A Survival Data Analysis of the Longevity of Potholes in Boston

Daniel Shank

April 29, 2016

Introduction

Boston 311 is a service launched in August of last year which expanded on the access individuals have to report service requests to the City of Boston. Originally limited to the Mayor's 24-hour hotline, citizens now have the option to report service requests through mobile app, the website Boston.gov/311, and over Twitter. Moreover, every single service request is publically available via online database, containing a point by point story which details the constant upkeep of Boston.

This analysis attempts to study a phenomenon that many commuters are intimately familiar with, the lifespan of potholes. Using survival data analysis, a comprehensive model of the lifespan of potholes in Boston was created.

Data Collection

All 311 data from Boston is made publically availble, but not all calls have the same attributes, with many potential variable assignments missing. This would occur in a small subset of the data, for instance, some potholes were labelled as having occured in Ward 0, when no such ward exists in Boston. For this reason, the data was stored locally in a NoSQL database using MongoDB. This was primarily to let us quickly manipulate the couple hundreds of thousands of service requests.

Once inside the Mongo database, Javascript queries were used to filter service requests under the categories of, "Request for Pothole Repair", "Boston Water and Sewer Pothole", and "Pothole Repair (Internal)". Using the open and close dates of each ticket, the lifetime of each pothole was recorded. No open tickets were used from the database, so no censored data was used. The source of the tickets was also available, as well as the ward in which the pothole was located, letting the model take into account location and method of detection. Once a suitable collection was created, the tens of thousand remaining service requests which were potholes and had reasonable entries for ward number and source of service request, were formatted into a csv and imported into R.

Instead of considering how long the

Description of the Suvival Model

Consider potholes in Boston to be individuals in a population, with a survival probability density function $f(t)$, which describes the length of time each individual will survive. Then let $S(t) = 1 - F(t)$ be the suvivor function, the probability that an individual is still alive at time t. We also define $h(t)$ to be the hazard function, an indication of the instantaneous risk of death. Throughout the analysis, we will attemt to accomodate more complex models of the hazards function.

$$h(t) = \frac{f(t)}{1 - F(t)} \tag{1}$$

We first select the simplest model of the hazards function, which predicts $h(x)$ to be a constant λ with respect to time. This results in an exponential distribution of survival times, with

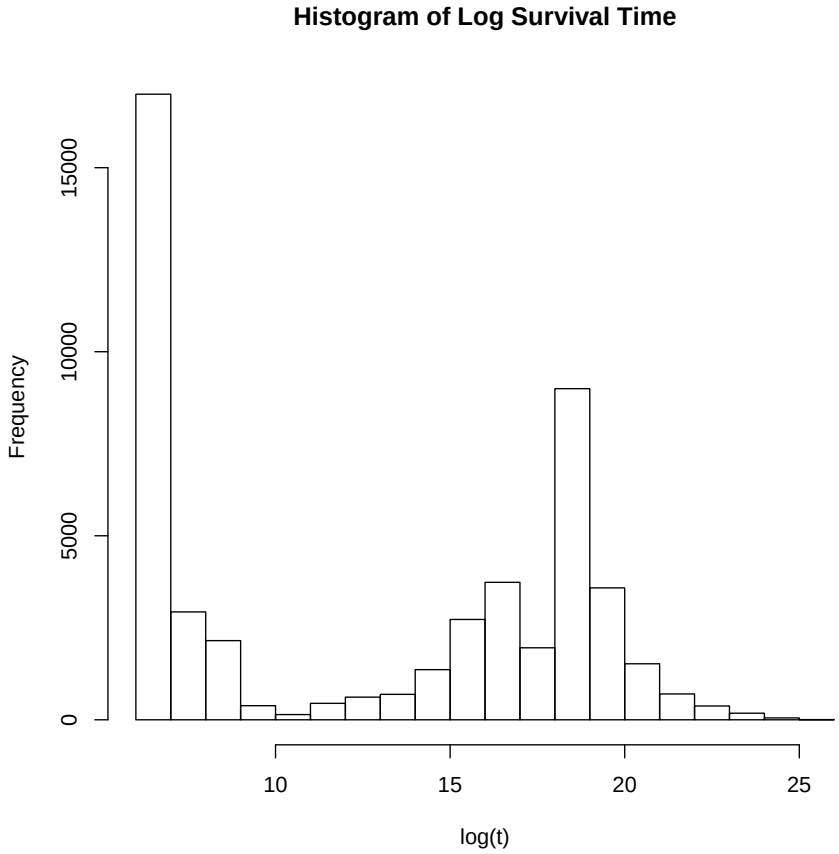
$$\begin{aligned} h(t) &= \lambda \\ f(t) &= \lambda e^{-\lambda t} \\ S(t) &= e^{-\lambda t} \end{aligned}$$

Finding a fit envokes the same generalized linear model machinery as the Poisson distribution, and more complex modeling of the hazards function is found using the Cox Proportional hazards model.

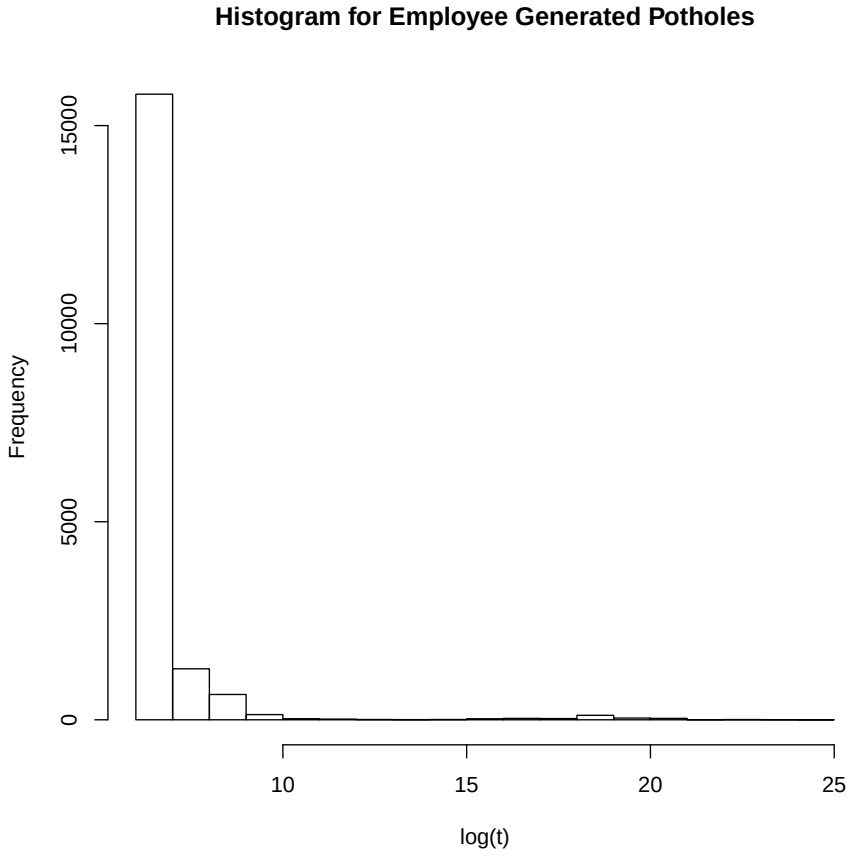
Initial Observations of Data

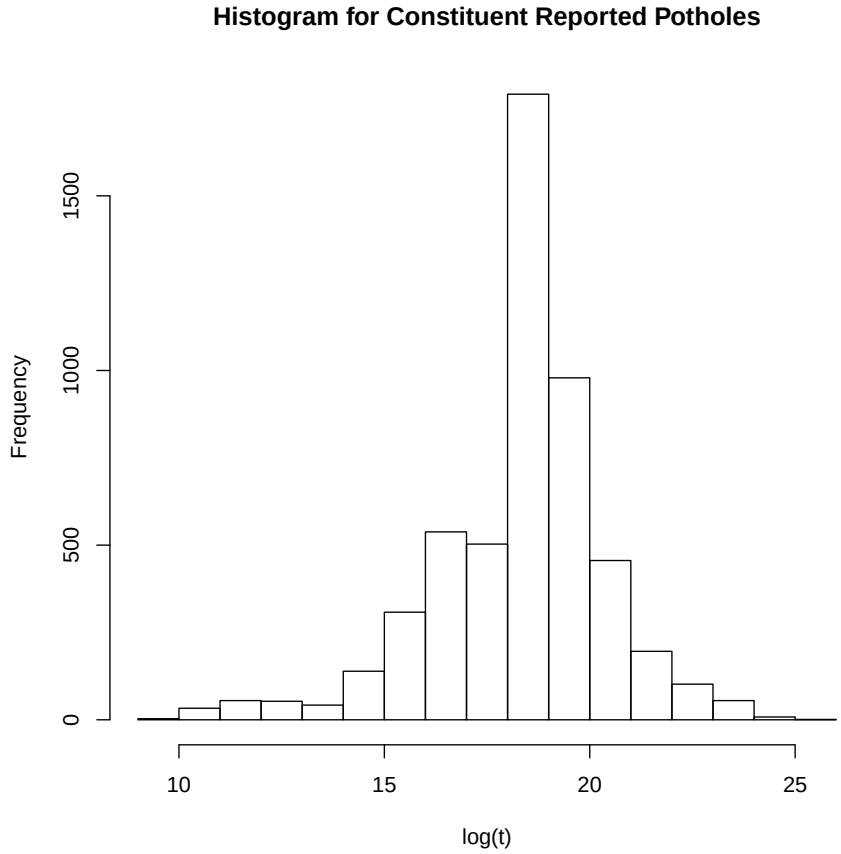
Looking at the histograms of survival times on a log scale, we can see two seperate peaks, indicating that our original assumption of an exponential distribution may not be accurate.

```
## Loading required package:  splines
```

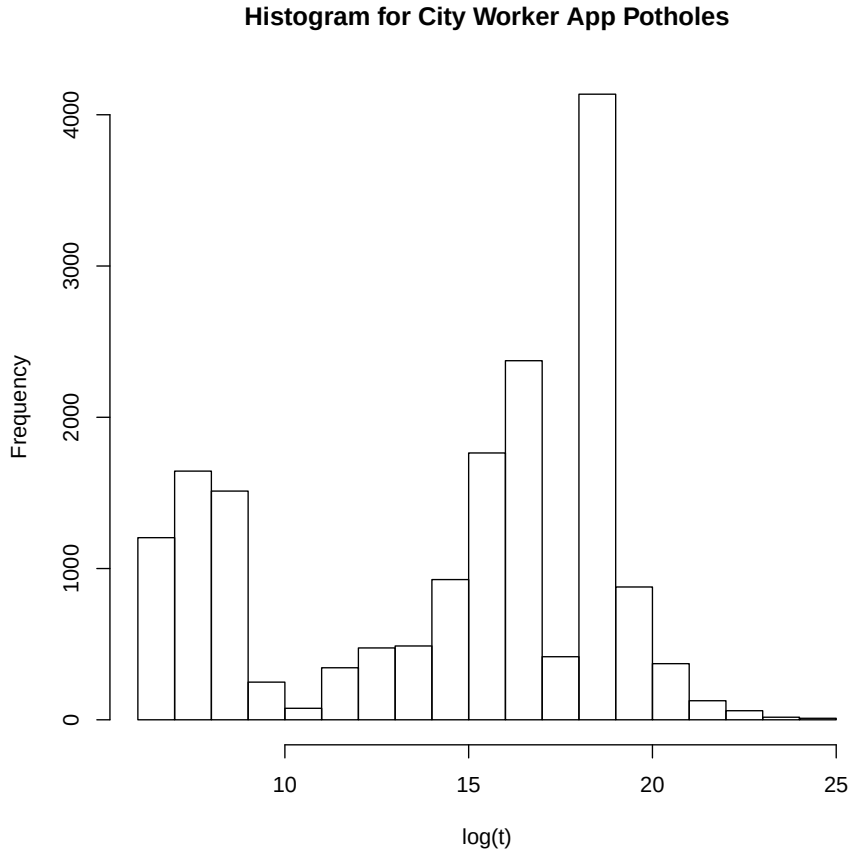


The first peak corresponds to potholes which are recorded in the database, and then filled in within the minute. These potholes are primarily employee generated, indicating that many of the potholes are detected by the city before any outside reports. The second spike refers to a time difference of approximately one day, since t has milliseconds as units. We'll see that considering these survival times by factors makes our assumptions much more reasonable.

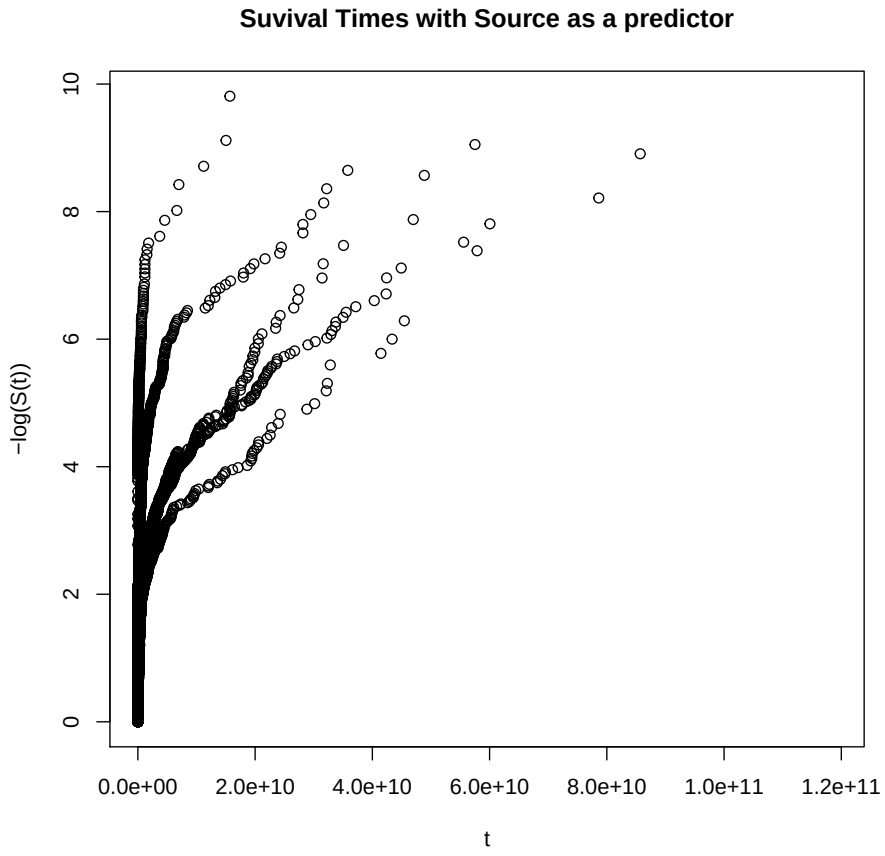

Histogram for Employee Generated Potholes
Frequency
0
5000
10000
15000
10
15
20
25
log(t)

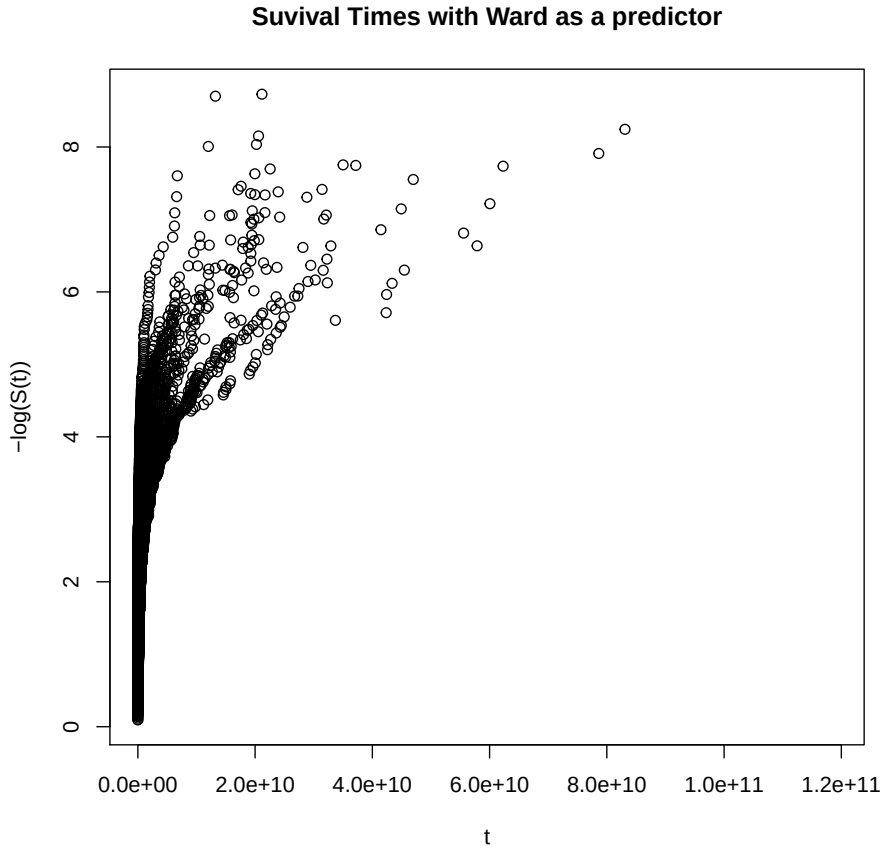


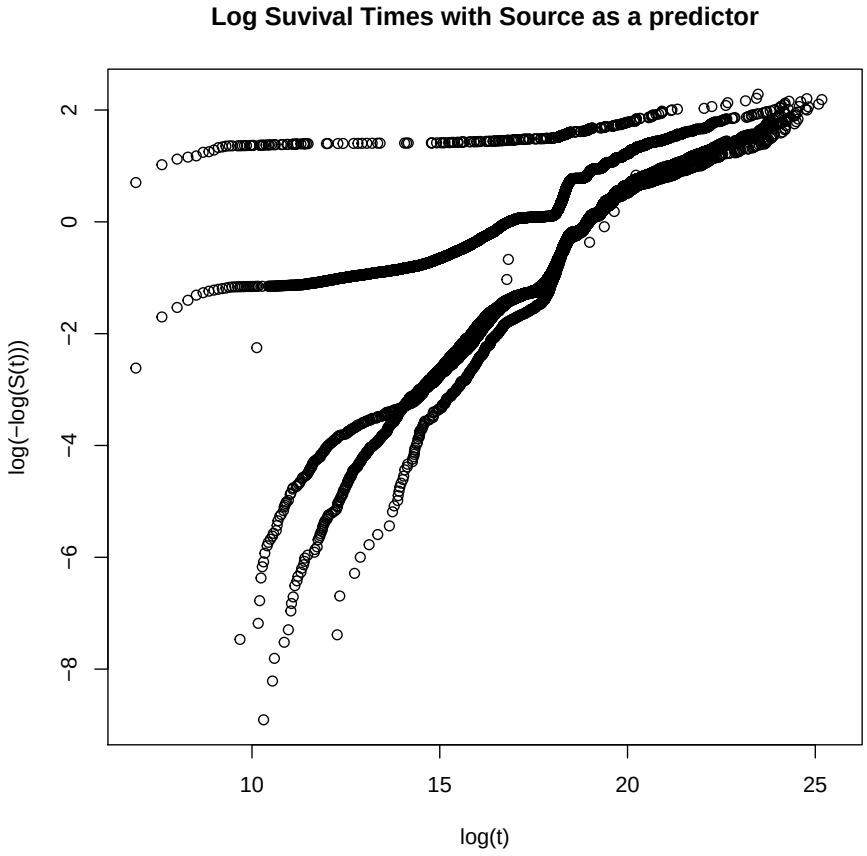
The internal record keeping has employees typically reporting opening and subsequently closing pothole cases quite quickly, with a small number on a day time scale. With constituent calls and similar outside sources such as the Citizen Connect App and Self Service, having curves centered on the day time scale. Interestingly, the City Worker App source contains both peaks, one on the day scale, and one on the order of less than a minute.

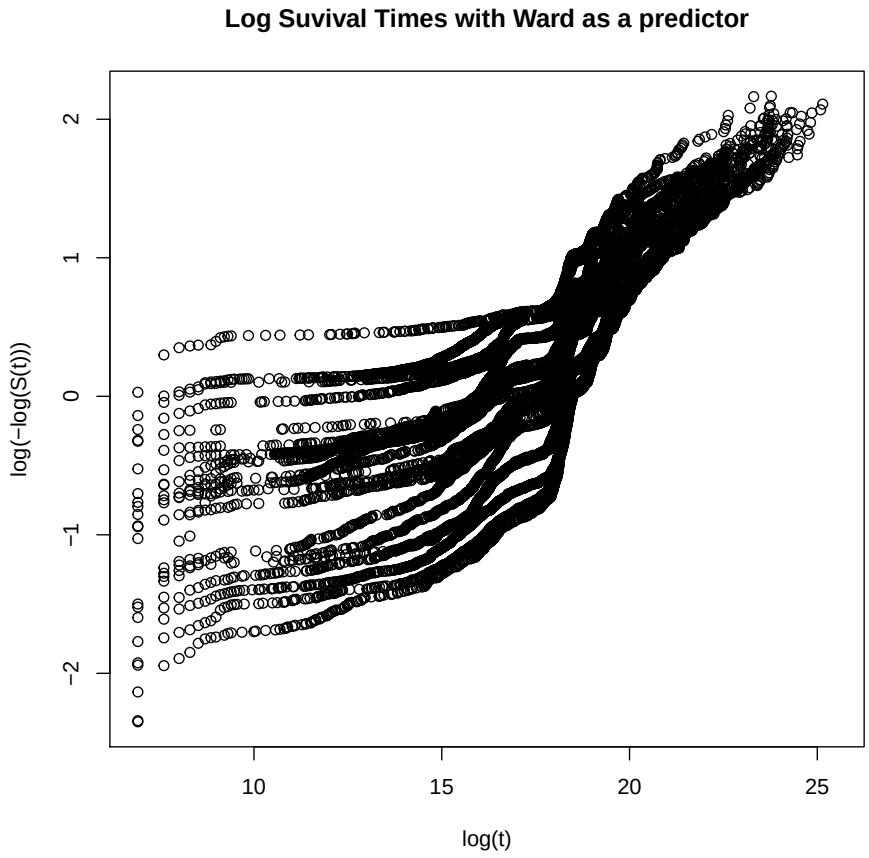


When we plot survival times by ward, we see again that many of the wards exhibit a two peak behavior. Moving onto testing our assumptions, we'll obtain the survival object as a response and make certain our assumptions about $\lambda(t)$ are accurate. This involves plotting t against $-log(\hat{S}(t))$, and if we want to see if an accelerated hazards model may be more appropriate, $log(t)$ against $log(-log(\hat{S}(t)))$. We obtain the following plots,


Suvival Times with Source as a predictor
−log(S(t))
0
2
4
6
8
10
0.0e+00
2.0e+10
4.0e+10
6.0e+10
8.0e+10
1.0e+11
1.2e+11
t


Suvival Times with Ward as a predictor
−log(S(t))
0
2
4
6
8
0.0e+00
2.0e+10
4.0e+10
6.0e+10
8.0e+10
1.0e+11
1.2e+11
t


Log Suvival Times with Source as a predictor
log(−log(S(t)))
2
0
−2
−4
−6
−8
10
15
20
25
log(t)



The linearity in the Weibull plot indicates that an accelerated hazards model with source as a predictor may be the most appropriate, with $h(t) = \lambda t^{\alpha} e^{x^T \beta}$. If we wanted to go further, there seems to be between factors and the slope α, so a further step would be to model α based off the factor level of the source of the pothole ticket.

Model Fitting

Initially fitting an exponential fit with source as a predictor,

```
## $alpha
## [1] 1
##
## $gmod
##
```

```
## Call:  glm(formula = w ~ x + offset(log(t)), family = poisson)
##
## Coefficients:
##         (Intercept)     xCity Worker App    xConstituent Call
##            -20.3202               1.6462               0.1421
## xEmployee Generated        xSelf Service             xTwitter
##              4.3027              -0.5449               0.8919
##
## Degrees of Freedom: 49549 Total (i.e. Null);  49544 Residual
## Null Deviance:     638000
## Residual Deviance: 524200  AIC: 623300
##
## $lhood
## [1] -950078.3
```

Using a chi-squared to test the residual against the null deviance with 5 df, our p-value is indistinguishable from zero,

```
pchisq(113800, 5, lower=F)

## [1] 0
```

and we get an idea of how much better our model is than simply using an average rate. The Cox Proportional Hazards model gives

```
## Call:
## coxph(formula = sl ~ source, data = holesource)
##
##
##                               coef exp(coef) se(coef)       z       p
## sourceCity Worker App      0.72431     2.063   0.0142  51.157 0.0e+00
## sourceConstituent Call    -0.00232     0.998   0.0180  -0.129 9.0e-01
## sourceEmployee Generated   2.76635    15.900   0.0157 176.350 0.0e+00
## sourceSelf Service        -0.11331     0.893   0.0275  -4.119 3.8e-05
## sourceTwitter             -0.09037     0.914   0.3165  -0.286 7.8e-01
##
## Likelihood ratio test=39878  on 5 df, p=0  n= 49550, number of events= 49550
```

Unfortunately, attempting to find a fit for the accelerated hazards did not converge in a reasonable amount of computing time, and a fit was unable to be found, but evidence indicates that a Weibull distribution would be appropriate for this data.

To summarize the two fits here, we can see the estimates for Constituent Call and Twitter seem to be a bit dubious. As we saw above, in the histogram of Constituent Reported Potholes, the two peak trend indicates Constituent Call should not explain a large amount of the deviance. There were only eight

potholes reported to the City of Boston via Twitter in the data set, so it is not suprising that its estimate is not particularly significant. Besides those two levels, the rest of the factor levels have parameters estimates within reasonable agreement.

If we say the estimates for Constituent Call and Twitter are not significantly different from zero using a Wald test, the hazards function can be found for each factor level using

$$h(t, x) = \lambda e^{x^T \beta} \tag{2}$$

to find the instanteous probability which a pothole at some time t is expected to be filled in.

Conclusions

There's a lot more to look into here, and although I was unable to obtain a fit for the accelerated hazards model, I believe that finding those estimates would explain a large amount of the deviance within the data. A follow up on using not only the source of the ticket, but a second look at using location as a predictor would be nice. Although using ward did not explain much of the data, it's possible that using a mixed-effects model with the longitude and latitude which are available in the data set would yield interesting results, specifically scaling the covariance between observations as a function of the distance between them. Another issue to address is that the times at which these potholes begin are the times at which the potholes are recorded by the city. This analysis should be interpretted as a model of how long the City of Boston takes to fill in potholes, not how long potholes last in the city themselves. The actual longevity of the potholes would require viewing every data point as left censored, something to explore in future analysis.

Code Appendix

```
# constant hazard: lambda(t) = lambda
ph.exp <- function (t, w, x) {
  gmod <- glm(w ~ x + offset(log(t)), family=poisson)
  mu <- fitted(gmod)
  lhood <- sum(w * log(mu) - mu + w * log(1 / t))
  list(alpha=1, gmod=gmod, lhood=lhood)
}

# accelerated failure: lambda(t) = lambda * alpha * t ^ (alpha - 1)
ph.weibull <- function (t, w, x, tolerance=1e-3) {
  alpha <- 1
  lhood <- Inf
  repeat {
```

```
    gmod <- glm(w ~ x + offset(alpha * log(t)), family=poisson)
    mu <- fitted(gmod)
    alpha <- sum(w) / sum((mu - w) * log(t))
    lhood.new <- sum(w * log(mu) - mu + w * log(alpha / t))
    if (abs(lhood.new - lhood) < tolerance) break
    lhood <- lhood.new
  }
  list(alpha=alpha, gmod=gmod, lhood=lhood)
}
library(survival)

holesource <- read.csv("~/Boston University/Spring 2016/MA 576/Final Project/wardsource.csv"
hist(log(holesource$date_diff), main="Histogram of Log Survival Time", xlab="log(t)")
hist(log(holesource$date_diff)[holesource$source=="Employee Generated"], main="Histogram for
hist(log(holesource$date_diff)[holesource$source=="Constituent Call"], main="Histogram for C
hist(log(holesource$date_diff)[holesource$source=="City Worker App"], main="Histogram for Ci

holesource$wardf <- as.factor(holesource$ward)

event <- ifelse(holesource$case_status=='Closed', 1, 0) # 'w' in our notes
sl <- Surv(holesource$date_diff, event=event) # survival object, to be used as response
sf <- survfit(sl ~ source, data=holesource) # Kaplan-Meier estimate of survival function
sf2 <- survfit(sl ~ wardf, data=holesource) # Kaplan-Meier estimate of survival function
plot(sf)
plot(sf2)
plot((sf$time), (-log(sf$surv)), main="Suvival Times with Source as a predictor", xlab="t",
plot((sf2$time), (-log(sf2$surv)), main="Suvival Times with Ward as a predictor", xlab="t",
plot(log(sf$time), log(-log(sf$surv)), main="Log Suvival Times with Source as a predictor",
plot(log(sf2$time), log(-log(sf2$surv)), main="Log Suvival Times with Ward as a predictor",

phe <- ph.exp(holesource$date_diff, event, holesource$source) # exp fit, initially
#phe2 <- ph.exp(holesource£date_diff, event, holesource£ward) # exp fit, initially
print(phe) # interpret coefficients
#print(phe2)

pchisq(113800, 5, lower=F)
coxph(sl ~ source, data=holesource)
```