$$E\left[\left(\frac{g(\boldsymbol{Y}, \theta)}{E[\partial g / \partial \theta]} \right)^2 \right]$$

Standardization: $g(\boldsymbol{Y}, \theta) / E[\partial g / \partial \theta]$

Why?

- (i) $g(\boldsymbol{Y}, \theta)$ and $c \times g(\boldsymbol{Y}, \theta)$ are equivalent.
- (ii) If $E_\theta[g(\boldsymbol{Y}, \theta)] = 0$ then we hope that
  $E_\theta[g(\boldsymbol{Y}, \theta + \delta)]$ is large.

Estimating Functions - Optimality ($\star\star\star$)

Godambe and Heyde (1987):

$$\mathcal{G}^{(1)} = \left\{ g \ : \ g(\boldsymbol{Y}, \theta) = \sum_i a_i(\theta)[Y_i - \mu_i(\theta)] \right\}$$

The class $\mathcal{G}^{(1)}$ is all *linear unbiased* estimating functions.

then, U, the quasi-score function,

$$U(\boldsymbol{\beta}) = \sum_i \left[\frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \right]^T V_i^{-1}(Y_i - \mu_i)$$

is optimal in the sense that:

1. Let $U^* = U/E[\partial U/\partial \beta]$ (standardized) and let g^* also represent

standardized g. Then,

$$E[(U^*)^2] \leq E[(g^*)^2]$$

2. Let s^* be the standardized score function. Then,

$$E[(U^* - s^*)^2] \leq E[(g^* - s^*)^2]$$

3. Equivalently,

$$\text{corr}(U^*, s^*) \geq \text{corr}(g^*, s^*)$$

EE Asymptotics ($\star\star\star$)

Fahrmeir and Kaufman (1985):

- In the classical linear model with iid errors.

$$\lambda_{\min} \sum_i \boldsymbol{X}_i^T \boldsymbol{X}_i \rightarrow \infty$$

where $\lambda_{\min}$ is the minimum eigenvalue, is necessary and sufficient for either weak or strong consistency of $\hat{\beta}$ the OLS estimator.

- Generalized Linear Model

(D) $\lambda_{\min} \mathcal{I}_n(\boldsymbol{\beta}) \rightarrow \infty$

(C) Bounded from below

$$F_n(\beta) - cF_n \qquad \text{is positive semidefinite}$$
$$\beta \in N(\delta), \qquad n_1 = n_1(\delta), n > n_1$$

(N) Convergence and continuity

$$\max_{\beta \in N(\delta)} \|V_n(\beta) - \mathcal{I}_n\| \quad \rightarrow \quad 0$$
$$V_n(\beta) \quad = \quad F_n^{-1/2} F_n(\beta) F_n^{-1/2}$$

Theorem 1: If (D) and (C) then the sequence of roots $\{\hat{\beta}_n\}$

(i) asymptotic existence

$$P[s_n(\hat{\beta}_n) = 0] \longrightarrow 1$$

(ii) weak consistency

$$\hat{\beta}_n \xrightarrow{P} \beta_0$$

Theorem 2: If (D) and (N) then

$$F_n^{-1/2} s_n \xrightarrow{d} N(0, I)$$

NOTE:

- This is for canonical link. If noncanonical then $H_n(\beta) \neq \mathcal{I}_n(\beta)$ and $\log L$ may not be convex.
- $F_n = F_n(\beta_0)$
- $N(\delta) = \left\{ \beta \ : \ \|F_n^{-1/2}(\beta - \beta_0)\| \leq \delta \ \right\}$

EE Asymptotics ($\star\star\star$)

Small and McLeish (1994): (Scalar θ)

$$
\begin{aligned}
U_n(\theta - \epsilon) &\rightarrow E_\theta[U_n(\theta - \epsilon)] > 0 \\
U_n(\theta + \epsilon) &\rightarrow E_\theta[U_n(\theta + \epsilon)] < 0
\end{aligned}
$$

Then the WLLN $\Longrightarrow$ exists a root in $(\theta - \epsilon, \theta + \epsilon)$.

General EE Consistency: (Crowder, 1986) Theorem 3.2

- $U_n(\theta) = \sum_i \frac{1}{n} U_i(\theta)$
- $\delta S(\theta_0, \epsilon)$ is boundary of a sphere

Then consistency results if:

(R1) $U_n(\theta)$ is continuous.

(R2) information condition

$$\inf_{\delta S(\theta,\epsilon)} (\theta_0 - \theta)^T E_{\theta_0}[U_n(\theta)]$$

(R3) uniform continuity

$$\sup_{\delta S(\theta_0,\epsilon)} \|U_n(\theta) - E_{\theta_0}[U_n(\theta)]\| \rightarrow 0$$

Q: (R2)?

$$
\begin{aligned}
(\theta_0 - \theta)^T E_{\theta_0}[U_n(\theta)] &\approx (\theta_0 - \theta)^T \left\{ E_{\theta_0}[U_n(\theta_0) + U_n'(\theta_0)(\theta - \theta_0)] \right\} \\
&= (\theta_0 - \theta)^T \left\{ -E_{\theta_0}[U_n'(\theta_0)] \right\} (\theta_0 - \theta)
\end{aligned}
$$

This will be $\geq \delta > 0$ if the "information" matrix is positive definite.

For EE's

$$
\frac{1}{n} \sum_i D_i^T V_i^{-1} D_i \rightarrow \mathcal{I}_\infty^{(M)}
$$

If $\mathcal{I}_\infty^{(M)}$ has its minimum eigenvalue > 0 then (R2) is satisfied.

Q: (R3)?

Crowder (1986) gives lemmas 2.2 and 3.2 that equate (R3) to conditions on $\frac{\partial}{\partial\theta}U_n(\theta)$:

$$P[\| \frac{\partial}{\partial\theta}U_n(\theta) - E_\theta U_n \| > G_\eta] < \eta$$

Again, for EEs this condition can be satisfied by conditions on information matrices (both model based and "true" information).

EEs and Asymptotic Normality

The derivation of the asymptotic distribution for the root of an estimating function is based on a simple Taylor's series expansion of the estimating function:

$$
\begin{aligned}
0 &= U_n(\hat{\theta}) \\
&\approx \\
&= \\
&=
\end{aligned}
$$

EEs and Empirical Variance Estimates

In the "sandwich variance" forms $A^{-1}BA^{-1}$ we find that the middle of the sandwich is simply

$$B = \frac{1}{n}\sum_{i=1}^{n} \text{var}\left[\boldsymbol{U}_i(\boldsymbol{\beta})\right]$$

We can either construct an explicit model that gives the form of B (see Exercises for the Poisson case), or we can use an **empirical variance estimator**:

$$\widehat{B} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}_i(\widehat{\boldsymbol{\beta}})\boldsymbol{U}_i(\widehat{\boldsymbol{\beta}})^T$$

- Under mild regularity conditions this leads to a consistent variance estimate for $\widehat{\boldsymbol{\beta}}$ without adopting an explicit variance model.

- In small or moderate samples we may choose to adopt a variance model.

- Huber (1967), White (1980, 1982), Royall (1986)

Summary

- $\exists$ lots of possible estimating functions to consider.

- Optimal estimating functions can be chosen from within certain classes of functions.

- Estimating functions provide semiparametric inference since means / covariances can be estimated without specification of a complete probability model.

- All (?) of frequentist estimation can be viewed via estimating functions.

Guyon, X. (1995)
Random Fields on a Network.

Springer-Verlag

Heyde, C.C (1997)
Quasi-likelihood And Its Application
Springer-Verlag

Small, C.G, and McLeish, D.L. (1994)
Hilbert Space Methods in Probability and Statistical Inference.
Wiley

Joint Modelling of Mean and Dispersion

McCullagh and Nelder (1989) - Chapter 10

Model:

$$
\begin{aligned}
E(Y_i) &= \mu_i(\boldsymbol{\beta}) \\
g(\mu_i) &= \boldsymbol{X}_i\boldsymbol{\beta}
\end{aligned}
$$

$$
\begin{aligned}
\text{var}(Y_i) &= \phi_i V(\mu_i) \quad (\star\star) \\
h(\phi_i) &= \boldsymbol{Z}_i\boldsymbol{\gamma}
\end{aligned}
$$

Estimation:

$$\boldsymbol{U}_1 = \sum_{i=1}^{n} \left[\frac{\partial \mu_i}{\partial \boldsymbol{\beta}}\right]^T [\phi_i V(\mu_i)]^{-1} (Y_i - \mu_i)$$

$$\boldsymbol{U}_2 = \sum_{i=1}^{n} \left[\frac{\partial \phi_i}{\partial \boldsymbol{\gamma}}\right]^T [\tau V_D(\phi_i)]^{-1} (d_i - \phi_i)$$

$$d_i = d(Y_i, \mu_i) \text{ such that } E(d_i) = \phi_i$$

$$\text{var}(d_i) = \tau V_D(\phi_i)$$

Example: `bod.regn()`

Parametric Models for Overdispersion

Crowder (1978) considered a hierarchical, or mixture model for overdispersed binomial data.

beta-binomial:

$$\begin{aligned} Y_i \mid p_i &\sim \text{binomial}(N_i, p_i) \\ p_i &\sim \text{beta}(\alpha, \beta) \end{aligned}$$

$$\begin{aligned} E(Y_i) &= N_i E(p_i) \; = \; N_i \mu \\ \text{var}(Y_i) &= N_i \mu (1-\mu) \left[1 + \rho (N_i - 1) \right] \end{aligned}$$

Beta-Binomial

$$P(Y_i = y) = \binom{N_i}{y} \frac{B(\alpha + y, N_i + \beta - y)}{B(\alpha, \beta)}$$

$$\mu =$$

$$\rho =$$

- Permits likelihood-based inference when data are overdispersed.

beta-binomial:

∘ Other mixture models are possible (ie. random intercepts GLMM), but the beta distribution is the conjugate distribution for the binomial.

∘ Permits regression estimation via a GLM for μ.

∘ The beta-binomial variance form leads to a different type of **weighting** when combining (Y_i, N_i) than that used by adopting a scale overdispersion model $(\phi N_i \mu(1-\mu))$.

- Regression inference, $g(\mu) = \boldsymbol{X}\boldsymbol{\beta}$, using the beta-binomial requires that the model is correct **conditional** on $\boldsymbol{X}_i$. (see Liang and Hanfelt 1994 for examples of trouble!)

Example: Teratology data and beta-binomial variance.

Parametric Models for Overdispersion

For count data the use of a mixture model to characterize extra-Poisson variation has also been suggested:

negative-binomial:

$$Y_i \mid \theta_i \sim \text{Poisson}(\ \theta_i\ )$$

$$\begin{aligned} \theta_i &= \exp(\ \boldsymbol{X}_i\boldsymbol{\beta} + \epsilon_i\ ) \\ \theta_i &= \exp(\ \boldsymbol{X}_i\boldsymbol{\beta}) \cdot \exp(\epsilon_i\ ) \\ &= \mu_i z_i \end{aligned}$$

$$\begin{aligned} z_i &\sim \Gamma(\text{shape} = \delta,\ \text{scale} = \gamma) \\ & \gamma = \delta \ \text{ so that } \ E(z_i) = 1 \end{aligned}$$

Negative Binomial

$$P(Y_i = y; \mu_i, \delta) = \frac{\Gamma(\delta + y)}{\Gamma(\delta)\Gamma(y+1)} \left(\frac{\delta}{\delta + \mu_i} \right)^{\delta} \left(\frac{\mu_i}{\delta + \mu_i} \right)^{y}$$

$$E(Y_i) = \mu_i$$

$$\mathsf{var}(Y_i) = \mu_i + \frac{1}{\delta}\mu_i^2$$

⋆ See pages 100-102 of Cameron and Trivedi for details.

Negative Binomial Models

- There are two common ways that the negative binomial model is parameterized in the regression context. Consider the distribution given on the previous page, with subscript i for both μ and δ:

$$P(Y_i = y; \mu_i, \delta_i) \quad = \quad \frac{\Gamma(\delta_i + y)}{\Gamma(\delta_i)\Gamma(y+1)} \left(\frac{\delta_i}{\delta_i + \mu_i} \right)^{\delta_i} \left(\frac{\mu_i}{\delta_i + \mu_i} \right)^{y}$$

Q: What assumptions are used to constrain the δ_i? Note that regression is used to structure μ_i.

Negative Binomial Models

NB-1 Model

$$E(Y_i \mid \boldsymbol{X}_i) \quad = \quad \mu_i$$

$$V(Y_i \mid \boldsymbol{X}_i) \quad = \quad \mu_i + \frac{1}{\delta_i}\mu_i^2 = \mu_i(1 + \frac{\mu_i}{\delta_i})$$

$$= \quad \mu_i \cdot \phi$$

assuming $\frac{\mu_i}{\delta_i} = \phi$

Negative Binomial Models

NB-2 Model

$$
\begin{aligned}
E(Y_i \mid \boldsymbol{X}_i) &= \mu_i \\
V(Y_i \mid \boldsymbol{X}_i) &= \mu_i + \frac{1}{\delta_i}\mu_i^2 = \mu_i(1 + \frac{\mu_i}{\delta_i}) \\
&= \mu_i(1 + \alpha \cdot \mu_i) \\
&\quad \text{assuming } \frac{1}{\delta_i} = \alpha
\end{aligned}
$$

Negative Binomial Models

General NB Model

$$
\begin{aligned}
E(Y_i \mid \boldsymbol{X}_i) &= \mu_i \\
V(Y_i \mid \boldsymbol{X}_i) &= \mu_i + \frac{1}{\delta_i}\mu_i^2 = \mu_i(1 + \frac{\mu_i}{\delta_i}) \\
&= \mu_i(1 + \alpha_i \cdot \mu_i) \\
h(\alpha_i) &= \boldsymbol{Z}_i\boldsymbol{\gamma}
\end{aligned}
$$

⋆ See pages 72-75 of Cameron and Trivedi for details.

Seizure Analysis using STATA

```
***********************************************************
*  seizure.do                                             *
***********************************************************
*                                                         *
* PURPOSE:  illustrate negative-binomial fitting          *
*                                                         *
* AUTHOR:  P. Heagerty                                    *
*                                                         *
* DATE:  15 Jan 2003                                      *
*                                                         *
***********************************************************

  *** READ DATA ***

infile id age base trt y1 y2 y3 y4 using "seizure.data"
```

```
  *** Create log(baseline+0.5) and log(age)

generate logB = ln( base + 0.5 )
generate logA = ln( age )




  *** POISSON REGRESSION ***

poisson y4 trt logA logB

poisgof

poisson y4 trt logA logB, robust
```

```
  *** Negative Binomial (2) ***

nbreg y4 trt logA logB, dispersion( mean )


  *** Negative Binomial (1) ***

nbreg y4 trt logA logB, dispersion( constant )


  *** Negative Binomial with dispersion model ***

gnbreg y4 trt logA logB, lnalpha( trt )

gnbreg y4 trt logA logB, lnalpha( trt logA )
```

```
COMMAND:  poisson y4 trt logA log

Poisson regression                                Number of obs   =         59
                                                  LR chi2(3)      =     331.93
                                                  Prob > chi2     =     0.0000
Log likelihood = -166.04363                       Pseudo R2       =     0.4999

------------------------------------------------------------------------------
          y4 |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         trt |  -.1439681   .1023428    -1.41   0.160    -.3445563    .0566202
        logA |   .3752063   .2339703     1.60   0.109    -.0833671    .8337797
        logB |   1.199049   .0692144    17.32   0.000     1.063391    1.334706
       _cons |  -3.382639   .9001401    -3.76   0.000    -5.146881   -1.618397
------------------------------------------------------------------------------
.
. poisgof

         Goodness-of-fit chi2  =  144.3189
         Prob > chi2(55)       =    0.0000
```

```
COMMAND:  poisson y4 trt logA log, robust

Poisson regression                              Number of obs   =         59
                                                Wald chi2(3)    =      78.86
                                                Prob > chi2     =     0.0000
Log likelihood = -166.04363                     Pseudo R2       =     0.4999

------------------------------------------------------------------------------
             |               Robust
          y4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         trt |  -.1439681   .1923317    -0.75   0.454    -.5209312     .232995
        logA |   .3752063    .329521     1.14   0.255     -.270643    1.021055
        logB |   1.199049   .1556253     7.70   0.000     .8940287    1.504069
       _cons |  -3.382639   1.273095    -2.66   0.008    -5.877859   -.8874198
------------------------------------------------------------------------------
```

```
COMMAND:  nbreg y4 trt logA log, dispersion( mean )

Negative binomial regression                      Number of obs   =         59
                                                  LR chi2(3)      =      60.41
                                                  Prob > chi2     =     0.0000
Log likelihood = -149.95938                       Pseudo R2       =     0.1677

------------------------------------------------------------------------------
          y4 |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         trt |  -.3062519   .1624457    -1.89   0.059     -.6246395    .0121357
        logA |   .2822898    .376851     0.75   0.454     -.4563246    1.020904
        logB |   1.091251   .1088676    10.02   0.000      .8778741    1.304627
       _cons |  -2.614126   1.396506    -1.87   0.061     -5.351227    .1229751
-------------+----------------------------------------------------------------
    /lnalpha |  -1.742645   .3823873                        -2.49211   -.9931791
-------------+----------------------------------------------------------------
       alpha |   .1750568   .0669395                        .0827352    .3703973
------------------------------------------------------------------------------
Likelihood ratio test of alpha=0:  chibar2(01) =   32.17 Prob>=chibar2 = 0.000
```

```
COMMAND:  nbreg y4 trt logA log, dispersion( constant )

Negative binomial (constant dispersion)           Number of obs   =         59
                                                  LR chi2(3)      =      54.41
                                                  Prob > chi2     =     0.0000
Log likelihood = -152.96274                       Pseudo R2       =     0.1510

------------------------------------------------------------------------------
          y4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         trt |  -.2040428   .1578151    -1.29   0.196    -.5133547    .1052692
        logA |   .3644169   .3464645     1.05   0.293     -.314641    1.043475
        logB |   1.128261   .1144604     9.86   0.000     .9039232      1.3526
       _cons |  -3.048889   1.356871    -2.25   0.025    -5.708307   -.3894705
-------------+----------------------------------------------------------------
    /lndelta |   .3793183   .3581039                     -.3225524    1.081189
-------------+----------------------------------------------------------------
       delta |   1.461288   .5232929                       .724298    2.948183
------------------------------------------------------------------------------
Likelihood ratio test of delta=0:  chibar2(01) =   26.16 Prob>=chibar2 = 0.000
```

```
COMMAND:  gnbreg y4 trt logA logB, lnalpha( trt )

Generalized negative binomial regression          Number of obs   =         59
                                                  LR chi2(3)      =      59.87
                                                  Prob > chi2     =     0.0000
Log likelihood = -148.69141                       Pseudo R2       =     0.1676

------------------------------------------------------------------------------
          y4 |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
y4           |
         trt |  -.3049821   .1657975    -1.84   0.066     -.6299393     .019975
        logA |   .2822315   .3717642     0.76   0.448     -.4464129    1.010876
        logB |   1.040298   .1076457     9.66   0.000      .8293165     1.25128
       _cons |  -2.447882    1.34919    -1.81   0.070     -5.092244    .1964814
-------------+----------------------------------------------------------------
lnalpha      |
         trt |   1.259242   .8122707     1.55   0.121     -.3327789    2.851264
       _cons |  -2.396513    .633884    -3.78   0.000     -3.638903   -1.154124
------------------------------------------------------------------------------
```

```
COMMAND:  gnbreg y4 trt logA logB, lnalpha( trt logA )

Generalized negative binomial regression          Number of obs   =         59
                                                  LR chi2(3)      =      59.95
                                                  Prob > chi2     =     0.0000
Log likelihood = -148.64905                       Pseudo R2       =     0.1678

------------------------------------------------------------------------------
          y4 |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
y4           |
         trt |  -.3093693   .1654708    -1.87   0.062     -.633686    .0149475
        logA |   .2548674   .3818872     0.67   0.505    -.4936178    1.003353
        logB |   1.041267   .1079694     9.64   0.000     .8296512    1.252883
       _cons |  -2.357008   1.376755    -1.71   0.087    -5.055398    .3413828
-------------+----------------------------------------------------------------
lnalpha      |
         trt |   1.184617   .8542226     1.39   0.166    -.4896282    2.858863
        logA |  -.5674455   1.983359    -0.29   0.775    -4.454757    3.319867
       _cons |  -.4834258   6.702122    -0.07   0.942    -13.61934    12.65249
------------------------------------------------------------------------------
```

NB Likelihood and Score Equations

- Note – these models assume specific variance forms with:
  - ▷ var$(y_i \mid x_i) = \mu_i + \alpha\mu_i^p$
  - ▷ **NB-1** assumes $p = 1$ and var$(y_i \mid x_i) = \mu_i \cdot (1 + \alpha)$
  - ▷ **NB-2** assumes $p = 2$ and var$(y_i \mid x_i) = \mu_i \cdot (1 + \alpha \cdot \mu_i)$
- Details on the likelihood, score equations, and model-based information is presented by Cameron and Trivedi pp. 71–75.
- The standard (common) hierarchical formulation assumes:
  - ▷ $[Y_i \mid z_i, \boldsymbol{X}_i] \sim \mathcal{P}\text{oisson}(\lambda_i \cdot z_i)$
  - ▷ $\log \lambda_i = \boldsymbol{X}_i\boldsymbol{\beta}$
  - ▷ $[z_i \mid \boldsymbol{X}_i]$ is scaled **gamma** with parameter δ_i

NB-2 likelihood

- The general form of the NB likelihood is given on p. 173 of lecture notes.
- This model assumes
  - ▷ $[z_i \mid \boldsymbol{X}_i] \sim \text{gamma}(\text{shape} = 1/\alpha, \text{scale} = 1/\alpha)$
- The resulting score equations define the MLEs $\widehat{\boldsymbol{\beta}}^{(2)}$ and $\widehat{\alpha}^{(2)}$ as the solution to:

$$\mathbf{0} = \sum_{i=1}^{n} (\boldsymbol{X}_i \mu_i)^T \left[\mu_i(1+\alpha\mu_i)\right]^{-1} (Y_i - \mu_i)$$

$$0 = \sum_{i=1}^{n} \left\{ \frac{1}{\alpha^2} \left(\log(1+\alpha\mu_i) - \sum_{j=0}^{Y_i-1} \frac{1}{(j+\alpha^{-1})} \right) + \frac{Y_i - \mu_i}{\alpha(1+\alpha\mu_i)} \right\}$$

NB-2 and maximum likelihood

- The information matrix takes a block-diagonal form:

$$\boldsymbol{A}^{(2)} = \begin{bmatrix} \boldsymbol{A}_\beta^{(2)} & 0 \\ 0 & A_\alpha^{(2)} \end{bmatrix}$$

- The model-based variance of $\widehat{\boldsymbol{\beta}}^{(2)}$ is therefore given by $\frac{1}{n}\left[\boldsymbol{A}_\beta^{(2)}\right]^{-1}$ where

$$\boldsymbol{A}_\beta^{(2)} = \frac{1}{n}\sum_i (\boldsymbol{X}_i\mu_i)^T \left[\mu_i(1+\alpha\mu_i)\right]^{-1} (\boldsymbol{X}_i\mu_i)$$

- The NB-2 model is a member of the **linear** exponential family (LEF) and therefore has certain robustness properties (e.g. what happens if distribution is not NB and/or variance is not correctly specified?)

NB-1 likelihood

- This model assumes
  - $[z_i \mid \boldsymbol{X}_i] \sim \text{gamma}(\text{shape} = \mu_i/\phi, \text{scale} = \mu_i/\phi)$
- The resulting score equations define the MLEs $\widehat{\boldsymbol{\beta}}^{(1)}$ and $\widehat{\alpha}^{(1)}$ as the solution to:

$$\mathbf{0} = \sum_{i=1}^{n} X_i^T \left\{ \left(\sum_{j=0}^{Y_i-1} \frac{(\phi-1)^{-1}\mu_i}{(j+(\phi-1)^{-1}\mu_i)} \right) + (\phi-1)^{-1}\mu_i \right\}$$

$$0 = \sum_{i=1}^{n} \frac{1}{(\phi-1)^2} \left\{ - \left(\sum_{j=0}^{Y_i-1} \frac{1}{[j+(\phi-1)^{-1}]} \right) \right.$$
$$\left. -(\phi-1)^{-2}\mu_i \log(\phi) - \frac{\phi-1}{\phi} + Y_i(\phi-1) \right\}$$

NB-1 and maximum likelihood

- The information matrix does not take a block-diagonal form.
- The NB-1 model is a **not** a member of the **linear** exponential family (LEF) and therefore **does not** have desirable robustness properties.
- Note: this model assumes the **heterogeneity** (extra-Poisson variability) does depend on $\boldsymbol{X}_i$ since the distribution of z_i depends on μ_i.

NB MLEs and Robustness?

- **Q**: what if the **mean** and **variance** models are correct but the **distribution** is incorrect (e.g. ZIP model is truth)?
- **Q**: what if the **mean** is correct but the **variance** model and the **distribution** is incorrect?
- **Answer**:
  - ▷ NB-2 MLE for β is consistent if mean model is correct. Model-based standard errors are not valid if either the variance model or the distributional assumption is violated.
  - ▷ NB-1 MLE for β is inconsistent unless data are negative binomial with assumed variance model.

NB MLE and Robustness?

- **NB-2 MLE**: $\widehat{\boldsymbol{\beta}}^{(2)}, \widehat{\alpha}^{(2)}$
  - ▷ Assume distribution is <u>not</u> NB but mean and variance are correctly specified.
  - ▷ $\widehat{\boldsymbol{\beta}}^{(2)} \longrightarrow \boldsymbol{\beta}_0$ (**justify?**)
  - ▷ $\widehat{\alpha}^{(2)} \longrightarrow \alpha^* \neq \alpha_0$
  - ▷ Model-based standard errors are given by $\boldsymbol{A}_{\beta}^{(2)}$ evaluated at the MLE $\widehat{\alpha}^{(2)}$ which is biased, and therefore the model-based standard errors are <u>not</u> consistent
- **NB-1 MLE**: $\widehat{\boldsymbol{\beta}}^{(1)}, \widehat{\alpha}^{(1)}$
  - ▷ **Trouble**.
  - ▷ $\widehat{\boldsymbol{\beta}}^{(1)} \longrightarrow \boldsymbol{\beta}^* \neq \boldsymbol{\beta}_0$ (justify?)
  - ▷ $\widehat{\phi}^{(1)} \longrightarrow \phi^* \neq \phi_0$

Summary of NB Maximum Likelihood Estimation

- Different parametrizations lead to different variance assumptions and different robustness properties.
- The Huber-White results allow us to study the properties of the MLE by considering the properties of the **score equations** – the defining estimating equations.
- We have seen that **sandwich variance** estimates can be used to correct standard errors for the Poisson regression MLE $\boldsymbol{\beta}^{(0)}$.
  - ▷ Obtain $\boldsymbol{A}$
  - ▷ Estimate $\boldsymbol{B}$
  - ▷ Report $\frac{1}{n}\boldsymbol{A}^{-1}\boldsymbol{B}\ \boldsymbol{A}^{-1}$

Summary of NB Maximum Likelihood Estimation

- We have used a couple of options regarding estimation of $\boldsymbol{B}$:
  - ▷ Assume a variance model and estimate additional variance parameters using method-of-moment estimators.
  - ▷ Use empirical variance of U_i, the EE contributions.
- **Q**: what about choosing an **efficient** estimator?
  - ▷ We have considered a **given** regression estimator such as the Poisson GLM MLE, $\widehat{\boldsymbol{\beta}}^{(0)}$, and used different methods to obtain a valid standard error – **unbiased inference**.
  - ▷ We may consider choice of regression estimator – such as using $\widehat{\boldsymbol{\beta}}^{(2)}$ – and the question of which estimator yields the most precise estimate of $\boldsymbol{\beta}$ arises – **efficient inference**.