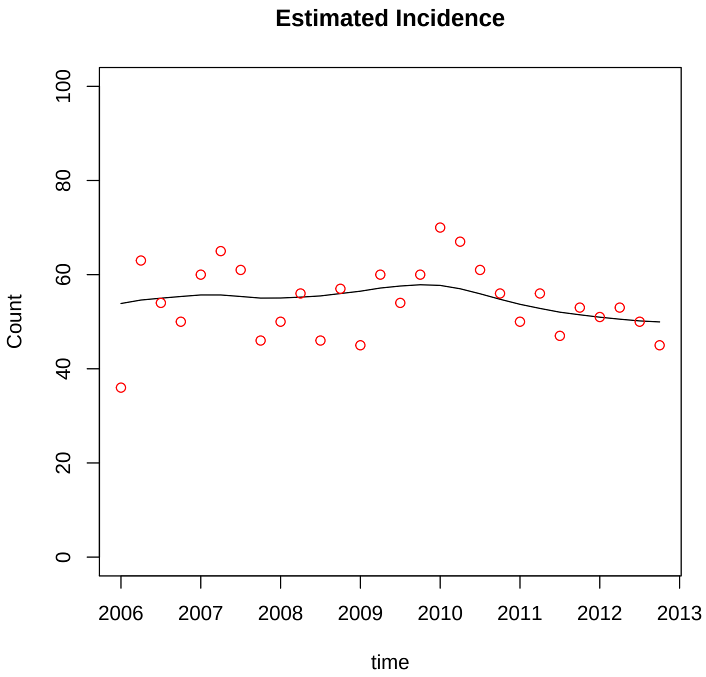
Esimtating the number of undiagnosed HIV+ MSM in King county

May 22, 2013

```
> source("data-cleaning.R")
> library(HIVBackCalc)
```

Using the methods outlined in the methodoly section, we estimate the incidence of HIV using back calculation.

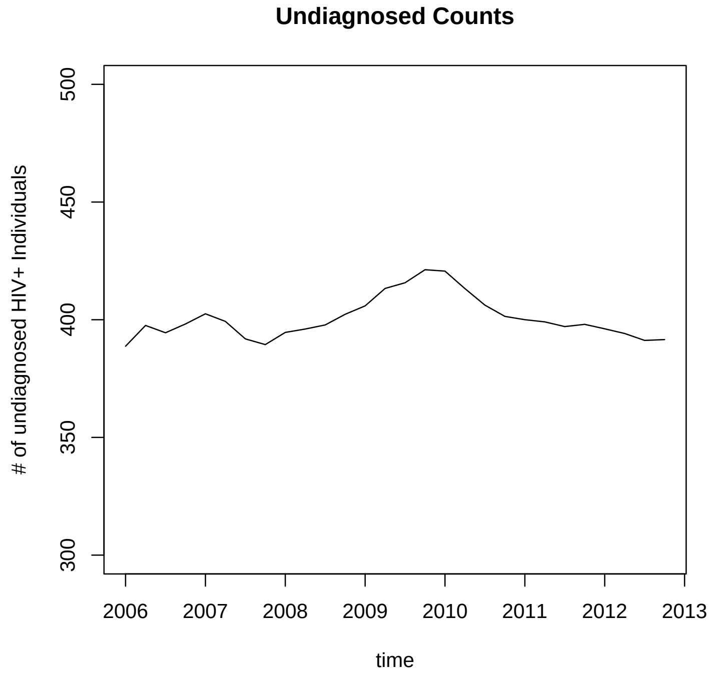
```
> intLength <- .25
> y <- c(rep(NA,100),table(msm$timeDx))
> obs <- !is.na(y)
> pid <- estimateProbDist(infPeriod=msm$infPeriod,intLength=intLength)
> mod <- estimateIncidence(y,pid,gamma=.1,tol=10^-4)
> plot(mod,time=c(2006,2012.75))
```



This indicates a steady rate of infection over the 2006 to 2012 period, with approximatly 55 individuals being infected each quarter.

Next we estiamte the counts of infected individuals who are undiagnosed.

```
> undiag <- estimateUndiagnosed(mod)
> time <- c(2006,2012.75)
> time <- seq(from=time[1],to=time[2],length.out=sum(obs))
> plot(time,undiag[obs],ylim=c(300,500),type="l",
+      main="Undiagnosed Counts",xlab="time",
+      ylab="# of undiagnosed HIV+ Individuals")
```



Finally we compare to the naive method. The naive method is to multiply incedence times average time to infection. Assuming random infections during the time between last test and diagnosis, the average time from infection to diagnosis in years is:

```
> ti <- with(msm,sort(infPeriod[!is.na(infPeriod) & infPeriod>0]))
> mean(ti/2)
```

```
[1] 1.851559
```

which, accoarding tot the formula that we were given, leads to an estimate of

```
> incidence <- mean(y,na.rm=TRUE)
> incidence * mean(ti/2)*12/3
```

```
[1] 402.5819
```

However, though this is in the right ballpark, I can't seem to derive a theoretical basis for its use. If we assume that tehre are no changes in incidence, then the diagnosis counts are estimates of the incidence counts, and the number of undiagnosed infections can be estimated with

$$\lim_{j\to\infty} y \sum_{i=0}^{j} P(D > j - i)$$

where y is the rate of infection, and D is a random variable representing the probability that someone infected with HIV waits D time intervals before being diagnosed. Using quarterly discrete time units, this yields

```
> j <- 1000
> z <- 0
> for(i in 1:j) z <- z + sum(pid((j-i+1):2000)) + .5*pid(j-i)
> z*incidence

[1] 408.5285
```

which is consistant with our model based estimates, as the incidence did not drastically change.

Using the continuous time empirical distribution, and again assuming constant incidence, we get the following estimate for undiagnosed

```
> #continuous density of time between infection and diagnosis
> pi <- function(i){
+   sapply(i,function(ii){
+     ints <- ti[ti>=ii]
+     sum(1/ints)/length(ti)
+   })
+ }
> uti <- unique(ti)
> p<-pi(uti) * diff(c(0,uti))
> cs <- cumsum(p)
> #cdf of density
> qi <- function(u){
+   uind <- rev(which(uti<=u))[1]
+   if(is.na(uind))
+     return(0)
+   cs[uind]
+ }
> n <- 10000
> m <- max(ti)
> s <- seq(from=0,to=m,length.out=n)
> l <- length(ti)
> v <- sum(sapply(s,function(x) (sum(1-qi(x)))))
> v * 4 * incidence / (n/m)

[1] 408.0196
```