36-617: Applied Linear Models

Multi-level glm's

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

Announcements

- Project 02 Schedule:
  - **Mon Nov 29 (or earlier):** Full IDMRAD paper first draft.
  - **Fri Dec 3:** Peer reviews due.
  - **Fri Dec 10 (or earlier):** Full IDMRAD paper final draft!
- Regular classes Nov 22 (today!), 29 & Dec 1
- No more graded hw's or quizzes
  - I may give one more ***ungraded*** hw, if I have time to set it up
  - End of semester feedback for me Nov 29
- *Reading for after Thanksgiving (no quiz…)*
  - *Sheather, Appendix: Nonparametric Smoothing*
  - *Additional Reading (I will place in Canvas under weeks 13 & 14)*

Outline

- Review glm's, e.g.
  - Logistic Regression
  - Poisson Regression
- Clustering, growth curves, overdispersion
- Multi-level glm's
  - A.k.a. generalized linear mixed effects regression models (glmer!)
- Example(s)
  - Clustering/growth curves: Hospital birth choices
  - Overdispersion: Roach data redux

Linear Regression, Logistic Regression

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

$$\begin{aligned} y_i &\overset{indep}{\sim} N(\theta_i, \sigma^2),\ i = 1, \ldots, n \\ \theta_i &= X_i\beta = \beta_0 X_{i0} + \cdots \beta_p X_{ip} \end{aligned}$$

  - 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 $N(*, \sigma^2)$ that describes unmodeled variation around $\theta_i = E[y_i]$

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

$$\begin{aligned} y_i &\overset{indep}{\sim} f(y_i|\mu_i, \ldots),\ i = 1, \ldots, n \\ \theta_i &= g(\mu_i) = X_i\beta = \beta_0 X_{i0} + \cdots \beta_k X_{ip} \end{aligned}$$

  - 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 $\mu_i = E[y_i]$

Logistic regression, Poisson regression

- The ***logistic regression*** model is:

$$y_i \overset{indep}{\sim} Binomial(n_i, p_i),\ i = 1, \ldots, n$$

$$\theta_i = \log \frac{p_i}{1 - p_i} = X_i\beta = \beta_0 X_{i0} + \cdots \beta_p X_{ip}$$

  - Each y ϵ {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 $f(y_i | p_i)$ = Binomial(n_i, p_i) that describes unmodeled variation around $p_i = E[y_i]$

- The ***Poisson Regression*** model is:

$$y_i \overset{indep}{\sim} Poisson(\lambda_i),\ i = 1, \ldots, n$$

$$\theta_i = \log \lambda_i = X_i\beta = \beta_0 X_{i0} + \cdots \beta_p X_{ip}$$

  - Each $y_i \epsilon$ {0, 1, 2, 3, …} has some mean $\lambda_i = E[y_i]$
  - Each $\theta_i = g(\lambda_i)$ has some linear structure [$g(\lambda_i) = \log(\lambda_i)$!]
  - There is a statistical distribution $f(y_i | \lambda_i)$ = Poiss(λ_i) that describes unmodeled variation around $\lambda_i = E[y_i]$

Clustering, growth curves, overdispersion

- Just as with linear models, glm data can involve
  - ***Clustering***: groups of observations more similar to each other within group than between groups
  - ***Growth curves***: the clusters are individuals, and the observations are measurements at successive time points
- And with glm's we also sometimes see
  - ***Overdispersion***: Although the variance should be a function of the mean ($Var_{Poiss}(y) = \lambda$; $Var_{Bern}(y)=p(1-p)$), when it is not, we need a way to model it

Multi-level glm's

- Level 1 (a glm, modeling the data itself):

$$\begin{aligned} y_i &\overset{indep}{\sim} f(y_i|\mu_i, \ldots), \; i = 1, \ldots, n \\ \theta_i &= g(\mu_i) = X_i\alpha = \alpha_{0j[i]}X_{i0} + \cdots \alpha_{pj[i]}X_{ip} \end{aligned}$$

- Level 2 (modeling level 1 coefficients):

$$\begin{aligned} \alpha_{0j} &= \beta_{00} + \beta_{01}W_{j1} + \cdots + \beta_{0q}W_{jq} + \eta_0 \;, \quad \eta_0 \sim N(0, \tau_0^2) \\ \alpha_{1j} &= \beta_{10} + \beta_{11}W_{j1} + \cdots + \beta_{1q}W_{jq} + \eta_1 \;, \quad \eta_1 \sim N(0, \tau_1^2) \\ &\vdots \qquad\qquad \vdots \\ \alpha_{pj} &= \beta_{p0} + \beta_{p1}W_{j1} + \cdots + \beta_{pq}W_{jq} + \eta_p \;, \quad \eta_p \sim N(0, \tau_p^2) \end{aligned}$$

- Can fit with glmer() from the lme4() R package…

Example 1: Deliver babies in a hospital or at home?

- `hosp.txt` contains data from Lillard & Panis (2000)'s study of the decisions of 501 mothers to give birth in a hospital or elsewhere, for 1060 births:

```
'data.frame': 1060 obs. of 6 variables:
$ hospital: int 0 0 1 0...   1 = hospital birth, 0 = elsewhere
$ loginc : num 4.33 5.62...  Log_e of family income (log dollars)
$ distance: num 1.7 7.9...   distance (miles) to nearest hospital
$ dropout : int 0 0 0 0 0... 0 = mom completed hs , 1 = did not
$ college : int 1 0 0 0 0... 1 = mom attended coll, 0 = did not
$ mom : int 1 2 2 2 2...     unique identifier for each mother
```


Lillard, L. A., & Panis, C. W. (2000). aML multilevel multiprocess statistical software, release 1.0. (current version: http://www.applied-ml.com/) *Los Angeles: EconWare*.

Example 1: Hospital Birth Choices

- See R handout/demonstration
  - hosp-births.r

Example 2: Cockroach Eradication

- `roachdata.csv` contains data from an experiment on the effectiveness of an "integrated pest management system" in apartment buildings in a particular city (from G&H).

```
#  $ 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]
```

Example 2: Cockroach Eradication

- See R handout/demonstration
  - roachdata.r

Summary

- Review glm's, e.g.
  - Logistic Regression
  - Poisson Regression
- Clustering, growth curves, overdispersion
- Multi-level glm's
  - A.k.a. generalized linear mixed effects regression models (glmer!)
- Example(s)
  - Clustering/growth curves: Hospital birth choices
  - Overdispersion: Roach data redux