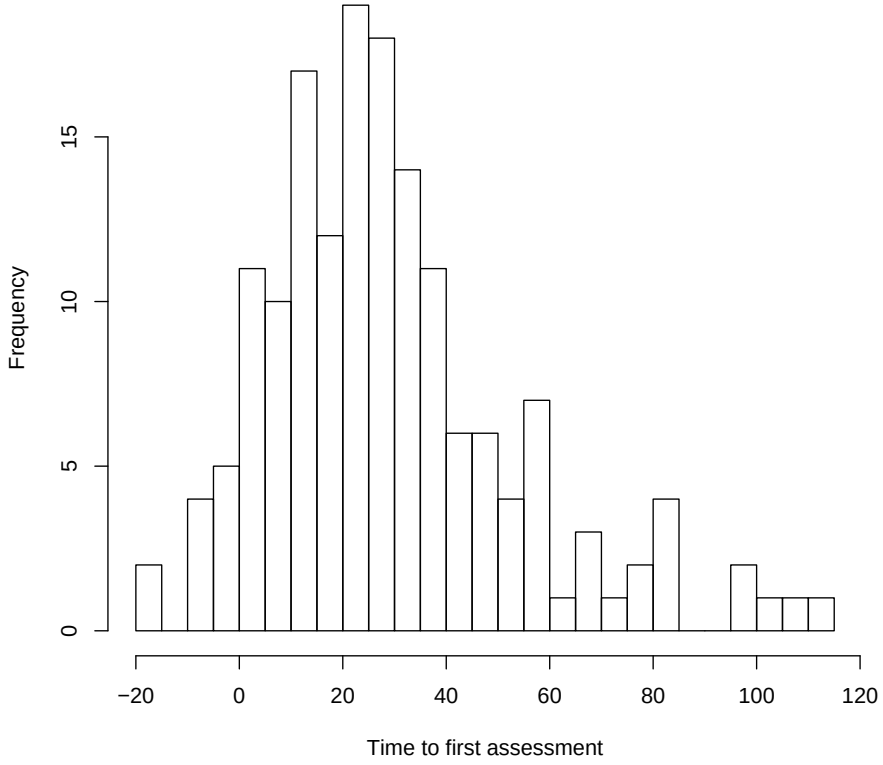
Analysis notes

March 25, 2014

V0.0 PN Mo March 24th
Data were crudely transformed to input, using max year for any timespans.
Subject to further discussion...

```
require(dplyr, quietly = T, warn.conflicts = F)

year_table <- tbl_df(read.csv("~/Work/Dropbox/First year of assessment/V2_as_input.csv"))


year_table <- year_table %.% mutate(time = YFA - Development)
```

```
hist(year_table$time, 20, main = "", xlab = "Time to first assessment")
```



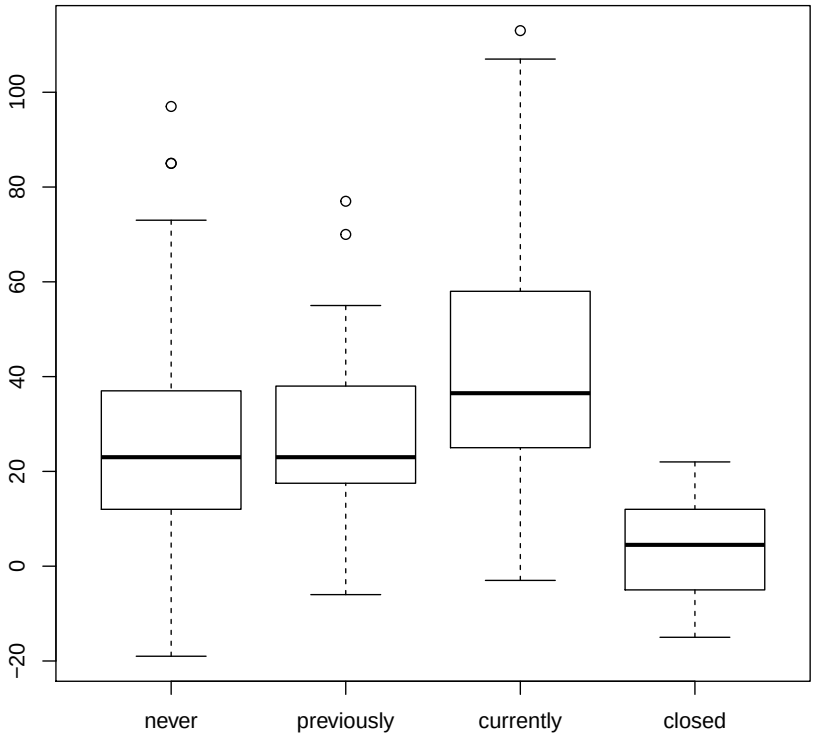
Time since *Development* doesn't look normally distributed, nor is it log-normal/gamma etc. The negative values make it quite different froma classical 'survival analysis' (time since diagnosis, for instance, is always positive).

Try a linear model to have a first look at potential patterns in the data: order rebuiling factor, not sure if that's necessary, but seems intuitive.

```
year_table$Rebuild <- ordered(year_table$Rebuild, levels = c("never", "previously",
    "currently", "closed"))

with(year_table, table(Rebuild))

## Rebuild
##      never previously  currently     closed
##        114         31         38          7

with(year_table, table(Region))
```

```
## Region
## Alaska   USEC   USNE   USSE   USWC
##     51      8     47     34     50

# set north-east as reference treatment
RC <- contr.treatment(levels(year_table$Region), base = 3)
```

```
with(year_table, boxplot(time ~ Rebuild))
```



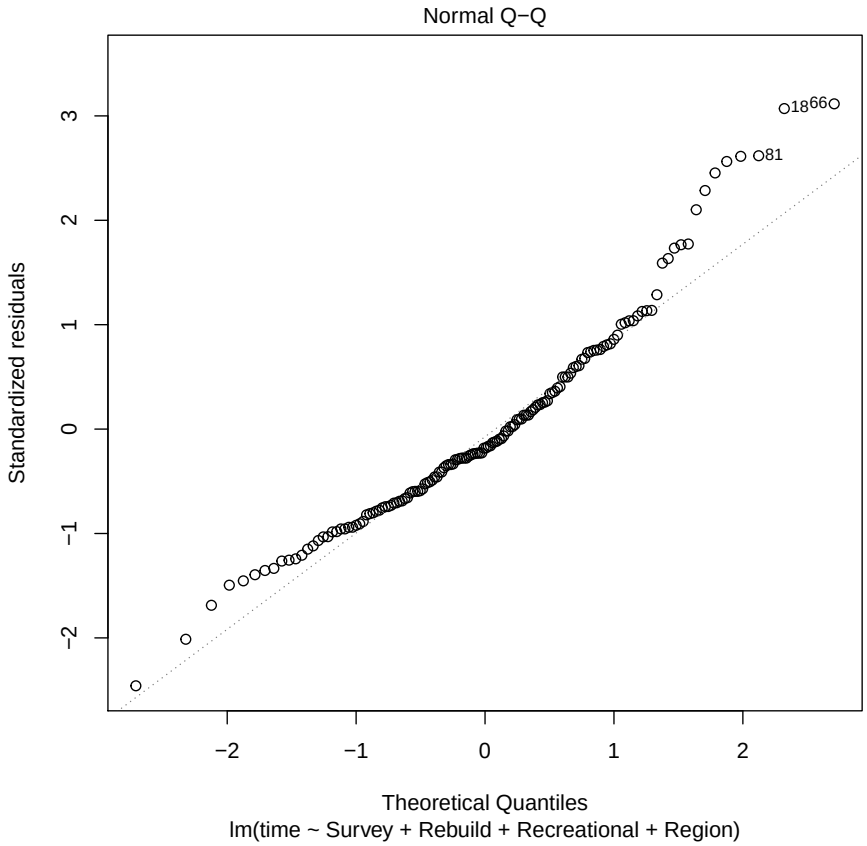
Try a simple normal linear model without any cencoring

```
ylm <- with(year_table, lm(time ~ Survey + Rebuild + Recreational + Region,
    contrasts = list(Region = RC)))

summary(ylm)
```

```
##
## Call:
## lm(formula = time ~ Survey + Rebuild + Recreational + Region,
##     contrasts = list(Region = RC))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -53.43 -15.13  -3.72  11.95  66.85
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   51.87486  434.44191    0.12   0.9051
## Survey        -0.00979    0.22085   -0.04   0.9647
## Rebuild.L     -8.80945    6.62829   -1.33   0.1860
## Rebuild.Q    -13.99104    6.04720   -2.31   0.0222 *
## Rebuild.C    -12.95189    4.38245   -2.96   0.0037 **
## Recreational   0.11786    0.30308    0.39   0.6980
## RegionAlaska -19.26538    6.14131   -3.14   0.0021 **
## RegionUSEC   -11.55973    9.29544   -1.24   0.2158
## RegionUSSE    -6.18902    6.68521   -0.93   0.3562
## RegionUSWC     0.11536    6.22994    0.02   0.9853
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.2 on 138 degrees of freedom
##   (42 observations deleted due to missingness)
## Multiple R-squared:  0.23,Adjusted R-squared:  0.179
## F-statistic: 4.57 on 9 and 138 DF,  p-value: 2.74e-05
```

The contrast in the status is with respect to the base case (never been in a rebuiling plan). Look at fit:

```
plot(ylm, w = 2)
```



The linear model fits ok, but explains only 17.9% of the total variance. Some initial questions:

- Is this the right response to look at?
- Is there a better strategy to give a single number to the development date? OR perhaps cencoring at that end of the distribution is appropriate as well.
- What could be additional factors: Size of the fishery in catch volume or revenue? Life-history/taxonomy?