Likelihood Ratio Test

Jeffrey Wu

2023-08-16

FILTER TO ONE AGE GROUP AND ONE CAUSE OF DEATH: 75-84 yr old / CLRD

Convert λ which is population weighted rate to θ

Let $\lambda =$ death rate per 100,000 people and let $\theta = \frac{\lambda N}{100000}$ (one per county)

Write likelihood for one county each month then sum over all months

Truncated Poisson:

$P(X=0) = \frac{\theta^0 e^{-\theta}}{0!} = e^{-\theta}$

$P(1 \leq X \leq 10) = P(X \in TC) = \sum_{x=1}^{10} \frac{\theta^x e^{-\theta}}{x!} = e^{-\theta} p(\theta)$

$P(X=x) = \frac{\theta^x e^{-\theta}}{x!} for x \geq 11$

Combine these together to get the likelihood for our truncated Poisson distribution

$l(\vec{x_c}) = \Pi_{y=1}^{6} \Pi_{m=1}^{12} (e^{-\theta})^{I(x_{cym}=0)} (p(\theta)e^{-\theta})^{I(x_{cym} \in TC)} (\frac{\theta^{x_{cym}} e^{-\theta}}{x_{cym}!})^{I(x_{cym} \geq 11)}$

Calculate log likelihood:

$logl(\vec{x_c}) = \sum_{y=1}^{6} \sum_{m=1}^{12} [-\theta_c + logp(\theta_c) * I(x_{cym} \in TC) + (x_{cym} log\theta_c - log(x_{cym}!)) * I(x_{cym} \geq 11)]$

```
#Defining log likelihood for truncated Poisson
model1_ll = function(x,theta){
  v = 1:10
  ptheta = sum(theta^v / gamma(v+1))
  ll = 0

  for(i in 1:length(x)){
    value = -theta + log(ptheta)*(x[i] > 0 & x[i] <= 10)  + (x[i]*log(theta))*(x[i] >= 11) #no need for
    ll = ll + value
  }

  return(ll)
}
```

Truncated ZIP

$P(X=0) = w + (1-w)e^{-\theta}$

$P(1 \leq X \leq 10) = P(X \in TC) = (1-w) \sum_{x=1}^{10} \frac{\theta^x e^{-\theta}}{x!} = (1-w)e^{-\theta} p(\theta)$

$P(X=x) = (1-w) \frac{\theta^x e^{-\theta}}{x!} for x \geq 11$

Combine these together to get the likelihood for our truncated Poisson distribution

$l(\vec{x_c}) = \Pi_{y=1}^{6}\Pi_{m=1}^{12}(w+(1-w)e^{-\theta})^{I(x_{cym}=0)}((1-w)e^{-\theta}p(\theta))^{I(x_{cym}\in TC)}((1-w)\frac{\theta^x e^{-\theta}}{x_{cym}!})^{I(x_{cym}\geq 11)}$

Calculate log likelihood:

$logl(\vec{x}) = \sum_{y=1}^{6}\sum_{m=1}^{12}[log(w+(1-w)e^{-\theta}) * I(x_{cym} = 0) + log(log(1-w) + log(p(\theta_c)) - \theta) * I(x_{cym} \in TC) + (log(1-w) + x_{cym}log\theta_c - log(x_{cym}!)) * I(x_{cym} \geq 11)]$

```
#Define log likelihood of truncated ZIP
model2_ll = function(x,theta,w){
  v = 1:10
  ptheta = sum( (theta^v) / (gamma(v+1)) )
  ll = 0

  for(i in 1:length(x)){

    value1 = (log(w + (1-w)*exp(-theta))) * (x[i] == 0)
    value2 = (log(1-w) + log(ptheta) - theta) * (x[i] > 0 & x[i] <= 10)
    value3 = (log(1-w) + x[i]*log(theta) - theta) * (x[i] >= 11) #no need for constant term log(x[i]!)

    ll = ll + (value1 + value2 + value3)
  }

  return(ll)
}
```

LRT function for Truncated Poisson vs ZIP

Load data:

Define this LRT into a function:

Our hypotheses for these LRTs are:

H0: standard Poisson model is appropriate fit for the mortality dataset

HA: a zero-inflated Poisson model is a better fit for the mortality dataset than the standard Poisson model

```
ws = seq(0.01,0.99,length.out=100)

#agegroups are of the form: Less than 1 year , 55 - 64 years , 85 years and over
#cause can take values: Chronic lower respiratory diseases OR Influenza and pneumonia

PoissonLRT = function(dataset, agegroup = "55 - 64 years",cause = "Chronic lower respiratory diseases",

  x = dataset %>% filter(Age == agegroup, Cause_of_Death == cause) %>% filter(County == county) %>% arra
  x = x$Total_Deaths
  maxval = max(x)
  if(maxval == 0){maxval = 1}

  thetas = seq(0.001,maxval,length.out=100)

  ###Find maximum likelihood for model 1

  result1 = matrix(0,nrow=length(thetas),ncol=2)
  count = 0

  for (theta in thetas){
      count = count+1
```

```
      result1[count,] = c(theta, model1_ll(x,theta))
  }

  result1 = data.frame(result1)
  colnames(result1) = c("theta","ll1")
  idx1 = which(result1$ll1 == max(result1$ll1))

  ll1 = result1[idx1,]$ll1

  ###Find maximum likelihood for model 2

  result2 = matrix(0,nrow=length(thetas)*length(ws),ncol=3)
  count = 0

  for (theta in thetas){
    for (w in ws){
      count = count+1
      result2[count,] = c(theta, w , model2_ll(x,theta,w))
    }
  }

  result2 = data.frame(result2)
  colnames(result2) = c("theta","w","ll2")
  idx2 = which(result2$ll2 == max(result2$ll2))

  ll2 = result2[idx2,]$ll2

  ###Perform LRT

  TS = 2*(ll2 - ll1) #distributed chi-sq df1
  pvalue = 1-pchisq(TS,df = 1)

  decision = (pvalue < alpha)

  result_vec = c(county,round(as.numeric(result1$theta[idx1]),2),
                 round(as.numeric(ll1),2),round(as.numeric(result2$theta[idx2]),2),
                 round(as.numeric(ll2),2),round(as.numeric(TS),2),
                 round(as.numeric(pvalue),2),decision) #TRUE means reject H0

  return(result_vec)
}

test = PoissonLRT(dataset = mortality2,county = "Alpine")
```

PERFORM LRT FOR EVERY COUNTY

```
LRT_results = matrix(NA,nrow = 58, ncol = 8)
counties = unique(mortality2$County)

for (i in 1:58){
  LRT_results[i,] = PoissonLRT(mortality2,agegroup = "75 - 84 years",county = counties[i])
}
```

```r
LRT_results = data.frame(LRT_results)
colnames(LRT_results) = c("County","Theta Model 1","Likelihood Model 1",
                          "Theta Model 2","Likelihood Model 2","Test Statistic",
                          "p-value","Reject H0?")

###ROUND TO TWO DECIMAL PLACES
head(LRT_results,10)
```

```
##          County Theta Model 1 Likelihood Model 1 Theta Model 2
## 1       Alameda         10.51             623.61         10.51
## 2        Alpine          0.03              -5.61           1.3
## 3        Amador          0.64             -47.65          1.27
## 4         Butte          3.45              -5.75             5
## 5     Calaveras          0.76             -47.65          1.39
## 6        Colusa          0.15             -28.56          0.46
## 7  Contra Costa          9.76             493.29          9.76
## 8     Del Norte          0.55             -46.95          1.03
## 9     El Dorado          2.03             -26.72          2.18
## 10       Fresno          9.55             472.59          9.55
##    Likelihood Model 2 Test Statistic p-value Reject H0?
## 1              622.92          -1.39       1      FALSE
## 2               -5.23           0.76    0.38      FALSE
## 3              -47.65           0.02     0.9      FALSE
## 4               -3.45            4.6    0.03       TRUE
## 5              -47.65           0.01    0.94      FALSE
## 6              -28.55           0.01    0.93      FALSE
## 7               492.6          -1.39       1      FALSE
## 8              -46.95              0    0.99      FALSE
## 9              -26.72              0       1      FALSE
## 10              471.9          -1.39       1      FALSE
```

##WHICHEVER MODEL WINS COMPARE WITH MODEL 3 (TRUNCATED ZIP MODEL FOR EACH QUARTER)

To maximize likelihood here, just call model2_ll 4x with the subsetted quarterly datasets and get 4 max lls -> add those 4 max ll values together to get ll3