Generalized Linear Models

Overdispersion

Demetris Athienitis

Model Fit

From the properties of the χ^2 distribution, we know that

$$E(\chi^2_\nu) = \nu$$

For a well fitting model we expect

$$X^2 \approx \text{Residual d.f.}$$

However, there are cases of concern where

$$X^2 \gg \text{Residual d.f.}$$

Could use G^2 (Residual Deviance) as an alternative, since $X^2 \approx G^2$, but not as efficient in detecting overdispersion.

Reasons

1. Badly fitting model
   - omitted terms/variables
   - incorrect relationship (link)
   - outliers
2. Variation greater than predicted by model that leads to *overdispersion*
   - count data: $V(Y) > \mu$
   - binomial data: $V(Y) > n\pi(1-\pi)$

Causes of Overdispersion

- variability of experimental material - individual level variability
- correlation between individual responses, e.g. litters of rats
- cluster sampling, e.g. areas; schools; classes; children
- aggregate level data
- omitted unobserved variables
- excess zero counts (structural and sampling zeros)

Consequences

With correct mean model we have consistent estimates of β but:

- incorrect standard errors
- selection of overly complex models

Remark

Overdispersion is much more common for count data, especially due to the restriction by the Poisson model $E(Y) = V(Y)$.

Checking overdispersion

- Check whether $X^2 \gg df$, or $\frac{X^2}{df} \gg 1$
- Fit a different model with additional parameters that allow variance to be greater and test the significance of those parameters
  - count data: Negative Binomial, parameter θ is introduced and estimated via MLE

$$V(Y) = \mu + \left(\frac{1}{\theta}\right)\mu^2$$

  - binomial data: Beta-Binomial, parameter ρ is introduced and estimated via MLE

$$V(Y) = n\pi(1-\pi)[1+(n-1)\rho]$$

In R

- Negative Binomial: `glm.nb{MASS}`
- Beta-Binomial: `betabinomial{VGAM}`

Example (Homicide)

1308 individuals who where classified as “Black” or “White” were asked: ”How many homicide victims have you personally known?”

	Number of victims						
Race	0	1	2	3	4	5	6
Black	119	16	12	7	3	2	0
White	1070	60	14	4	0	0	1

```
> head(homicide)  #data entered in ‘‘shorter’’ format
  nvics  race Freq
1     0 Black  119
2     1 Black   16
3     2 Black   12
4     3 Black    7
5     4 Black    3
6     5 Black    2
```

Example

```
> homicide=transform(homicide,race=relevel(race,"White"))
> hom.poi=glm(nvics~race,family=poisson(link="log"),
+ weights=Freq,data=homicide)
> summary(hom.poi)
.
    Null deviance: 962.80  on 10  degrees of freedom
Residual deviance: 844.71  on  9  degrees of freedom
```

Checking for overdispersion via $X^2/(df) \gg 1$ we first notice that the way the data was entered, the degrees of freedom is not 9 but actually 1308-2=1306

```
> sum(resid(hom.poi,type="pearson")^2)/
+  (sum(homicide$Freq)-length(hom.poi$coefficients))
[1] 1.745692
```

Some evidence of overdispersion is apparent.

Example

```
> library(MASS)
> hom.nb=glm.nb(nvics~race,weights=Freq,data=homicide)
> summary(hom.nb)
.
              Theta:  0.2023
          Std. Err.:  0.0409
```

$$\left(\frac{1}{\hat{\theta}}\right) \approx 5$$

seems substantial in. Much better now,

```
> sum(resid(hom.nb,type="pearson")^2)/
+  (sum(homicide$Freq)-length(hom.nb$coefficients))
[1] 1.090373
```

Example

More examples in notes.

Remark

The Beta-Binomial application is omitted here but an alternative method that does not use a likelihood approach but merely the structure between the mean and variance are the

- count data: Pseudo-Poisson
- binomial data: Pseudo-Binomial

but as result likelihood ratio tests are not possible.

We learned

- What is overdispersion
- Why it may occur
- How to identify it
- Possible model remedies