36-463/663: Multilevel & Hierarchical Models

Generalized Linear Models

Brian Junker
132E Baker Hall
brian@stat.cmu.edu

Outline

- Linear Regression, Logistic Regression
- Generalized Linear Models (GLM)
- Example: Poisson Regression
  - Exposure and Offsets
  - Overdispersion
  - Zero-inflation
- I've been slow to get hw solutions out but it should be better now
  - HW01 and HW02 solutions are online now, and HW03 solutions will be out soon
- HW04 (due next week) is posted online

Linear Regression, Logistic Regression

- The ***linear regression*** model is:

$$
\begin{aligned}
y_i &\overset{indep}{\sim} N(\theta_i, \sigma^2),\ i = 1, \ldots, n \\
\theta_i &= X_i\beta = \beta_1 X_{i1} + \cdots \beta_k X_{ik}
\end{aligned}
$$

  - Each $y_i \in (-\infty, \infty)$ has some mean $\theta_i = E[y_i]$
  - Each θ_i has some linear structure
  - There is a statistical distribution N(*, σ^2) that describes unmodeled variation around $\theta_i = E[y_i]$
- The ***logistic regression*** model is:

$$
\begin{aligned}
y_i &\overset{indep}{\sim} Bernoulli(p_i),\ i = 1, \ldots, n \\
\theta_i &= \log \frac{p_i}{1-p_i} = X_i\beta = \beta_1 X_{i1} + \cdots \beta_k X_{ik}
\end{aligned}
$$

  - Each $y \in \{0, 1\}$ has some mean $p_i = E[y_i]$
  - Each $\theta_i = g(p_i)$ has some linear structure [g(p) = log p/(1-p) !]
  - There is a statistical distribution Bernoulli(*) that describes unmodeled variation around $p_i = E[y_i]$

Generalized Linear Models

- The ***generalized linear model (glm)*** is:

$$
\begin{aligned}
y_i &\overset{indep}{\sim} f(y_i|\mu_i, \ldots),\ i = 1, \ldots, n \\
\theta_i &= g(\mu_i) = X_i\beta = \beta_1 X_{i1} + \cdots \beta_k X_{ik}
\end{aligned}
$$

  - Each y_i has some mean $\mu_i = E[y_i]$
  - Each $\theta_i = g(\mu_i)$ has some linear structure [$g(\mu)$ is the "link function"]
  - There is a statistical distribution $f(y_i|\mu_i, \ldots)$ that describes unmodeled variation around $\mu_i = E[y_i]$
  - *There may be other parameters "…" in* $f(y_i|\mu_i, \ldots)$ *but the "main" parameter is* $\mu_i = g^{-1}(\theta_i) = g^{-1}(X_i\beta)$
- For *ordinary linear regression*
  - $f(y_i|\mu_i, \ldots) = N(\mu_i, \sigma^2)$ [$\mu_i = E[y_i]$]
  - $g(\mu) = \mu$ [the "identity link function"]
- For *logistic regression*
  - $f(y_i|p_i) = \text{Bernoulli}(p_i)$ [$p_i = E[y_i]$]
  - g(p) = log p/(1-p) [the "logit link function"]

Some Other GLM's

- Poisson Regression Model
  - $y_i \in \{0, 1, 2, 3, \ldots\}$
  - $f(y_i | \lambda_i) = \text{Poiss}(\lambda_i)$ $[\lambda_i = E[y_i]]$
  - $\theta_i = \log(\lambda_i) = X_i \beta$

    (confusingly: G&H use θ where I use λ... sorry!)
- Logistic-Binomial Model (aka logistic regression)
  - $y_i \in \{0, 1, \ldots, n_i\}$
  - $f(y_i | p_i, n_i) = \text{Binomial}(n_i, p_i)$
  - $\theta_i = \log p_i/(1-p_i) = X_i\beta$

A Few More GLM's

- Probit Regression Model
  - $y_i \in \{0, 1\}$
  - $f(y_i | p_i) = \text{Bernoulli}(p_i)$
  - $\theta_i = \Phi^{-1}(p_i) = X_i\beta$
- Ordered Multinomial Logit Model
  - $y_i \in \{1, 2, \ldots, K\}$
  - $f(y_i | p_{i1}, \ldots, p_{iK})$: $P[y_i>k] = p_{ik}$ $\quad k = 1, \ldots, K-1$
  - $\theta_i = \log p_{ik}/(1-p_{ik}) = X_i\beta - c_k$ $\quad k = 1, \ldots, K-1$
  - This is one kind of "***multinomial regression***" model

    -- there are many others!

Poisson Regression Example

- Poisson Regression Model
  - $y_i \epsilon \{0, 1, 2, 3, \ldots\}$
  - $f(y_i | \lambda_i) = \text{Poiss}(\lambda_i)$ $[\lambda_i = E[y_i]]$
  - $\theta_i = \log(\lambda_i) = X_i \beta$
- We will fit this model to data, and then look at some modifications of the model involving
  - offsets
  - overdispersion
  - zero-inflation

  (the same kinds of modifications can be helpful with logistic regression and other GLM's…)

Poisson Regression – The Data

- Data from an experiment on the effectiveness of an "integrated pest management system" in apartment buildings in a particular city (from G&H Ch 8).

```
roachdata <- read.csv ("roachdata.csv")

str(roachdata)

 'data.frame':   262 obs. of  6 variables:
  $ X        : int  1 2 3 4 5 6 7 8       [observation number]
  $ y        : int  153 127 7 7 0 0       [# of roaches trapped after expmt]
  $ roach1   : num  308 331.25 1.67       [# of roaches before experiment]
  $ treatment: int  1 1 1 1 1 1 1 1       [pest mgmt tx in this apt bldg?]
  $ senior   : int  0 0 0 0 0 0 0 0       [apts restricted to sr citzns?]
  $ exposure2: num  0.8 0.6 1 1 1.14      [avg # of trap-days per apt for y]
```

Poisson Regression – Fitting the Model

```
> glm.0 <- glm (y ~ roach1 + treatment + senior,
  family=poisson)
> summary(glm.0)

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.136e+00  2.124e-02  147.64   <2e-16 ***
roach1        6.444e-03  8.832e-05   72.97   <2e-16 ***
treatment    -5.124e-01  2.465e-02  -20.79   <2e-16 ***
senior       -3.760e-01  3.355e-02  -11.21   <2e-16 ***
```

$$\begin{aligned}
\lambda_i &= E[Y_i] \\
\log \lambda_i &= 3.14 + 0.00064(roach1) - 0.5(treatment) - 0.38(senior) \\
\lambda_i &= \exp(3.14 + 0.00064(roach1) - 0.5(treatment) - 0.38(senior)) \\
&= \exp(3.14)\exp(0.00064(roach1))\exp(-0.5(treatment))\exp(-0.38(senior))
\end{aligned}$$

Poisson Regression – Interpreting the Coefficients

- ***Intercept = 3.14***: exp(3.14) = 23.10 is the average # of roaches trapped *after* the experiment, in an apt bldg with no roaches *before* the experiment (roach1=0), no treatment (treatment=0) and not a seniors' building (senior=0).
  - In this case there are about 60 buildings with no roaches at the start of the experiment, so this is probably a meaningful description
- ***roach1 = 0.00644***: exp(0.00644) = 1.006 is the factor increase in average roaches caught *after* the experiment, per roach caught *before* the experiment (does this make sense?).
- ***treatment = -0.512***: exp(-0.512) = 0.60 is the factor reduction in average roaches caught *after* the experiment, due to treatment
- ***senior = -0.38***: exp(-0.38) = 0.68 is the factor reduction in the average roaches caught *after* the experiment, due to being a senior bldg

Poisson Regression - Exposure

- We have not made use of exposure2 = average number of trap-days
  - If twice as many traps, expect to catch 2x roaches
  - If 3 times as many days, expect to catch 3x roaches
- To accommodate this multiplicative effect, we can try

$$\lambda_i = u_i e^{X_i \beta}$$

where u_i = exposure2.

Poisson Regression – Exposure

- Taking logs, the "linear regression" form is

$$\log(\lambda_i) = \log(u_i) + X_i \beta$$

This is like including $\log(u_i)$ in the model, and basically forcing its coefficient to be exactly 1.

- In R we accomplish this with the "offset" argument
- This makes interpretation of the coefficients easier
  - coefficients measure deviations from expected counts under the various numbers of trap-days
  - This "unconfounds" exposure from treatment, bldg type, etc.

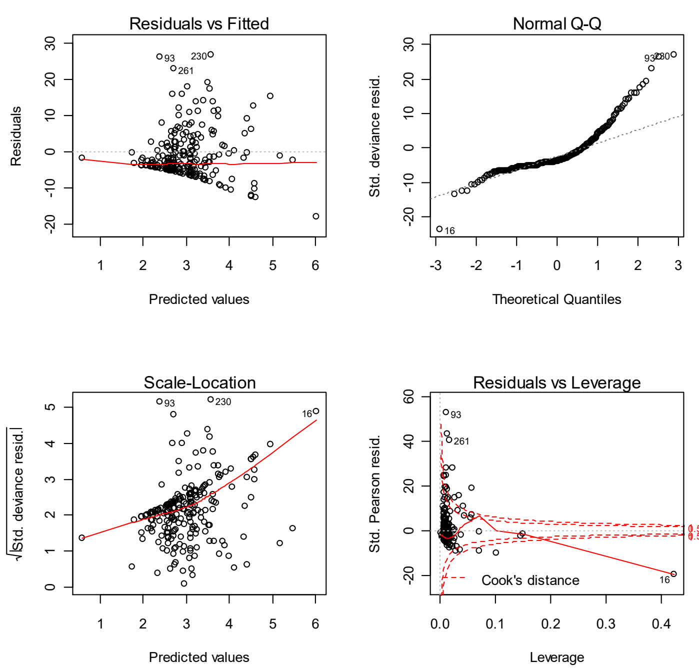
Poisson Regression – Exposure and Offsets

```
> glm.1 <- glm (y ~ roach1 + treatment +
   senior, family=poisson,
   offset=log(exposure2))
> round(cbind(glm.0=coef(glm.0),
   glm.1=coef(glm.1)),4)
```

	glm0	glm1
(Intercept)	3.1360	3.0892
roach1	0.0064	0.0070
treatment	-0.5124	-0.5167
senior	-0.3760	-0.3799

Poisson Regression – Looking at Residuals

par(mfrow=c(2,2))
plot(glm.1)

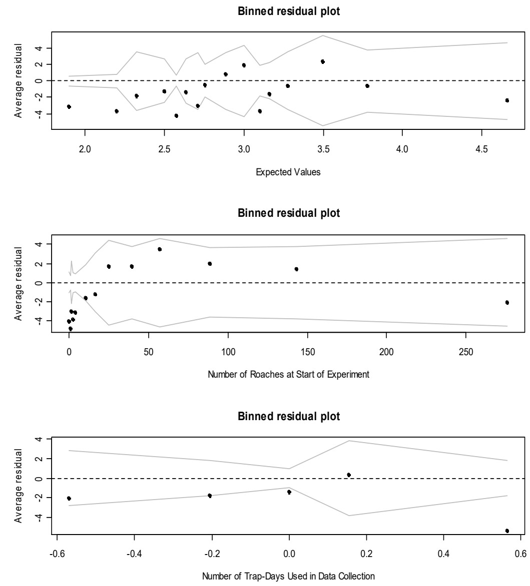


Poisson Regression – Looking at Residuals

```
par(mfrow=c(3,1))
xvar <- predict(glm.1)
yvar <- resid(glm.1)
binnedplot(xvar,yvar)

xvar <- roach1
binnedplot(xvar,yvar,xlab
   ="Number of Roaches at
   Start of Experiment")

xvar <- log(exposure2)
binnedplot(xvar,yvar,xlab
   ="Number of Trap-Days
   Used in Data
   Collection")
```



Poisson Regression – Testing Lack of Fit

- If $y_i \sim$ Poisson(λ_i) then the standardized residual

$$z_i = \frac{y_i - \lambda_i}{\sqrt{\lambda_i}}$$

is approximately normal, so that

$$\sum_{i=1}^{n} z_i^2$$

should follow a χ^2 distribution on n – k df

  - n = sample size, k = number of betas in the model

Poisson Regression – Testing Lack of Fit

```
> E.y. <-
    predict(glm.1,type="response")
> z <- (y - E.y.)/sqrt(E.y.)

> test.statistic <- sum(z^2)

> n <- length(y)
> k <- length(coef(glm.1))

> pchisq(test.statistic,n-
    k,lower.tail=F)

 [1] 0
```

```
> test.statistic

 [1] 16883.04      # this is *huge*!

> n-k

 [1] 258

>  test.statistic/(n-k)

 [1] 65.43815
```

We found that the residuals are extremely *overdispersed*: the variability of the z's is about 65 times what it should be!

Poisson Regression - Overdispersion

- We can adjust our inferences for overdispersion by adjusting the standard errors of the coefficients:

```
round(coef(summary(glm.1))[,1:2],2)
#              Estimate Std. Error
# (Intercept)      3.09       0.02
# roach1           0.01       0.00
# treatment       -0.52       0.02
# senior          -0.38       0.03

round(coef(summary(glm.1))[,1:2] %*%
  diag(c(1,sqrt(test.statistic/(n-k)))),2)
#               [,1] [,2]
# (Intercept)   3.09 0.17
# roach1        0.01 0.00
# treatment    -0.52 0.20
# senior       -0.38 0.27
```

After adjusting, every-thing remains signifi-cant, except for "senior" housing status.

Poisson Regression - Overdispersion

- We can also get R to estimate the overdispersed poisson regression model directly.

```
> glm.2 <- glm (y ~ roach1 + treatment + senior, family=quasipoisson,
  offset=log(exposure2))
> summary(glm.2)

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.0892463  0.1717721  17.985   <2e-16 ***
roach1       0.0069829  0.0007179   9.727   <2e-16 ***
treatment   -0.5167262  0.2001254  -2.582   0.0104 *
senior      -0.3798751  0.2703380  -1.405   0.1612
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 65.4403)
```


Poisson Regression – Zero Inflation

- If we explore the data a little more we find that there may be too many zeros for the Poisson model to fit well:

```
> stem(y)

  The decimal point is 1 digit(s) to the right of the |

   0 | 0000000000000000000000000000000000000000000000000000000000000+111
   2 | 0112445667789234778888
   4 | 02458889000399
   6 | 0234993777
   8 | 081
  10 | 249
  12 | 0127756
  14 | 8039
  16 | 189
  18 | 33
  20 | 3
  22 |
  24 | 3
  26 |
  28 | 3
  30 |
  32 |
  34 | 7
```

```
> length(y)
[1] 262
> (lambda <- mean(y))
[1] 25.65
> tbl <- NULL
> for (k in 0:5) {
+ tbl <- rbind(tbl,c(obs=sum(y==k),
+ exp=262*exp(-25.65)* 25.65^k /
+ factorial(k)))
+ }
```

```
> tbl
     obs          exp
[1,]  94 1.899537e-09
[2,]  20 4.872313e-08
[3,]  11 6.248742e-07
[4,]  10 5.342674e-06
[5,]   7 3.425990e-05
[6,]   7 1.757533e-04
```

Poisson Regression – Zero Inflation

- In cases like this it can also be useful to separately model
  - What distinguishes zero-cockroach buildings from others; and
  - what drives cockroach counts in the buildings that have them
- We combine a logistic regression analysis and a Poisson regression analysis to try to answer these questions

Poisson Regression – Zero Inflation

```
> some.cockroaches <-
   ifelse(y>0, 1, 0)

> zero.fit <-
   glm(some.cockroaches ~ roach1
   + treatment + senior +
   exposure2, family=binomial)

> display(zero.fit)

glm(formula = some.cockroaches ~
   roach1 treatment + senior +
   exposure2, family = binomial)
             coef.est coef.se
(Intercept)  0.85     0.57
roach1       0.03     0.01
treatment   -0.64     0.30
senior      -0.86     0.31
exposure2   -0.20     0.48
---
  n = 262, k = 5
  residual deviance = 281.7,
   null deviance = 342.0
   (difference = 60.3)
```

```
> glm.3 <- glm (y ~ roach1 + treatment
   + senior, family=quasipoisson,
  offset=log(exposure2),subset = (y>0))

> display(glm.3)

glm(formula = y ~ roach1 + treatment +
    senior, family = quasipoisson,
     subset = (y > 0), offset =
    log(exposure2))
            coef.est coef.se
(Intercept)  3.49     0.16
roach1       0.01     0.00
treatment   -0.47     0.19
senior      -0.22     0.26
---
  n = 168, k = 4
  residual deviance = 7764.6, null
   deviance = 10979.5 (difference =
   3214.9)
  overdispersion parameter = 61.2
```

Everything is a significant predictor, except for # of trap-days

Poisson Regression – Zero Inflation

- *A building with no roaches at the start of the experiment (roach1=0) in the treatment group (treatment=1) that is a seniors' building (senior=1) with 1.5 trap-days (exposure2=1.5) has probability*

  invlogit(0.85 + (0)*(0.03) + (-0.64)*(1) + (-0.86)*(1) + (1.5)*(-0.20)) = 0.28

  of having roaches at the end of the experiment
- *Given that the building does have roaches at the end, the expected number of roaches is*

  exp(log(1.5) + 3.48 + (0)*(0.0056) + (1)*(-0.47) + (1)*(-0.22)) = 25

Summary

- Linear Regression, Logistic Regression
- Generalized Linear Models (GLM)
- Example: Poisson Regression
  - Exposure and Offsets
  - Overdispersion
  - Zero-inflation
- HW04 is posted online