modeling injury data

the basics

Charles DiMaggio, PhD, MPH, PA-C

Professor of Surgery and and Population Health
New York University School of Medicine
Bellevue Hospital
Division of Trauma and Surgical Critical Care
462 First Avenue, New York, NY 10016

Spring 2017

- http://www.injuryepi.org/
- Charles.DiMaggio@nyumc.org

Outline

1. state mvc data
2. visualizing and plotting the data
3. introduction to poisson models
4. poisson models of state mvc data

traffic fatality data

loading and exploring the data

```
install.packages("AER") #applied econometrics in R
library(AER)

data(Fatalities)
?Fatalities
str(Fatalities)

#calculate incidence per state per year, plot as time series
table.deaths<-tapply(Fatalities$fatal, list(Fatalities$state,
Fatalities$year), sum)
table.exp<-tapply(Fatalities$milestot, list(Fatalities$state,
Fatalities$year), sum)

inc.dense<-table.deaths/table.exp*100
inc.dense
```

visualizing the data

```
plot.ts(t(inc.dense), plot.type="single") #need to transpose

inc.dense.df<-as.data.frame(inc.dense)

str(inc.dense.df)
mean(inc.dense.df$`1982`)
colMeans(inc.dense.df)

plot.ts(colMeans(inc.dense.df))

plot.ts(colMeans(inc.dense.df), ylim=c(0,3))
```

ggplot

```
# plot ggplot with loess line
library(ggplot2)
nat.rates<-data.frame(Rate=colMeans(inc.dense.df), Year=1982:1988)
p1<-ggplot(data=nat.rates, aes(x=Year, y=Rate))
p1+geom_line(aes(group=1))+ylim(0,3)+geom_smooth(aes(group=1))

# plot ggplot with regression line
p1+geom_smooth(method="lm", aes(group=1)) + geom_point() +ylim(0,3)

# add regression equation to plot
lm_eqn <- function(nat.rates){
    m <- lm(Rate ~ Year, nat.rates);
    eq <- substitute(italic(y) == a + b %.% italic(x)*","~~italic(r)^2~"="~r2,
         list(a = format(coef(m)[1], digits = 2),
              b = format(coef(m)[2], digits = 2),
             r2 = format(summary(m)$r.squared, digits = 3)))
    as.character(as.expression(eq));
}

p1+geom_smooth(method="lm", aes(group=1)) + geom_point() +ylim(0,3)
  + geom_text(x = 1984, y = 1, label = lm_eqn(nat.rates), parse = TRUE)
```

choropleth

```
nstall.packages ("choroplethr")
library(choroplethr)

states_map <- map_data("state")
region <- tolower(state.name)[-c(2,11)] # removing alaska hawaii
rate.vector <- rowMeans(inc.dense.df)
df<- data.frame(region,rate.vector)

pal <- colorRampPalette(c('grey10','darkgreen'))(100)
mp <- ggplot(df, aes(map_id=region))
     + geom_map(aes(fill=rate.vector), map=states_map)
     + coord_map("polyconic")
 + expand_limits(x = states_map$long, y = states_map$lat)
mp <- mp
 + scale_fill_gradient(low='grey90', high='darkgreen', limits=c(0,3))
mp
```

Poisson model

$$y_i \sim Poisson(\theta_i)$$
$$\theta_i = exp(X_i\beta)$$

- count data
- glm, log link
- θ constrained to be positive, fit on logarithmic scale
- each unit i is a *setting*, such as a time interval or spatial location, in which y_i events have occurred,
  - e.g. traffic crashes at intersection i in a given year
  - linear predictors X e.g. continuous measure average speed, indicator for traffic light
- note: if outcome is count or number of "successes" in some number of trials, standard to use binomial/logistic
  - if no natural limit on the number of outcomes, standard to use Poisson

offset variable

- makes sense to include a measure of *exposure*, υ

$$y_i \sim Poisson(\upsilon_i \theta_i)$$

- *log* υ called the *offset* variable
- a kind of baseline predictor in the model, equivalent to a regression coefficient with coefficient value fixed to 1

predictive interpretation Poisson regression coefficients

Gelman and Hill

- traffic crash model: effect of speed and traffic lights at intersections

$$y_i \sim Pois(e^{2.8+0.012X_{i1}-0.20X_{i2}})$$

- *intercept* (2.8) - crashes when speed is zero and no light, uninterpretable
- *speed coefficient* (X_{i1}) - expected difference on log scale for each addition mph average speed,
  - expected multiplicative increase is $e^{0.0012} = 1.012$, or 1.2% increase car crash rate for each 1 mph increase
  - might make more sense to multiply this by 10, so $e^{0.012} = 1.127$ for a 12.7% increase in crash rate per ten mph increase
- *traffic light indicator coefficient* (X_{i2}) - predictive difference of having a traffic light
  - multiply crash rate by $e^{-0.20} = 0.82$, or 18% reduction

Poisson regression of traffic fatalities

effect of law enforcement vs economic

```
model1 <- glm(fatal ~ year,
 offset = log(milestot),family = poisson, data=Fatalities)
summary(model1)
str(model1)
exp(model1$coefficients)
exp(coef(model1))
exp(confint(model1))

model2 <- glm(fatal ~ year+state,
 offset = log(milestot),family = poisson, data=Fatalities)
summary(model2)

model3 <- glm(fatal ~ year+state+jail,
 offset = log(milestot),family = poisson, data=Fatalities)
summary(model3)
exp(coef(model3))
exp(confint(model3))
```


YOUR
TURN

- run a Poisson model for the effect of a beer tax ("beertax") on traffic fatalities controlling for year and state
- include an offset variable for total miles driven
- compare the results to the those for the effect of mandatory jail sentences

beer tax model

```
model4 <- glm(fatal ~ year+state+beertax,
 offset = log(milestot),family = poisson, data=Fatalities)
summary(model4)
exp(coef(model4))
exp(confint(model4))
```

overdispersion in Poisson models

- Poisson variance is equal to mean, so s.d. is square root of the mean

$$E(y_i) = v_i\theta_i$$
$$sd(y_i) = \sqrt{v_i\theta_i}$$

- standardized residuals are

$$z_i = \frac{y_i - \hat{y}_i}{sd(\hat{y}_i)}$$
$$= \frac{y_i - v\hat{\theta}_i}{\sqrt{v_i\hat{\theta}_i}}$$

- if Poisson model true, expect z_i to have mean 0 and sd=1

testing for overdispersion

- compare sum of squares of z_i (Σz_i^2) to Chi square with n-k d.f.
- χ^2_{n-k} has average value of n-k, so $\frac{\Sigma z_i^2}{n-k}$ is an estimate of overdispersion
- values above 2 considered large
- R code from Gelman and Hill
  - set n to *nrow(data)* and k to the number of predictors

```
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")
```

- goodness of fit chi square test based on residuals and their df's

```
1 - pchisq(summary(model.pois)$deviance,
    summary(model.pois)$df.residual)
```

adjusting for overdispersion

- can multiply all regression s.e.'s by $\sqrt{overdispersion}$
- fit "quasipoisson" family or negative binomial model

```
model4 <- glm(fatal ~ year+state+beertax,
offset = log(milestot),family = poisson, data=Fatalities)
yhat <- predict(model4, type="response")
z <- (Fatalities$fatal-yhat)/sqrt(yhat)
sum(z^2)/(nrow(Fatalities)-(48+2))
#multiply s.e.'s by sqrt(5.897498), or...

library(MASS)
mod.nb<-glm.nb(fatal ~ year+state+beertax, offset(log(milestot
yhat <- predict(mod.nb, type="response")
z <- (Fatalities$fatal-yhat)/sqrt(yhat)
sum(z^2)/(nrow(Fatalities)-(48+2))
```