Longitudinal Data: Fixed Covariates, Time-dependent Outcomes

Binary Outcomes Part 2

Alan Hubbard

Sept 25, 2018

Census-type count data

- Following, we examine a relatively large data set on counts of types of mortality episodes, type of crimes, and counts of "disasters" measured at county level across the United States over nd sp of onverting repeated binary outcomes to a count (number of events) repeatedly over the years 1994-2004.
- If we treat county as the random unit, then this could be an example of repeated measures data and we will revist this data in later chapters.
- For now, we will choose a single year, 2000, as the year of our analyses examining rates of suicide and accidents as related to other factors, including previous disasters. In stead of using dplyr to subset data before running the regression, we instead do the subseting within the regression functions.

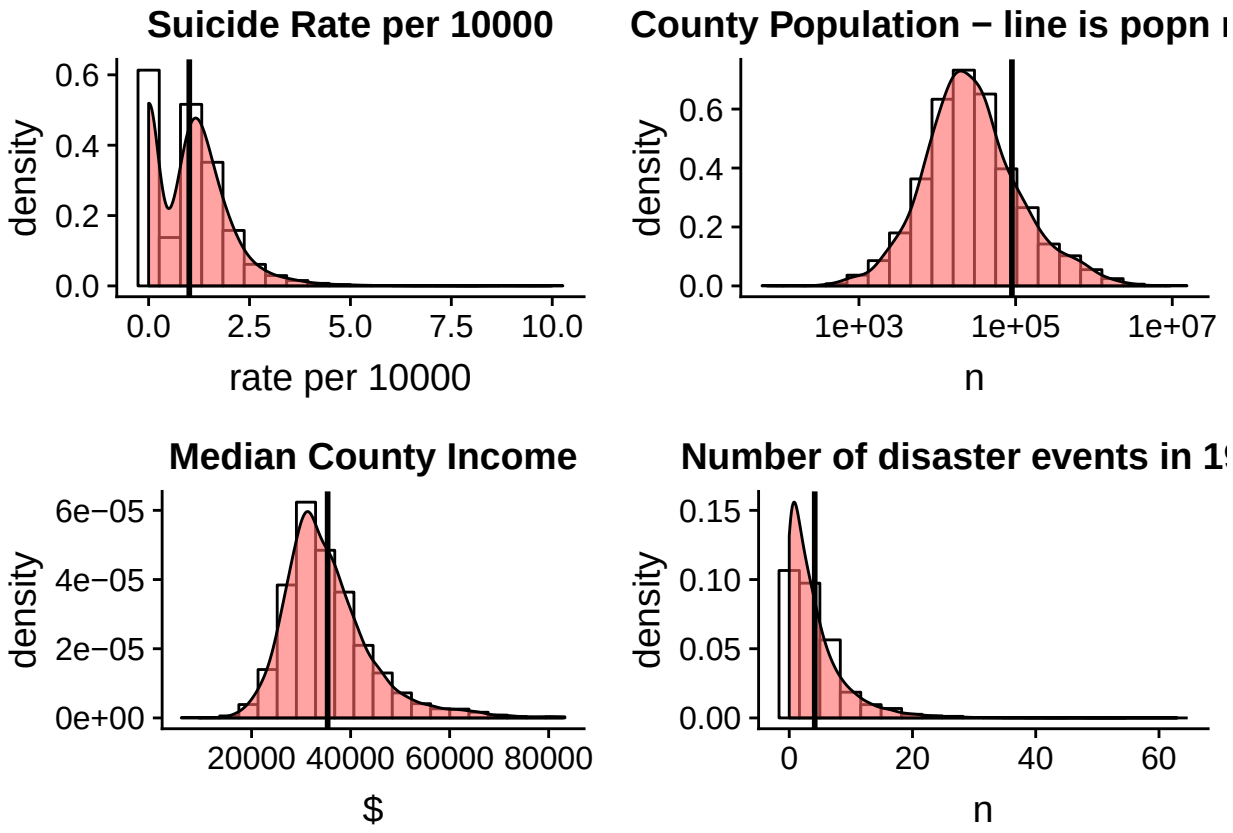
```
library(dplyr)
library(MASS)
library(tidyverse)
library(cowplot)
# get the regression output formatting function
source("glm_post_estimate.R")


dat=read_csv("allstack_04_14_08.csv")
dat=rename(dat,county_code=fips)
# read in county identifiers
county=read_csv("countycodes.csv")
# merge data
dat=left_join(dat,county, by = "county_code")
```

Data Exploration

- We will examine the distribution of counts of suicide by county and their association with other factors. The variable logpop00 is the natural log of the population size of the county.

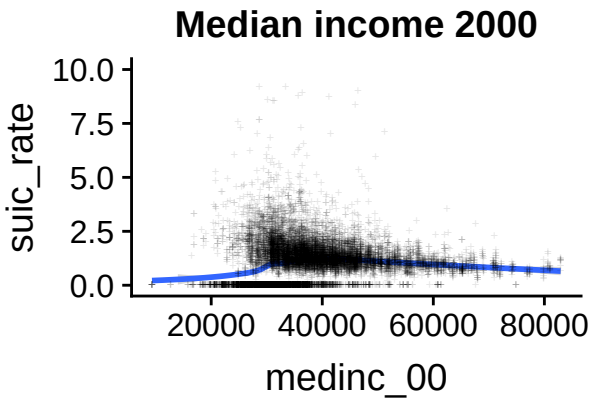
- We examine the distribution of the variables variables relevant to our goal, which is to determine the joint association of the number of disasters in the previous year (prev1i), the median income for the county (medinc_00), adjusting for differences in county size (logpop00).
- To look at distribution of suicide rates, we first make a variable that is the rate. Let Y be the count of suicides and T the population size of county, we make a new variable as the rate per 10000 people or $rate = 1000 * Y/T$. Then we examine histograms of the relevant variables.

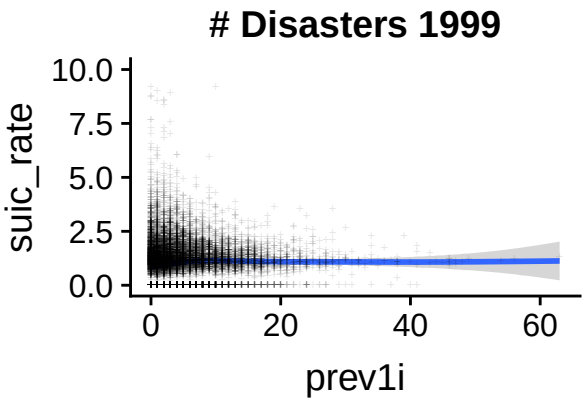
```r
dat2000=subset(dat,year=2000) %>%
  mutate(totpop=exp(logpop00)) %>%
  mutate(suic_rate=10000*nsui/totpop)
# given the very few observations > 10, we lump all of them at 10
p1 <- ggplot(dat2000, aes(x=pmin(suic_rate,10)))+
 geom_histogram(aes(y=..density..), colour="black", fill="white",bins=20) +
 geom_density(alpha=.6, fill="#FF6666",adjust=2) +ggtitle("Suicide Rate per 10000") +xlab("rate per 100
 geom_vline(xintercept=mean(dat2000$suic_rate,na.rm=T),size=1)

p2 <- ggplot(dat2000, aes(x=totpop)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white",bins=20) +
 geom_density(alpha=.6, fill="#FF6666",adjust=2)+
  ggtitle("County Population - line is popn mean") +xlab("n") +
   scale_x_log10() +
 geom_vline(xintercept=mean(dat2000$totpop,na.rm=T),size=1)

p3 <- ggplot(dat2000, aes(x=medinc_00)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white",bins=20) +
 geom_density(alpha=.6, fill="#FF6666",adjust=2)+
  ggtitle("Median County Income") +xlab("$") +
 geom_vline(xintercept=mean(dat2000$medinc_00,na.rm=T),size=1)

p4 <- ggplot(dat2000, aes(x=prev1i)) +
 geom_histogram(aes(y=..density..), colour="black", fill="white",bins=20) +
 geom_density(alpha=.6, fill="#FF6666",adjust=2)+
  ggtitle("Number of disaster events in 1999") +xlab("n") +
 geom_vline(xintercept=mean(dat2000$prev1i,na.rm=T),size=1)

plot_grid(p1,p2,p3,p4)
```

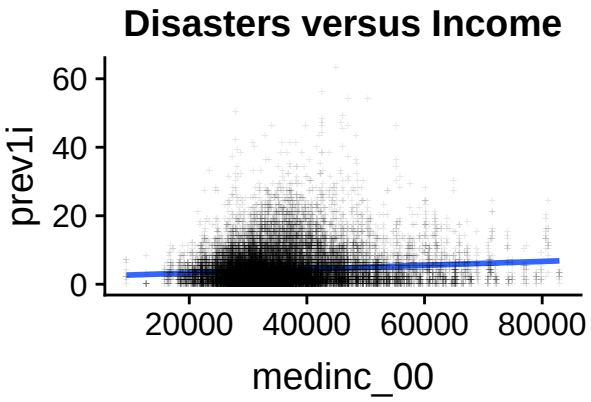


- One can see the suicide rate is strongly right skewed, more characteristic of a negative binomial distribution. We do not rely on particular distributions for the other variables, but such plots are good to look for outliers or other data issues.
- We next examine the relationship of the outcome to the two explanatory variables.

```
p1 <- ggplot(dat2000, aes(x=medinc_00,y=suic_rate)) +
 geom_smooth()+geom_point(shape="+",alpha=0.1)+ggtitle("Median income 2000") +
  scale_y_continuous(limits = c(0, 10))



p2 <- ggplot(dat2000, aes(x=prev1i,y=suic_rate)) +
 geom_smooth()+geom_point(shape="+",alpha=0.1)+scale_y_continuous(limits = c(0, 10))+ggtitle("# Disaste

p3 <- ggplot(dat2000, aes(x=medinc_00,y=prev1i)) +
 geom_smooth(method = "lm", fill = NA)+geom_point(shape="+",alpha=0.1)+ggtitle("Disasters versus Income


plot_grid(p1,p2,p3)
```







- One can see evidence of a non-linear associaton of income and suicide rate, less so with previous years disasters. There also appears to be relatively week relationship between disasters and income.

Poisson Regression

- We now examine the association of each within a single model of $log[E(Y \mid X_1, X_2, T)] = b_0 + log(T) + b_1X_1 + b_2X_2$, with X_1 being income, X_2 is previous disasters.
- We do so first assuming a Poisson distribution

```
glm_pois <- glm(nsui~medinc_00+prev1i+offset(logpop00),data=dat2000,family=poisson())
sum_glm_pois <- summary(glm_pois)
sum_glm_pois
```

```
##
## Call:
## glm(formula = nsui ~ medinc_00 + prev1i + offset(logpop00), family = poisson(),
##     data = dat2000)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -13.1494   -1.2532   -0.3491    0.9806   12.7923
##
## Coefficients:
##               Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -8.588e+00  8.782e-03 -977.927   <2e-16 ***
```

```
## medinc_00   -1.201e-05  1.925e-07  -62.378   <2e-16 ***
## prev1i      -1.331e-04  2.431e-04   -0.548    0.584
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 46821  on 14856  degrees of freedom
## Residual deviance: 42780  on 14854  degrees of freedom
##   (16523 observations deleted due to missingness)
## AIC: 89467
##
## Number of Fisher Scoring iterations: 4
```

```
summary(dat2000$medinc_00,na.rm=T)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    9333   29627   33759   35358   39389   82929
```

```
summary(dat2000$prev1i,na.rm=T)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   3.000   4.137   6.000  63.000
```

- Given the very large sample size, $m = 14857$ it is surprising that both coefficients are not significantly different from 0: t the estimate of β_1 is 0 with p-value of 0, whereas the estimate of β_2 is 0 with p-value of 0.584.
- These coefficients are estimates of the log ratio of means for an increase in the respective variables by 1 unit.
- However, for both variables, 1 unit is not an interesting increase, so though associations are significant, the coefficients are small as expected.
- We can easily get ratios for bigger changes, say a change equivalent to the interquarile range of the predictors. We do that below.

```
iqr_inc = IQR(dat2000$medinc_00,na.rm=T)
iqr_dis = IQR(dat2000$prev1i,na.rm=T)
#
comps <- rbind(c(0,iqr_inc,0),c(0,0,iqr_dis))
irr_pois <- glm.post.estimate(glm_pois,comps,labs=c("income","disasters"),exponentiate = T)
irr_pois
```

```
##           Ratio.est CI              pvalue
## income    "0.889"   "0.886 - 0.893" "0.000"
## disasters "0.999"   "0.997 - 1.002" "0.584"
```

- These ratios represent $exp(9762 * b_1)$ and $exp(5 * b_2)$.
- Thus, an increase in income equivalent to the IQR of the distribuiton of income is associated with a 0.889 change in the rate of suicide (a 11.1 % decrease in rate) whereas an equivalnet change in number of disasters results in almost no change (ratio very close to 1).

- We fit the same model assuming a negative binomial (NB) distribuiton next and then compare the two fits.

Negative binomial regression

```
library(MASS)
glm_NB <- glm.nb(nsui~medinc_00+prev1i+offset(logpop00),data=dat2000,na.action=na.omit)
sum_glm_NB <- summary(glm_NB)
sum_glm_NB
```

```
##
## Call:
## glm.nb(formula = nsui ~ medinc_00 + prev1i + offset(logpop00),
##     data = dat2000, na.action = na.omit, init.theta = 11.36297083,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.7298  -1.1354  -0.3834   0.4841   6.3187
##
## Coefficients:
##               Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -8.531e+00  1.615e-02 -528.237  < 2e-16 ***
## medinc_00   -1.136e-05  3.741e-07  -30.372  < 2e-16 ***
## prev1i      -2.730e-03  6.273e-04   -4.352 1.35e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(11.363) family taken to be 1)
##
##     Null deviance: 20816  on 14856  degrees of freedom
## Residual deviance: 19921  on 14854  degrees of freedom
##   (16523 observations deleted due to missingness)
## AIC: 76388
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  11.363
##           Std. Err.:  0.299
##
##  2 x log-likelihood:  -76380.089
```

```
irr_NB <- glm.post.estimate(glm_NB,comps,labs=c("income","disasters"),exponentiate = T)
irr_NB
```

```
##           Ratio.est CI              pvalue
## income    "0.895"   "0.889 - 0.901" "0.000"
## disasters "0.986"   "0.980 - 0.993" "0.000"
```

- NB regression produces very similar estimates to those produced by Poisson regression, but the small change in the coefficient related to disasters now results in it being a significant associaiton.

- Like above, these ratios represent $exp(9762 * b_1)$ and $exp(5 * b_2)$.

- Besides the small shift in the estimate for the ratio of means for a in crease in disasters from the POisson to Negative Binomial, one might consider the results to be significantly different.

- But if you look at the CI's for both, though one contains 1 and the other doesn't, it's a very small CI close to the null, thus, little evidence of a significant adjusted (for income) association of linear number of disasters and suicide rate.
- The the CI's returned by both Poisson and NB appear small in absolute number, they are proportionally larger for the NB regression, which is what one expects - it typically returns more conservative inference than the Poisson assumption (implies greater residual variability).
- Equivalently, the standard errors are larger for the associations (coefficients) for the NB versus poisson, e.g., $SE(\beta_1) = 0$ in the NB regression is much bigger than that estimated in the Poisson regression (0).

- We can also look at whether the difference is statistically significant.

```
library("lmtest")
lrtest(glm_NB, glm_pois)
```

```
## Likelihood ratio test
##
## Model 1: nsui ~ medinc_00 + prev1i + offset(logpop00)
## Model 2: nsui ~ medinc_00 + prev1i + offset(logpop00)
##   #Df LogLik Df Chisq Pr(>Chisq)
## 1   4 -38190
## 2   3 -44731 -1 13081  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The p-value is very small, suggesting that the NB is a significantly bette fit than the Poisson distribution. This matches what one would anticipate, given the highly skewed distribution of suicide counts.