Generalized Linear Models

Roberta De Vito

BROWN
Public Health

A generalized linear model

1. $y = (y_1, \ldots, y_n)$
2. X and coefficients β
3. a link function g so that $\hat{y} = g^{-1}(X\beta)$
4. a data distribution
5. other parameters, for example?

Link function

What is the link function $g(u)$ for the linear regression?

a the identity: u^{-1}

b the logarithm: $log^{-1}(u)$

c the square: u^{-2}

Poisson regression model

- $i \rightarrow$ setting (location, time interval, index street)
- y_i the number of events (n. of traffic accident)
- $y_i \sim Poisson(\theta_i)$
- $\theta_i = exp(X_i\beta)$

Interpreting the coefficient

$$y_i \sim Poisson(exp(2.8 + 0.012\, X_{i1} - 0.20\, X_{i2})$$

$X_{i1} \rightarrow$ average speed (in miles per hour, mph) on the nearby street

$X_{i2} = 1$ if the intersection has a traffic signal, or 0 otherwise

Interpreting the coefficient

$$y_i \sim Poisson(exp(2.8 + 0.012\, X_{i1} - 0.20\, X_{i2})$$

$X_{i1} \rightarrow$ average speed (in miles per hour, mph) on the nearby street

$X_{i2} = 1$ if the intersection has a traffic signal, or 0 otherwise

- the intercept not interpretable
- $e^{0.012} = 1.012 \rightarrow 1.2\%$ in the rate of traffic accidents per mph
- $e^{-0.20} = 0.82 \rightarrow$ reduction of 18% with traffic signal

Data set: Police Stop

stop and frisk data (with noise added to protect confidentiality)

`ID`	code for the identification of the sample
`stops`	NYC police stops
`arrests`	number of arrests in the previous year
`precincts`	numbered 1-75
`ethnicity`	1=black, 2=hispanic, 3=white
`crime type`	1=violent, 2=weapons, 3=property, 4=drug

Question 2 and question 3 in prismia!

The Poisson model: Police Stop

R output

```
glm(formula = stops ~ factor(eth), family=poisson,
    offset=log(arrests))
```

Coef.	coef est	coef sd	p val
Intercept	-0.58	0.0038	$< 2e-16$
factor(eth)2	0.07	0.006	$< 2e-16$
factor(eth)3	-0.16	0.008	$< 2e-16$

Model	Deviance
Null	183981
Eth. covariate	183297

The ethnicity coefficient in the Poisson model

Compared to the baseline category 1 (blacks) how many stops category 2 (hispanics) has (coef: 0.07)?

1. 7% more
2. 70 % more
3. 7% less

The ethnicity coefficient in the Poisson model

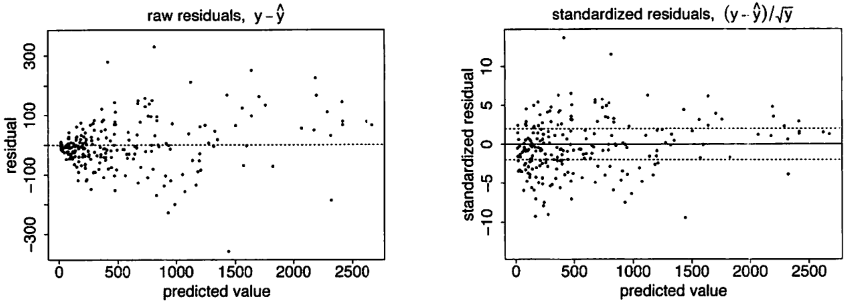
Compared to the baseline category 1 (blacks) how many stops category 3 (whites) has (coef: -0.16)?

1. 16 % less
2. 14% less
3. 85% more

Model with precincts

```
glm(formula = stops ~ factor(eth) + factor(precinct), family=poisson,
    offset=log(arrests))
                     coef.est coef.se
(Intercept)           -4.03    0.05
factor(eth)2           0.00    0.01
factor(eth)3          -0.42    0.01
factor(precinct)2     -0.06    0.07
factor(precinct)3      0.54    0.06
. . .
factor(precinct)75     1.41    0.08
  n = 225, k = 77
  residual deviance = 2828.6, null deviance = 44877 (difference = 42048.4)
  overdispersion parameter = 18.2
```

Testing for overdispersion in a Poisson regression model



The test

$$\text{Estimated overdispersion} = \frac{1}{n-k}\sum_{i=1}^{n} z_i^2$$

This is a χ^2_{n-k}

```
yhat <- predict (glm.police, type="response")
z <- (stops-yhat)/sqrt(yhat)
cat ("overdispersion ratio is ", sum(z^2)/(n-k), "\n")
cat ("p-value of overdispersion test is ", pchisq (sum(z^2), n-k), "\n")
```

The estimated overdispersion factor is 2700/148=18.2, and the p-value is 1.

Adjusting Inference for overdispersion

- multiply the standard errors by $\sqrt{18.2}$
- example whites

$$\text{before:} \quad -0.42 \pm 0.01 \qquad \text{now:} \quad -0.42 \pm 0.04$$

- confidence intervals

$$e^{-0.42 \pm 1.96 \times 0.04} = [0.61, 0.71]$$

Binary Data as a special case of count-data model

Logit or Probit?

$$\Pr(y_i = 1) = \Phi(X_i\beta)$$

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i < 0 \end{cases} \\ z_i &= X_i\beta + \epsilon_i \\ \epsilon_i &\sim \mathrm{N}(0,1), \end{aligned}$$

Probit or logit?

```
> fit.probit <- glm (switch ~ dist100, family=binomial(link="probit"))
> summary(fit.probit)

Call:
glm(formula = switch ~ dist100, family = binomial(link = "probit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4409  -1.3055   0.9669   1.0312   1.6674

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.37781    0.03730   10.13  < 2e-16 ***
dist100     -0.38741    0.06034   -6.42 1.36e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4118.1  on 3019  degrees of freedom
Residual deviance: 4076.3  on 3018  degrees of freedom
AIC: 4080.3

Number of Fisher Scoring iterations: 4
```

Probit or logit?

```
> fit.logit <- glm (switch ~ dist100, family=binomial(link="logit"))
> summary(fit.logit)

Call:
glm(formula = switch ~ dist100, family = binomial(link = "logit"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4406  -1.3058   0.9669   1.0308   1.6603

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.60596    0.06031  10.047  < 2e-16 ***
dist100     -0.62188    0.09743  -6.383 1.74e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4118.1  on 3019  degrees of freedom
Residual deviance: 4076.2  on 3018  degrees of freedom
AIC: 4080.2

Number of Fisher Scoring iterations: 4
```

Logit Distribution