STATISTICS 538, Lecture #6

Overdispersion

November 10, 2010

Overdispersion: common phenomenon in situations where the obvious first-choice glm is binomial or Poisson

Data may inherently be more variable than predicted by the model's mean-variance relationship.

Yields large deviance.

Standard errors too small if problem ignored?

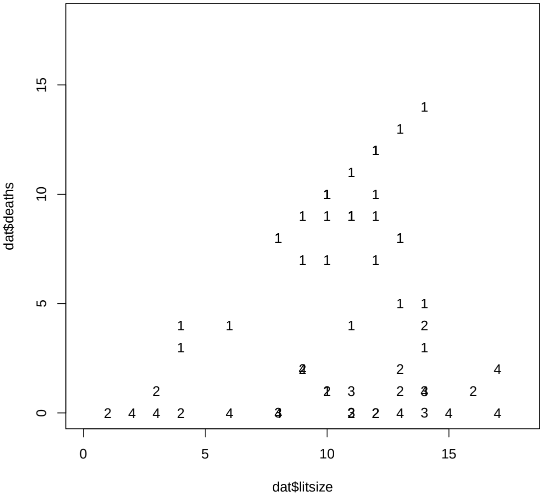
Example

```
### data from Agresti book, www.stat.ufl.edu
### 58 pregnant female rats on iron-deficient diets
### mortality per litter in offspring
### treatment groups 1:none
###                  2:iron supplement days 7 and 10
###                  3:iron supplement days 0 and 7
###                  4:iron supplement weekly

> dat <- read.table("rats.txt", header=F, row.names=1,
           col.names=c("", "trt", "litsize", "deaths") )

> dat$trt <- factor(dat$trt,
               labels=c("none", "7/10", "0/7", "wkly"))
```

plot(litsize, deaths, pch=as.character(as.numeric(trtmnt)))



Quasi-likelihood

Replace known dispersion ϕ with estimated $\hat{\phi} = (n - p)^{-1} D$.
(or a better estimator!)

Won't affect $\hat{\beta}$, but will boost standard errors by factor of $\sqrt{\hat{\phi}}$.

Recognition of additional uncertainty because of extra variability in data.

Back to rats data

```
dat$trt <- relevel(dat$trt, ref="7/10")
fit1 <- glm( cbind(deaths,litsize-deaths)~trt,
            family=binomial, data=dat)

            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.1785     0.3046  -7.153 8.51e-13 ***
trtnone       3.3225     0.3308  10.043  < 2e-16 ***
trt0/7       -1.1537     0.7814  -1.476    0.140
trtwkly      -0.8071     0.5503  -1.467    0.142

(Dispersion parameter for binomial family taken to be 1)
Residual deviance: 173.45  on 54  degrees of freedom
AIC: 252.92
```

Now with family=quasibinomial instead of family=binomial

```
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    -2.1785     0.5155  -4.226 9.23e-05 ***
trtnone         3.3225     0.5600   5.933 2.18e-07 ***
trt0/7         -1.1537     1.3227  -0.872    0.387    
trtwkly        -0.8071     0.9315  -0.867    0.390    

(Dispersion parameter for quasibinomial family taken
to be 2.864945)
Residual deviance: 173.45  on 54  degrees of freedom
AIC: NA
```

Comfortable doing inference without a real model?

Say Y is a random variable taking values in $\{0, \ldots, n\}$ with mean np and variance $\phi np(1-p)$. Can show $\phi < n$.

Note that in the Rats data ex., $\hat{\phi} = 2.86$, while $n_i = 1, 2$ for some litters.

What about using real models with $v(\mu) > \mu(1-\mu)$ for "binomial" data, or $v(\mu) > \mu$ for "Poisson" data?

Negative-Binomial as alternative to Poisson

This family of distributions can be parameterized as $E(Y) = \mu$, $Var(Y) = \mu + \mu^2/\theta$.

So have a GLM, but with an unknown parameter in the variance function (does complicate fitting algorithm).

Connection to usual parameterization, $Y \sim$ number of failures in sequence of independent trials performed until α successes are seen, where p is success probability for each trial?

More relevant representation of NB distribution as a **mixture**.

Simulate and fit some negative binomial data

```
set.seed(17)
n <- 80
x1 <- rnorm(n); x2 <- .8*x1 + sqrt(1-.8^2)*rnorm(n)
y <- rnbinom(n, mu=exp(1 + 0.3*x1), size=3)
dat2 <- data.frame(y=y, x1=x1, x2=x2)

fit3 <- glm(y ~ . , family=poisson, data=dat2)
fit4 <- glm(y ~ . , family=quasipoisson, data=dat2)
library("MASS")
fit5 <- glm.nb(y ~ x1+x2, data=dat2)
```

fit3

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.15947    0.06349  18.262   <2e-16 ***
x1           0.21035    0.09241   2.276   0.0228 *
x2          -0.01709    0.09490  -0.180   0.8571

(Dispersion parameter for poisson family taken to be 1)
Residual deviance: 167.92  on 77  degrees of freedom
AIC: 382.47
```

fit4

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.15947    0.09016  12.860   <2e-16 ***
x1           0.21035    0.13123   1.603    0.113
x2          -0.01709    0.13477  -0.127    0.899

(Dispersion parameter for quasipoisson family taken
to be 2.016572)
Residual deviance: 167.92  on 77  degrees of freedom
AIC: NA
```

fit5

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.15885    0.08983  12.900   <2e-16 ***
x1           0.21493    0.13282   1.618    0.106
x2          -0.01731    0.13635  -0.127    0.899

(Dispersion parameter for Negative Binomial(3.1055)
family taken to be 1)
Residual deviance: 92.740  on 77  degrees of freedom
AIC: 362.31

              Theta:  3.11
          Std. Err.:  1.06
```