Gov 2001 Section 8: Continuing with Binary and Count Outcomes

Konstantin Kashin[1]

March 27, 2013

[1]Thanks to Jen Pan, Brandon Stewart, Iain Osgood, and Patrick Lam for contributing to this material.

Outline

Replication Paper

- You will receive a group to re-replicate tonight
- Re-replication due Wednesday, April 3 at 7pm
- Aim to be helpful, not critical!
- Any questions about expectations?

Outline

Why Zero-Inflation?

- What if we knew that something in our data were mismeasured?
- For example, what if we thought that some of our data were sytematically zero rather than randomly zero? This could be when:
  1. Some data are spoiled or lost
  2. Survey respondents put “zero” to an ordered answer on a survey just to get it done.

If our data are mismeasured in some systematic way, our estimates will be off.

A Working Example: Fishing

You're trying to figure out the probability of catching a fish in a park from a survey. People were asked:

- How many children were in the group
- How many people were in the group
- Whether they caught a fish.

A Working Example: Fishing

The problem is, some people didn't even fish! These people have systematically zero fish.

The Model

We're going to assume that whether or not the person fished is the outcome of a Bernoulli trial.

$$Y_i = \begin{cases} 0 & \text{with probability } \psi_i \\ \text{Logistic} & \text{with probability } 1 - \psi_i \end{cases}$$

The Model

We can write out the distribution of Y_i as:

$$P(Y_i = y_i|\beta, \psi_i) \begin{cases} \psi_i + (1 - \psi_i)\left(1 - \frac{1}{1+e^{-X\beta}}\right) & \text{if } y_i = 0 \\ (1 - \psi_i)\left(\frac{1}{1+e^{-X\beta}}\right) & \text{if } y_i = 1 \end{cases}$$

And we can put covariates on ψ:

$$\psi = \frac{1}{1 + e^{-z_i\gamma}}$$

Deriving the Likelihood

The likelihood function is proportional to the probability of Y_i:

$$\begin{aligned}
L(\beta, \psi_i | Y_i) &\propto P(Y_i | \beta, \psi_i) \\
&= \left[\psi_i + (1 - \psi_i)\left(1 - \frac{1}{1 + e^{-X_i\beta}}\right)\right]^{1-Y_i} \\
&\quad \left[(1 - \psi_i)\left(\frac{1}{1 + e^{-X_i\beta}}\right)\right]^{Y_i} \\
&= \left[\frac{1}{1 + e^{-z_i\gamma}} + \left(1 - \frac{1}{1 + e^{-z_i\gamma}}\right)\left(1 - \frac{1}{1 + e^{-X_i\beta}}\right)\right]^{1-Y_i} \\
&\quad \left[\left(1 - \frac{1}{1 + e^{-z_i\gamma}}\right)\left(\frac{1}{1 + e^{-X_i\beta}}\right)\right]^{Y_i}
\end{aligned}$$

Deriving the Likelihood

Multiplying over all observations we get:

$$L(\beta, \gamma|Y) = \prod_{i=1}^{n} \left[\frac{1}{1+e^{-z_i\gamma}} + \left(1 - \frac{1}{1+e^{-z_i\gamma}}\right)\left(1 - \frac{1}{1+e^{-X_i\beta}}\right)\right]^{1-Y_i} \left[\left(1 - \frac{1}{1+e^{-z_i\gamma}}\right)\left(\frac{1}{1+e^{-X_i\beta}}\right)\right]^{Y_i}$$

Deriving the Likelihood

Taking the log we get:

$$\begin{aligned}
\ln L &= \sum_{i=1}^{n} \left\{ Y_i \ln\left[(1-\psi)\left(\frac{1}{1+e^{-X_i\beta}} \right) \right] + \right. \\
&\quad \left. (1-Y_i)\ln\left[\psi + (1-\psi)\left(1 - \frac{1}{1+e^{-X_i\beta}}\right)\right] \right\} \\
&= \sum_{i=1}^{n} \left\{ Y_i \ln\left[\left(1 - \frac{1}{1+e^{-z_i\gamma}}\right)\left(\frac{1}{1+e^{-X_i\beta}} \right) \right] + \right. \\
&\quad \left. (1-Y_i)\ln\left[\frac{1}{1+e^{-z_i\gamma}} + \left(1 - \frac{1}{1+e^{-z_i\gamma}}\right)\left(1 - \frac{1}{1+e^{-X_i\beta}}\right)\right] \right\}
\end{aligned}$$

Let's program this in R

Load and get the data ready:

```
fish <- read.table("http://www.ats.ucla.edu/stat/R/dae/fish.csv"
X <- fish[c("child", "persons")]
Z <- fish[c("persons")]
X <- as.matrix(cbind(1,X))
Z <- as.matrix(cbind(1,Z))
y <- ifelse(fish$count>0,1,0)
```

Let's program this in R

Write out the Log-likelihood function

```
ll.zilogit <- function(par, X, Z, y){
beta <- par[1:ncol(X)]
gamma <- par[(ncol(X)+1):length(par)]
phi <- 1/(1+exp(-Z%*%gamma))
pie <- 1/(1+exp(-X%*%beta))
sum(y*log((1-phi)*pie) + (1-y)*(log(phi + (1-phi)*(1-pie))))
}
```

Let's program this in R

Optimize to get the results

```
par <- rep(1,(ncol(X)+ncol(Z)))
out <- optim(par, ll.zilogit, Z=Z, X=X,y=y, method="BFGS",
      control=list(fnscale=-1), hessian=TRUE)
```

```
out$par
[1]  1.507470 -2.686476  1.447307  1.876404 -1.247189
```

Plotting to See the Relationship

These numbers don't mean a lot to us, so we can plot the predicted probabilities of a person having not fished.

First, we have to simulate our gammas:

```
varcv.par <- solve(-out$hessian)
library(mvtnorm)
sim.pars <- rmvnorm(10000, out$par, varcv.par)
sim.z <- sim.pars[,(ncol(X)+1):length(par)]
```

Plotting to See the Relationship

These numbers don't mean a lot to us, so we can plot the predicted probabilities of a group having not fished.

We then generate predicted probabilities that different sized groups did not fish.

```
person.vec <- seq(1,4)
Zcovariates <- cbind(1, person.vec)
exp.holder <- matrix(NA, ncol=4, nrow=10000)
for(i in 1:length(person.vec)){
exp.holder[,i] <- 1/(1+exp(-Zcovariates[i,]%*%t(sim.z)))
}
```

Plotting to See the Relationship

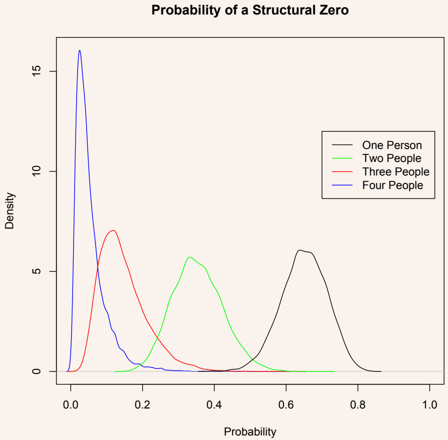
These numbers don't mean a lot to us, so we can plot the predicted probabilities of a group having not fished.

Using these numbers, we can plot the densities of probabilities, to get a sense of the probability and the uncertainty.

```
plot(density(exp.holder[,4]), col="blue", xlim=c(0,1),
    main="Probability of a Structural Zero", xlab="Probability")
lines(density(exp.holder[,3]), col="red")
lines(density(exp.holder[,2]), col="green")
lines(density(exp.holder[,1]), col="black")
legend(.7,12, legend=c("One Person", "Two People",
   "Three People", "Four People"),
     col=c("black", "green", "red", "blue"), lty=1)
```

Plotting to See the Relationship

These numbers don't mean a lot to us, so we can plot the predicted probabilities of a group having not fished.



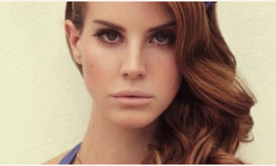
Outline

The Poisson Distribution

It's a discrete probability distribution which gives the probability that some number of events will occur in a fixed period of time.

Examples:

1. number of terrorist attacks in a given year
2. number of publications by a professor in a career
3. number of times word "hope" is used in a Barack Obama speech
4. number of songs on a pop music CD

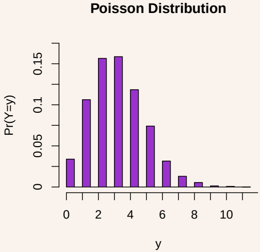

The Poisson Distribution

-Here's the probability density function (PDF) for a random variable Y that is distributed Pois(λ):

$$\Pr(Y = y) = \frac{\lambda^y}{y!} e^{-\lambda}$$

-Suppose $Y \sim$ Pois(3). What's $\Pr(Y = 4)$?

$$\Pr(Y = 4) = \frac{3^4}{4!} e^{-3} = 0.168.$$



The Poisson Distribution

One more time, the probability density function (PDF) for a random variable Y that is distributed $\text{Pois}(\lambda)$:

$$\Pr(Y = y) = \frac{\lambda^y}{y!} e^{-\lambda}$$

Using a little bit of geometric series trickery, it isn't too hard to show that $E[Y] = \sum_{y=0}^{\infty} y \cdot \frac{\lambda^y}{y!} e^{-\lambda} = \lambda$.

It also turns out that $\text{Var}(Y) = \lambda$, a feature of the model we will discuss later on.

The Poisson Distribution

Poisson data arises when there is some discrete event which occurs (possibly multiple times) at a constant rate for some fixed time period.

This constant rate assumption could be restated: the probability of an event occurring at any moment is independent of whether an event has occurred at any other moment.

Derivation of the distribution has some other technical first principles, but the above is the most important.

The Poisson Model for Event Counts

1. The stochastic component:

$$Y_i \sim Pois(\lambda_i)$$

2. The systematic component:

$$\lambda_i = exp(X_i\beta)$$

The likelihood is therefore:

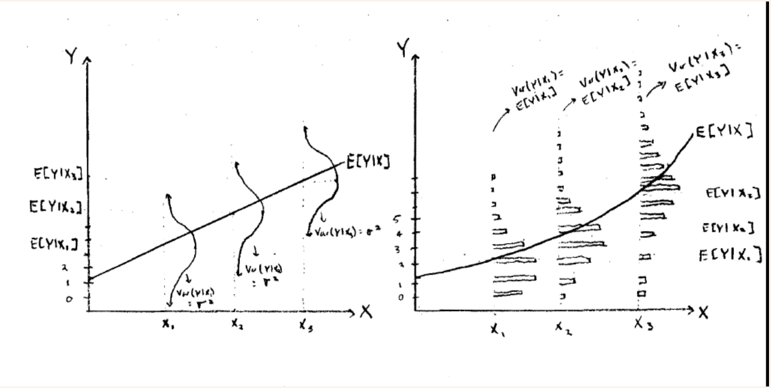
$$\mathrm{L}(\beta|X, y) = \prod_{i=1}^{n} \frac{\lambda_i^{y_i}}{y_i!} \mathrm{e}^{-\lambda_i}$$

The Poisson Model for Event Counts

And the log-likelihood

$$
\begin{aligned}
\ln \mathrm{L}(\beta|X, y) &= \sum_{i=1}^{n} y_i \ln \lambda_i - \ln(y_i!) - \lambda_i \\
&= \sum_{i=1}^{n} y_i \ln(exp(X_i\beta) - \ln(y_i!) - exp(X_i\beta) \\
&= \sum_{i=1}^{n} y_i(X_i\beta) - exp(X_i\beta)
\end{aligned}
$$

Comparing with the Linear Model



Comparing with the Linear Model

Possible dimensions for comparison:

1. distribution of $Y|X$
2. shape of the mean function
3. assumptions about $Var(Y|X)$
4. calculating fitted values
5. meaning of intercept and slope

Generally: the linear model (OLS) is biased, inefficient, and inconsistent for count data!

Example: Civil Conflict in Northern Ireland

Background: a conflict largely along religious lines about the status of Northern Ireland within the United Kingdom, and the division of resources and political power between Northern Ireland's Protestant (mainly Unionist) and Catholic (mainly Republican) communities.

The data: the number of Republican deaths for every month from 1969, the beginning of sustained violence, to 2001 (at which point, most organized violence had subsided). Also, the unemployment rates in the two main religious communities.

Example: Civil Conflict in Northern Ireland

Example: Civil Conflict in Northern Ireland

The model: Let Y_i = # of Republican deaths in a month. Our sole predictor for the moment will be: U_C = the unemployment rate among Northern Ireland's Catholics.

Our model is then:

$$Y_i \sim Pois(\lambda_i)$$

and

$$\lambda_i = E[Y_i | U_i^C] = exp(\beta_0 + \beta_1 * U_i^C).$$

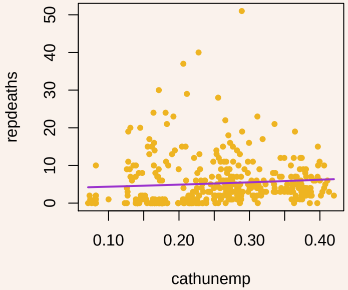
Estimate (just as we have all along!)

```
mod <- zelig(repdeaths ~ cathunemp,
         data = troubles, model = "poisson")

> summary(mod)$coefficients
            Estimate Std. Error  z value     Pr(>|z|)
(Intercept) 1.295875  0.1805327 7.178064 7.070547e-13
cathunemp   1.406498  0.6689819 2.102445 3.551432e-02
```

Our fitted model

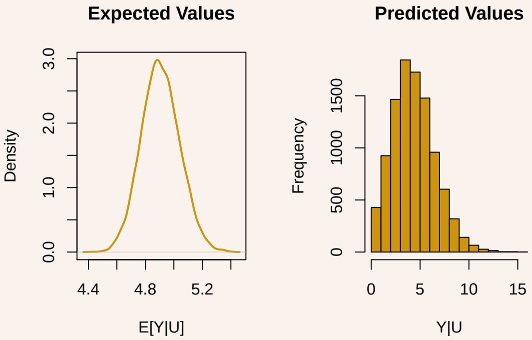
$$\lambda_i = E[Y_i | U_i^C] = exp(1.296 + 1.407 * U_i^C).$$



Some fitted and predicted values

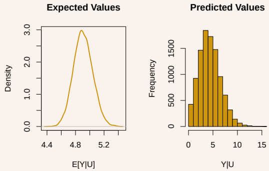
Suppose U_C is equal to .2.

```
mod.coef <- coef(mod); mod.vcov <- vcov(mod)
beta.draws <- mvrnorm(10000, mod.coef, mod.vcov)
lambda.draws <- exp(beta.draws[,1] + .2*beta.draws[,2])
outcome.draws <- rpois(10000, lambda.draws)
```



Some fitted and predicted values

Is the difference between expected and predicted values clear? What kind of uncertainty is accounted for in each of the two distributions?

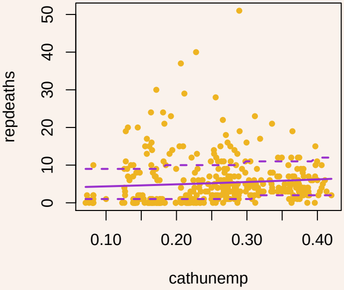


Estimation uncertainty for expected values.

Both estimation uncertainty and fundamental uncertainty for predicted values.

Overdispersion

36% of observations lie outside the 2.5% or 97.5% quantile of the Poisson distribution that we are alleging generated them.



Outline

The Negative Binomial Model

The variance of the Poisson distribution is only equal to its mean if the probability of an event occurring at any moment is independent of whether an event has occurred at any other moment, and if the occurrence rate is constant.

We can perturb this second assumption (constant rate) in order to derive a distribution which can handle both violations of the constant rate assumption and violations of the independence of events (or no contagion) assumption.

The trick is to assume that λ varies, within the same observation span, according to a new parameter we will introduce call ς.

Alternative Parameterization

Here's the new stochastic component:

$$\begin{aligned} Y_i|\lambda_i, \zeta_i &\sim Poisson(\zeta_i \lambda_i) \\ \zeta_i &\sim Gamma\left(\frac{1}{\sigma^2 - 1}, \frac{1}{\sigma^2 - 1}\right) \end{aligned}$$

Note that Gamma distribution has a mean of 1. Therefore, $Poisson(\zeta_i\lambda_i)$ has mean λ_i. Note that the variance of this distribution is $\sigma^2 - 1$. This means that as σ^2 goes to 1, the distribution of ζ_i collapses to a spike over 1.

Alternative Parameterization

Using a similar approach to that described in UPM pgs. 51-52 we can derive the marginal distribution of Y as

$$Y_i \sim Negbin(\lambda_i, \sigma^2)$$

where

$$f_{nb}(y_i|\lambda_i, \sigma^2) = \frac{\Gamma\left(\frac{\lambda_i}{\sigma^2-1} + y_i\right)}{y!\Gamma\left(\frac{\lambda_i}{\sigma^2-1}\right)} \left(\frac{\sigma^2 - 1}{\sigma^2}\right)^{y_i} (\sigma^2)^{-\frac{\lambda_i}{\sigma^2-1}}$$

Notes:

1. $\lambda_i > 0$ and $\sigma > 1$
2. $E[Y_i] = \lambda_i$ and $Var[Y_i] = \lambda_i \sigma^2$. What value of σ^2 would be evidence *against* overdispersion?
3. We still have the same old systematic component: $\lambda_i = exp(X_i \beta)$.

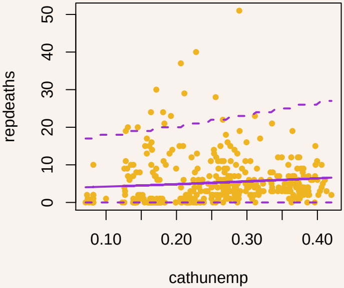
Estimates

```
mod <- zelig(repdeaths ~ cathunemp, data = troubles,
             model = "negbin")
summary(mod)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.2959     0.1805   7.178 7.07e-13 ***
cathunemp     1.4065     0.6690   2.102   0.0355 *
---
Signif. codes:  0 *** 0.001 **  0.01 *  0.05 . 0.1  1
              Theta:  0.8551
          Std. Err.:  0.0754
```

Overdispersion Handled!

5.68% of observations lie at or above the 95% quantile of the Negative Binomial distribution that we are alleging generated them.



Other Models

Note that there are many other count models:

- Generalized Event Count (GEC) Model
- Zero-Inflated Poisson
- Zero-Inflated Negative Binomial
- Zero-Truncated Models
- Hurdle Models