Analysis_DielArrival

Anna Steel

October 21, 2016

Read in non-rediscretized data (just filtered, prior to splitting into bursts) and add release metadata

```
load("Maestros/AllFish_FiltSec3Hold.RData")
 dat = red6; rm(red6)

fishrel = read.csv("C:/Users/Anna/Documents/GitHub/Fremont16/Maestros/TagReleaseList.csv",
                   colClasses=c("RelTime"="character"))
  fishrel$datetime = as.POSIXct(paste(fishrel$RelDate, fishrel$RelTime), format="%Y-%m-%d %H%M", tz="
  names(fishrel)[3] <- "id"
  names(fishrel)[ncol(fishrel)] <- "datetime.Rel"

 dat = merge(dat, fishrel, all.x=T)
  dat$RelHr = as.POSIXlt(dat$datetime.Rel)$hour
```

Store sunset and sunrise times

- Referenced from: http://aa.usno.navy.mil (mean for range of first two releases: 2/22 - 3/8/2016)

```
sunrise = 06.63
sunset = 17.98
```

Store distance between Tisdale Weir release site and top of receiver array

- Calculated from Google Earth positions, referencing CFTC receivers at known locations / rkm
- The rkm are calculated with the golden gate at rkm 0, and Chipps Island at km 69.5

```
travdist_km = 55.1
```

Run several more data cleaning and organizing steps

Extract first and last detections for each fish

Calculate hour of day (decimal hours) for time when fish were released & when fish arrived at array; calculate transit time (a.k.a. 'delay') between

- Add code for night or day, using sunrise/sunset times incorporated above, to both release and arrival.
- Also calculate passage time - may be useful later

Calculate mean and median transit times

```
## [1] "mean = 40.3765043149376"
```

```
## [1] "sd = 27.0904710894926"
```

```
## [1] "median = 35.4965787826313"
```

```
##   RelEv  mean_hr     sd_hr median_hr
## 1     1 44.45239 37.514802  35.42802
## 2     2 36.26252  4.868016  36.47698
```

```
##   RelEv Rel.hrDayfac  mean_hr     sd_hr median_hr
## 1     1            6 43.57106 46.441281  34.98257
## 2     1           11 43.44061 46.753894  34.37964
## 3     1           17 49.19190 34.780700  40.32683
## 4     1           23 41.87227 12.894312  37.07309
## 5     2            3 40.41726  2.567019  39.91390
## 6     2            9 38.44610  2.976854  37.70178
## 7     2           15 35.63652  5.233195  33.69668
## 8     2           21 30.93178  1.061992  30.73722
```

Removed one fish with passage time of 13+ days; next longest was 3 hours (183 min)

```
fl.df3 = fl.df2[fl.df2$Id != 36472,]
```

Mean and median transit times WITHOUT outlier

- outlier in Release Event 1, at 23:00
- very similar to values calculated with outlier

```
## [1] "mean = 40.3719695063276"
```

```
## [1] "sd = 27.1219369733865"
```

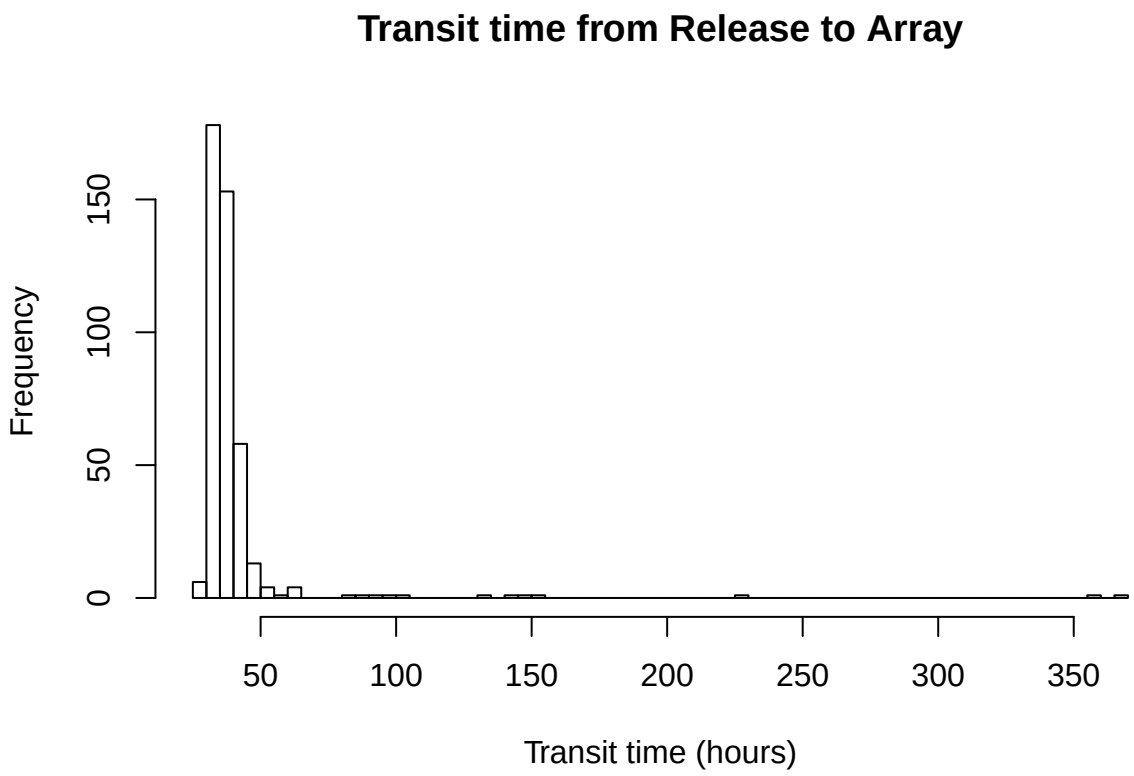
```
## [1] "median = 35.4812222011222"
```

```
##   RelEv  mean_hr     sd_hr media_hrn
## 1     1 44.46230 37.602068  35.41951
## 2     2 36.26252  4.868016  36.47698
```

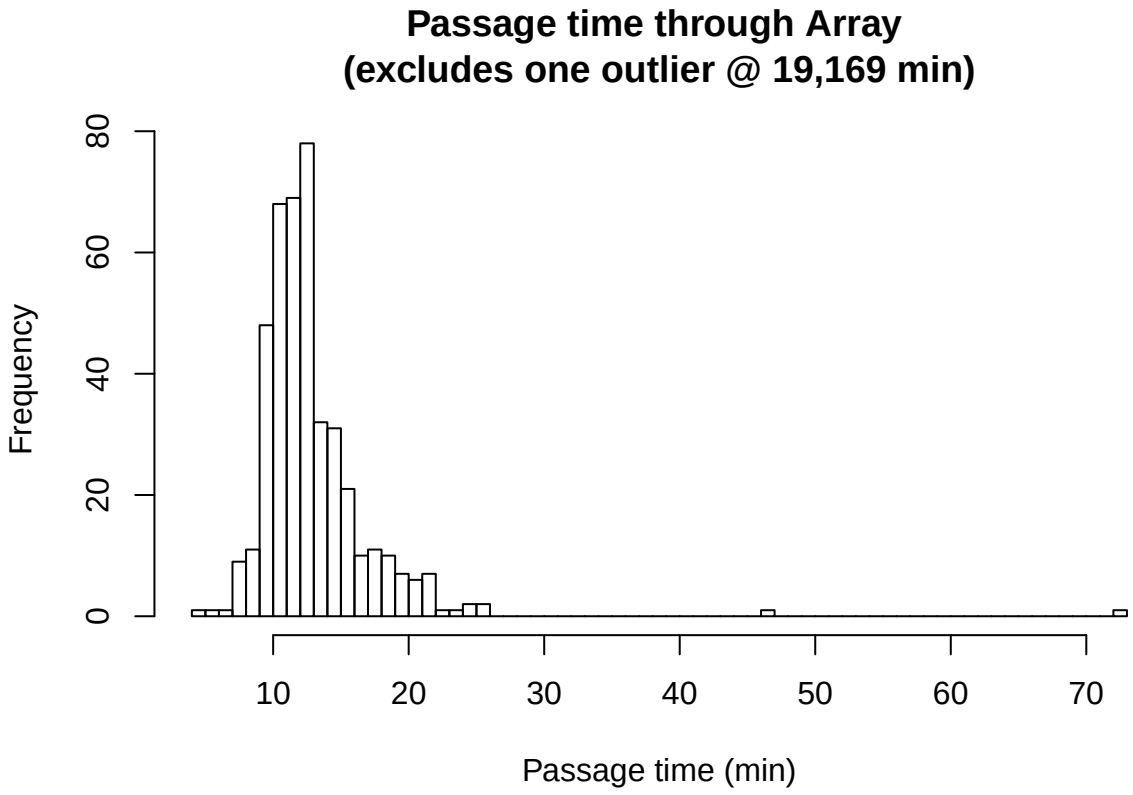
```
##   RelEv Rel.hrDayfac  mean_hr     sd_hr media_hrn
## 1     1            6 43.57106 46.441281  34.98257
## 2     1           11 43.44061 46.753894  34.37964
## 3     1           17 49.19190 34.780700  40.32683
## 4     1           23 41.86409 13.013009  37.04043
## 5     2            3 40.41726  2.567019  39.91390
## 6     2            9 38.44610  2.976854  37.70178
## 7     2           15 35.63652  5.233195  33.69668
## 8     2           21 30.93178  1.061992  30.73722
```

Plots to visualize initial transit time and passage time

```
hist(fl.df3$Delay.hr,
     main="Transit time from Release to Array",
     xlab="Transit time (hours)", ylab="Frequency", breaks=50)
```

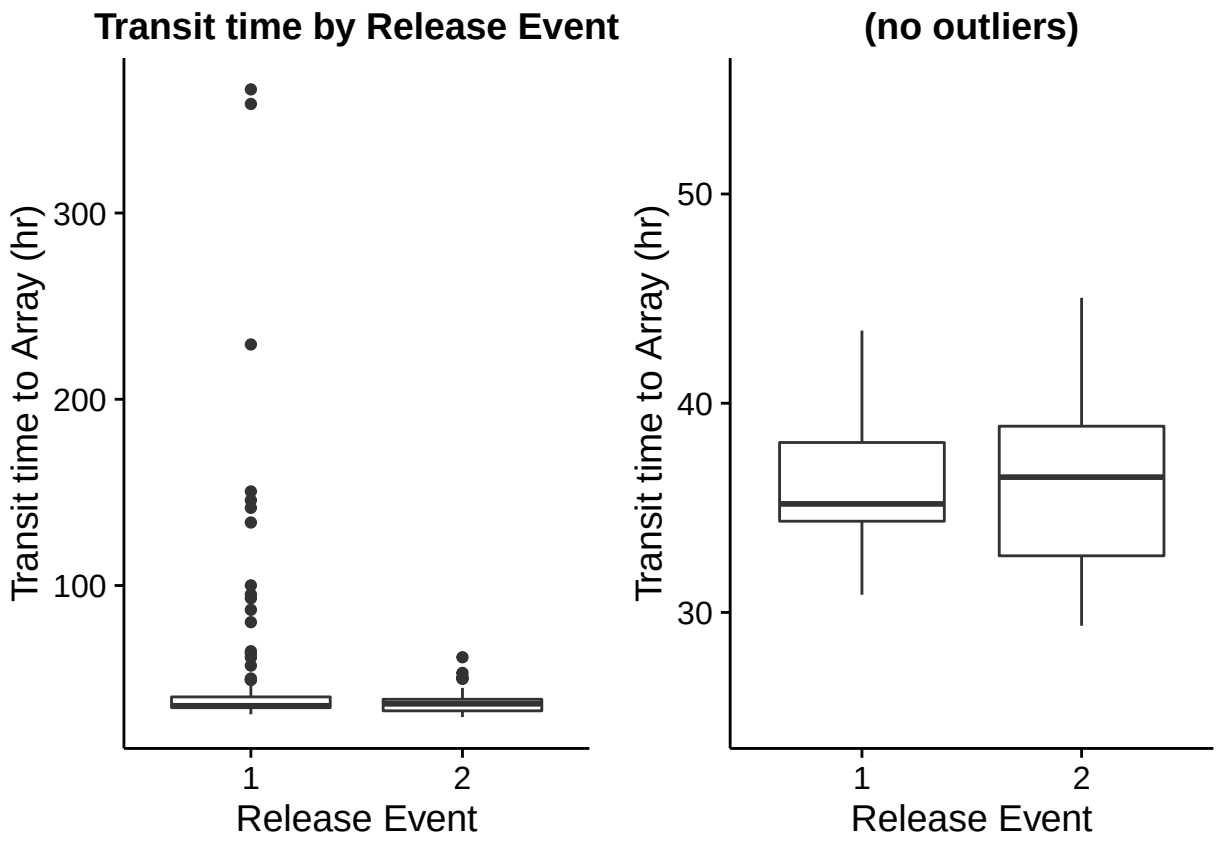


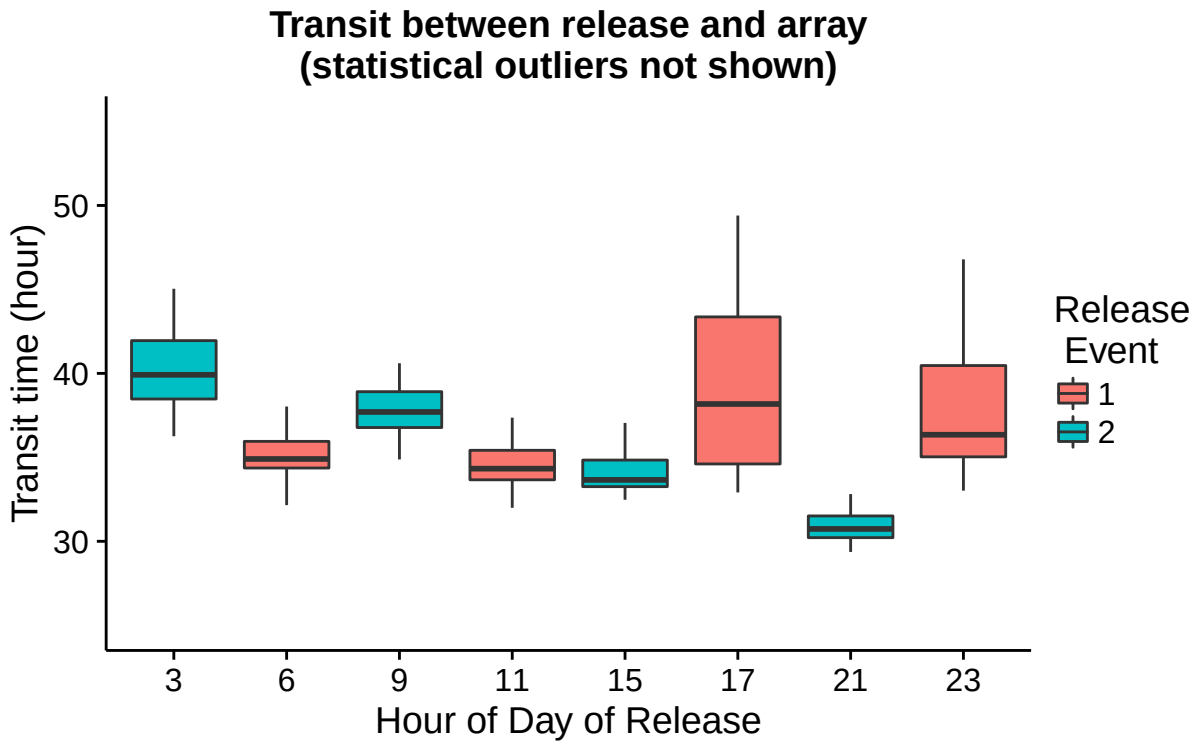

```
hist(fl.df3$passtime.min,
     main="Passage time through Array \n(excludes one outlier @ 19,169 min)",
     xlab="Passage time (min)",ylab="Frequency", breaks=50)
```

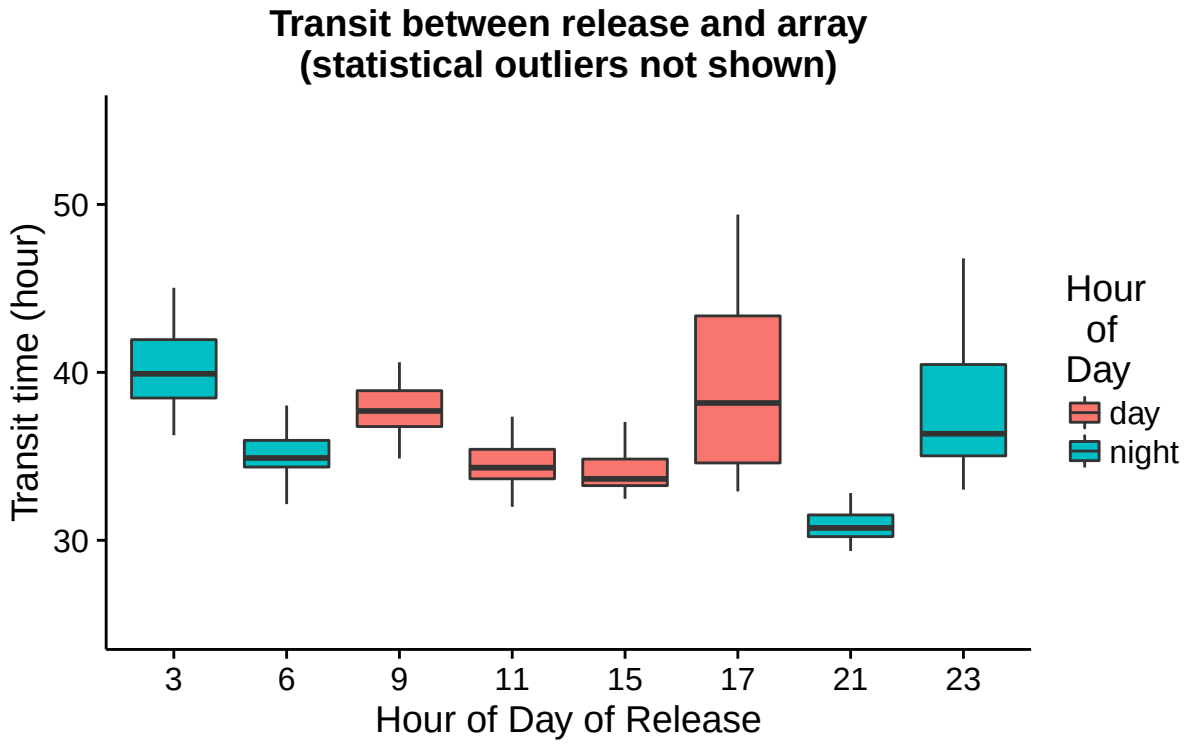


Differences by Release Event or Release Hour

- note: it would be nice to annotate boxes with respective sample sizes


Transit time by Release Event
Transit time to Array (hr)
300
200
100
1
2
Release Event
(no outliers)
Transit time to Array (hr)
50
40
30
1
2
Release Event





- note: consider creating this second set of boxplots in conjunction with a hydrograph to illustrate relationship between transit time and stage

Mean and Median of arrival times, overall and by release time groups

- requires circular statistics; use packages psych (mean/sd) and circular (median)

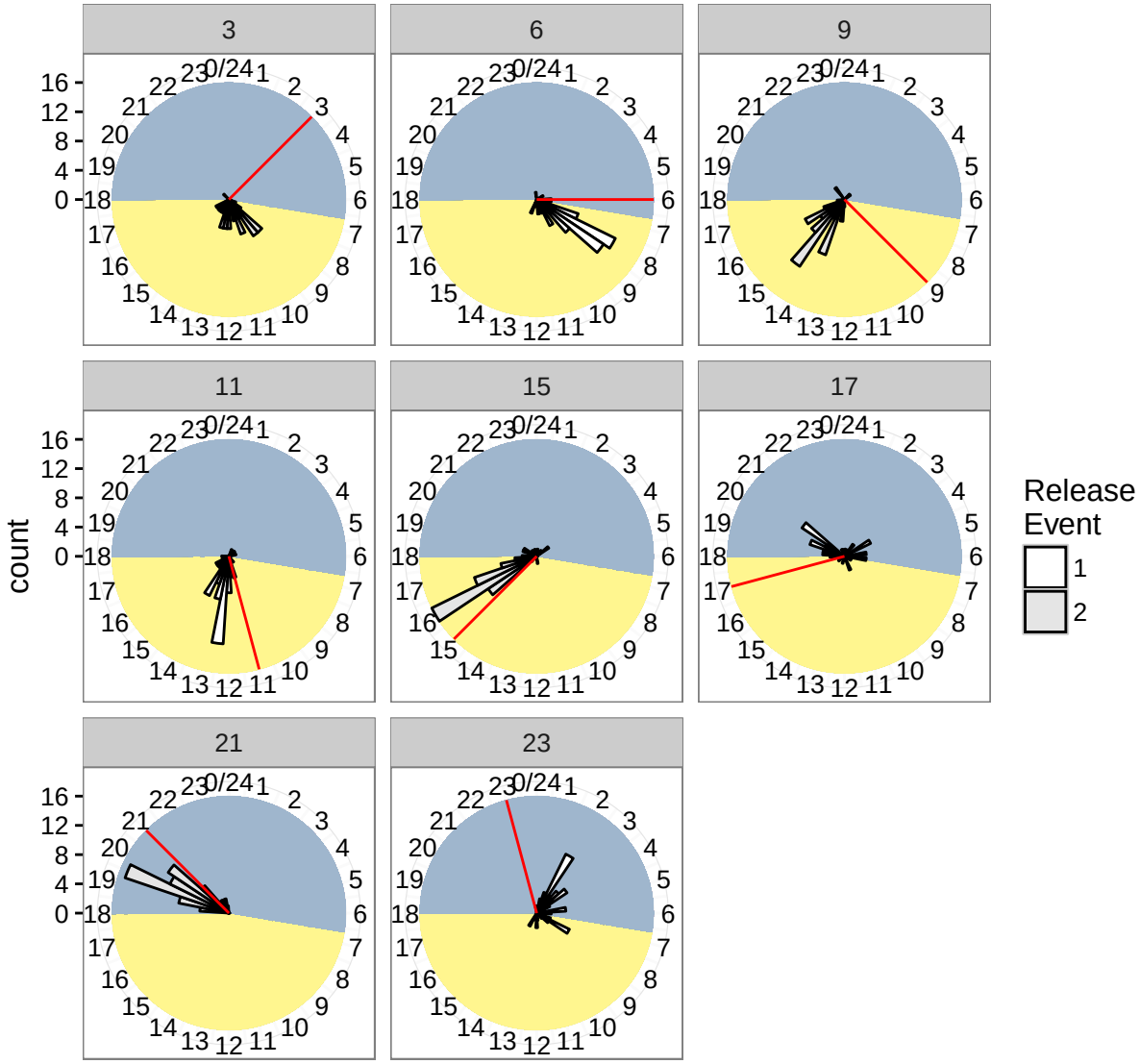
```
## [1] "mean = 13.8087646327308"
```

```
## [1] "sd = 1.61934078575087"
```

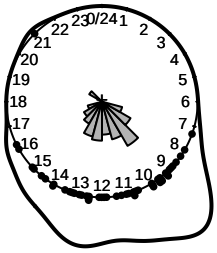
```
## [1] "median = 13.5"
```

```
##   Rel.hrDayfac median.F.hr mean.F.hr   sd.F.hr
## 1            3    11.00000 11.349139 0.6415828
## 2            6     8.40000  8.645294 0.5289553
## 3            9    14.50000 14.719170 0.6036528
## 4           11    12.72000 13.060375 0.6941681
## 5           15    16.30000 16.791960 0.6984048
## 6           17    20.63333 23.206370 1.5831653
## 7           21    19.80000 20.012994 0.2794529
## 8           23     3.70000  4.437168 0.9361980
```

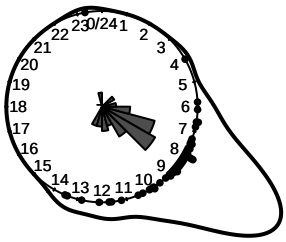
Circular plots: same data, two visualizations



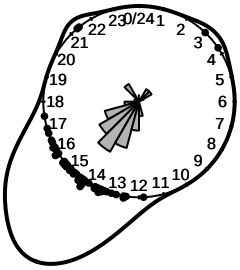
Release Event 2
Release Hour 3:00



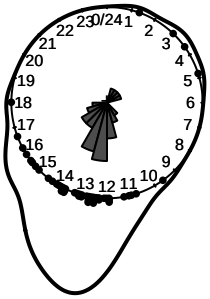
Release Event 1
Release Hour 6:00



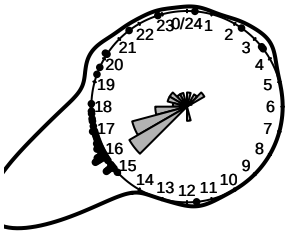
Release Event 2
Release Hour 9:00



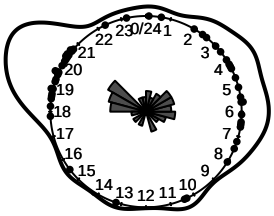
Release Event 1
Release Hour 11:00



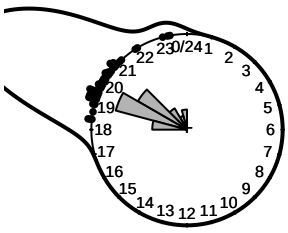
Release Event 2
Release Hour 15:00



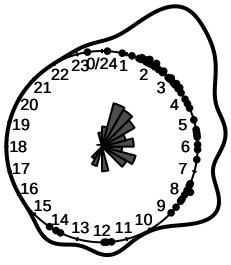
Release Event 1
Release Hour 17:00



Release Event 2
Release Hour 21:00



Release Event 1
Release Hour 23:00



- the blobs around the circle are kernel density lines, but the smoothing parameter is simply the default; if these graphics are going to be used for anything other than general exploration of the data I should revisit the smoothing parameter selection process.

Preliminary Exploration of circular statistics for diel questions

The following are derived from the test statistics I ran for 2015 to compare runs

- Much/all of the following was coded with guidance from the text book "Circular Statitics in R' by Arthur Pewsey et al (2013)

To select an appropriate statistical distrubition for the circular data we want to know if the data are symetrical (we know they are not uniform from looking at the plots, so won't bother to test this statistically, for now). If we do not reject symmetry, we may use the Jones-Pewsey or vonMises distributions, but if we do reject symmetry we may need to use the more flexible Batschelet distribution.

Test for 'reflective symmetry'

We can use he test proposed by Pewsey (2002) which is suitable for sapmle sizes of 50 or more (ours are n=51 - 56 in each release hour)

```
##   Relhrfac  teststat       pval
## 1        3 1.8235068 0.06822667
## 2        6 0.3442692 0.73064381
## 3        9 0.9718112 0.33114448
## 4       11 1.1539002 0.24854109
## 5       15 2.3837405 0.01713768
## 6       17 1.0828236 0.27888673
## 7       21 2.2563366 0.02404956
## 8       23 3.1894962 0.00142521
```

- NOTE: this uses template=clock24 and rotation=clock which may pose a problem; The functions may be expecting radians measured *counter-clockwise* from zero (in mathematic terms, so zero = *positive X-axis*). Here I use radians measured *clockwise* from the *top of the unit circle*. Before moving along or using these values, clarify this.
- Aside from that concern, the results are mixed -> releases at 15:00 (rel2), 21:00 (rel2) and 23:00 (re11) are not statistically symmetrical, but the release at 17:00 (rel1) is, dispite not resembling a normal distribution but rather being bimodal. Interesting. Not sure if any of this will be used in a report, so I'm not pursuing it at this moment.

Is the delay in arrival time related to release time?

I tried to use the guidance in Pewsey 2013 textbook to fit a cosine regression model, but the data on delay time are too skewed for it to fit the assumptions. Additionally, I'm not sure it's clear what the model will tell you because the release time is split into two predictor variables - in this case neither are significant, and the model doesn't particularly look nice.

Perhaps this indicates that there is NOT a significant effect of release time on travel time - ie: there isn't a strong or clear diel effect.

Regardless, I'm also not sure if time sunk into this exercise is valuable, so it will be put on the back burner for now.

Basic cosine model:

```r
# calculate a circular correlation as first pass at this relationship
circadian.linear.cor(fl.df3$Delay.hr, fl.df3$Rel.hrDay)
```

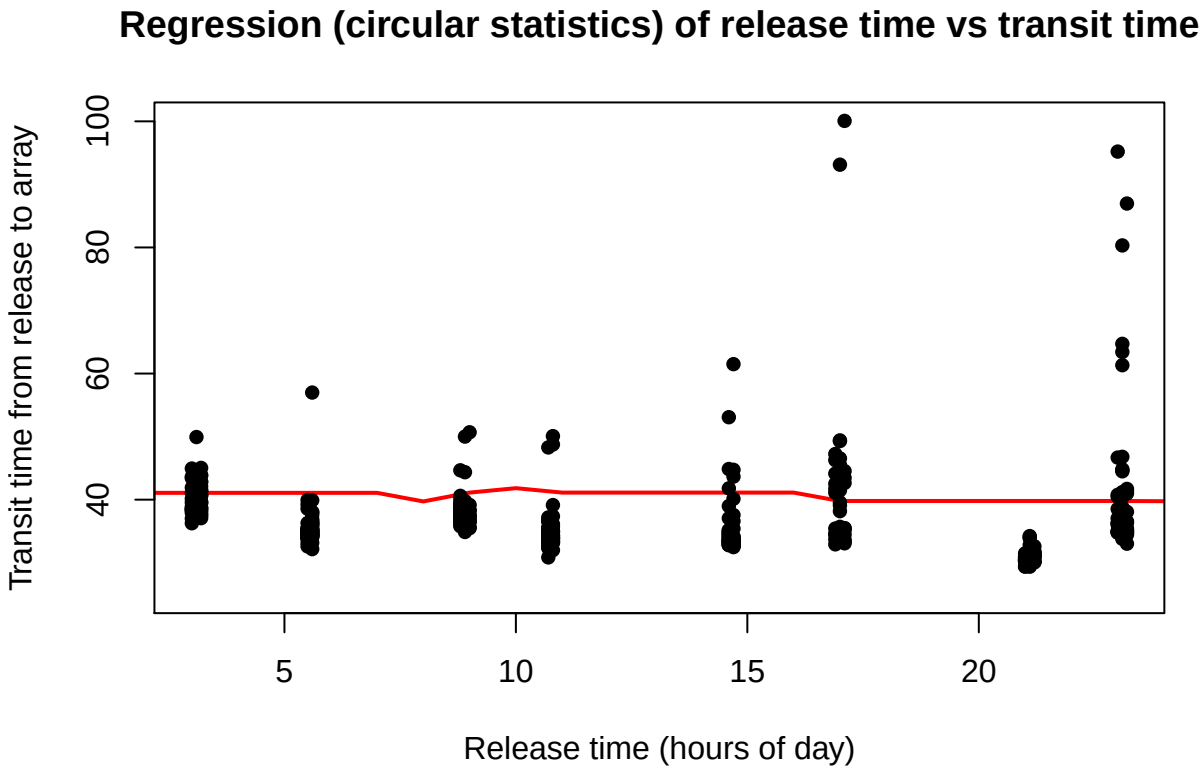
```
## [1] 0.398826
```

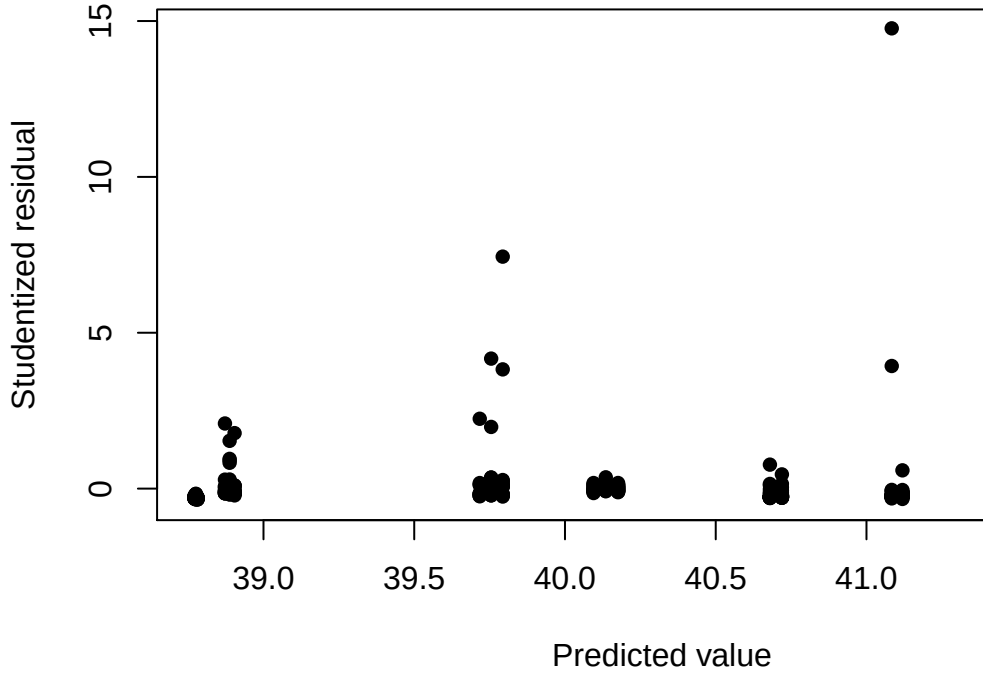
```r
# Next step: regression of a linear response on a circular predictor
# basic cosine model: x = a + b1*cos(2pi/24*Rel.hr*Day) + b2*sin(2pi/12*Rel.hrDay) + e
omega = 2*pi/24
cosvar = cos(omega*fl.df3$Rel.hrDay)
sinvar = sin(omega*fl.df3$Rel.hrDay)

delaymod = lm(fl.df3$Delay.hr ~ cosvar + sinvar)
summary(delaymod)
```

```
##
## Call:
## lm(formula = fl.df3$Delay.hr ~ cosvar + sinvar)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10.98  -7.09  -4.84  -1.23 325.32
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  40.3270     1.3127  30.721   <2e-16 ***
## cosvar       -1.2572     1.8634  -0.675    0.500
## sinvar        0.9289     1.8518   0.502    0.616
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.16 on 426 degrees of freedom
## Multiple R-squared:  0.001734,   Adjusted R-squared:  -0.002953
## F-statistic:  0.37 on 2 and 426 DF,  p-value: 0.691
```

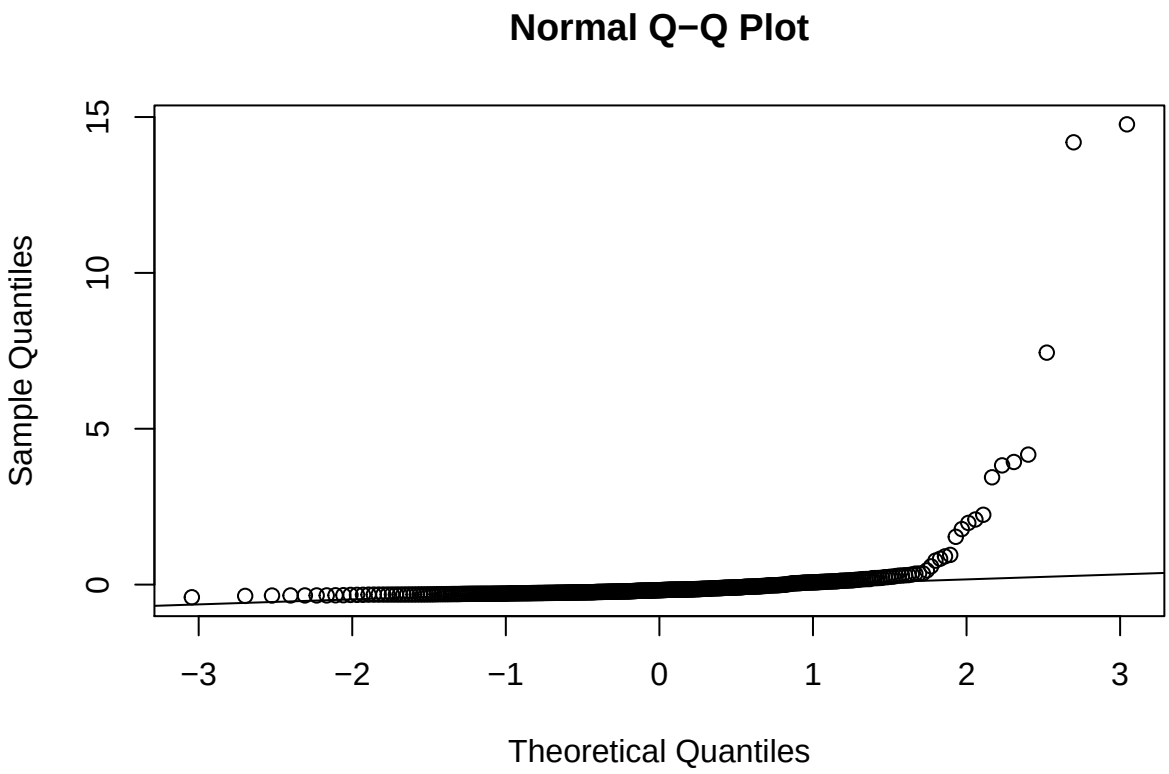
```r
plot(fl.df3$Rel.hrDay, fl.df3$Delay.hr,
     xlab="Release time (hours of day)",
     ylab="Transit time from release to array",
     main="Regression (circular statistics) of release time vs transit time",
     pch=16, ylim=c(25,100),
     lines(predict(delaymod), lwd=2, col="red") )
```


Regression (circular statistics) of release time vs transit time
Transit time from release to array
40
60
80
100
5
10
15
20
Release time (hours of day)



Diagnostics of basic cosine model:

```
##
##  Shapiro-Wilk normality test
##
## data:  delayresid
## W = 0.20786, p-value < 2.2e-16
```




```
##
##  Bartlett test of homogeneity of variances
##
## data:  delayresid and fl.df3$Rel.hrDay
## Bartlett's K-squared = 1263.4, df = 20, p-value < 2.2e-16
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  delayresid and fl.df3$Rel.hrDay
## Fligner-Killeen:med chi-squared = 96.729, df = 20, p-value =
## 4.827e-12
```

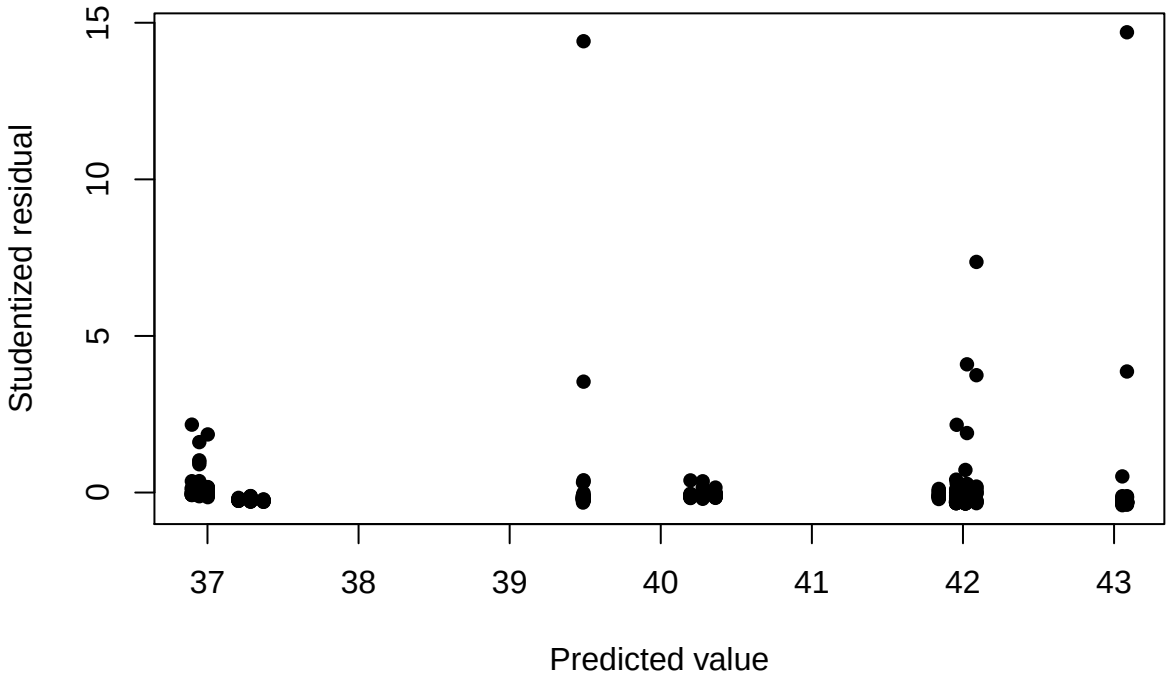
Doesn't meet assumptions, due to outliers with long delay times. But is it close enough?

Extended cosine model (additional sin & cos parameters):

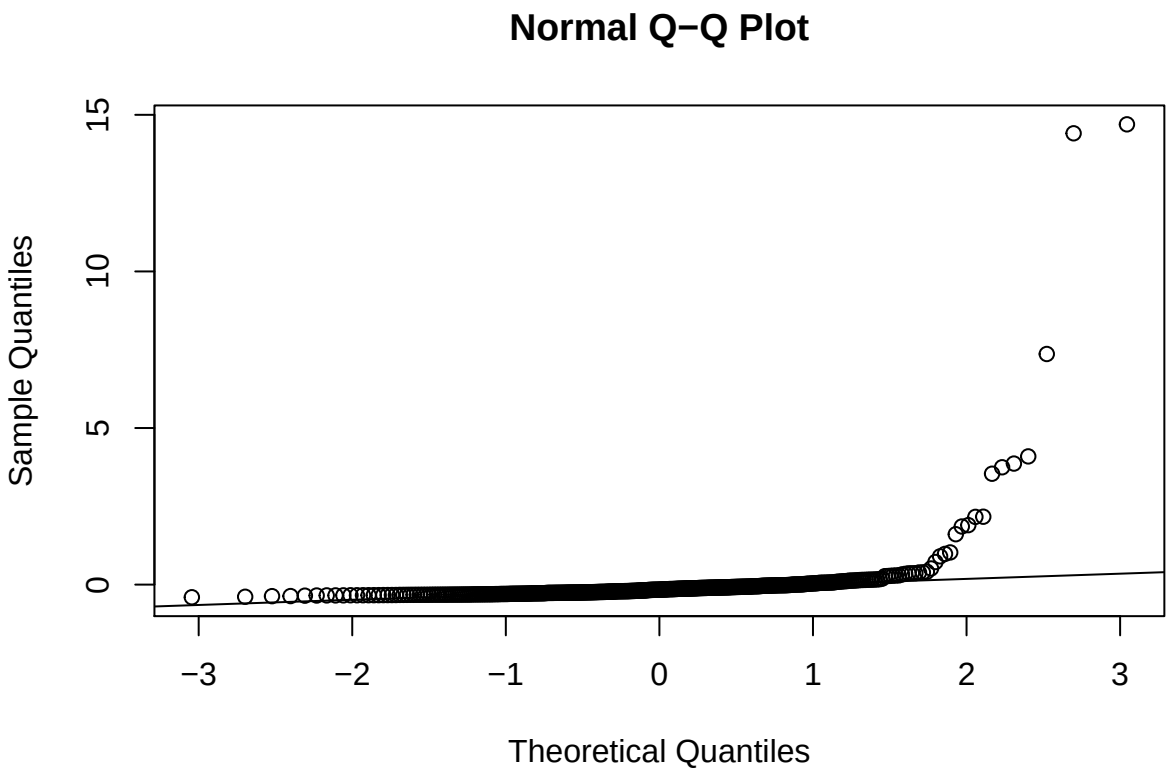
```
##
## Call:
## lm(formula = fl.df3$Delay.hr ~ cosvar + sinvar + cos2var + sin2var)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10.90  -7.11  -4.38  -1.04 323.32
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  40.3958     1.3156  30.706   <2e-16 ***
## cosvar       -1.1179     1.8659  -0.599    0.549
## sinvar        0.8803     1.8753   0.469    0.639
## cos2var      -1.5999     2.2506  -0.711    0.478
## sin2var       1.6122     1.8261   0.883    0.378
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.16 on 424 degrees of freedom
## Multiple R-squared:  0.006461,   Adjusted R-squared:  -0.002912
## F-statistic: 0.6893 on 4 and 424 DF,  p-value: 0.5997
```

Diagnostics of extended cosine model:




```
##
##  Shapiro-Wilk normality test
##
## data:  delay2resid
## W = 0.20857, p-value < 2.2e-16
```




```
##
##  Bartlett test of homogeneity of variances
##
## data:  delay2resid and fl.df3$Rel.hrDay
## Bartlett's K-squared = 1264.6, df = 20, p-value < 2.2e-16
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  delay2resid and fl.df3$Rel.hrDay
## Fligner-Killeen:med chi-squared = 96.776, df = 20, p-value =
## 4.734e-12
```

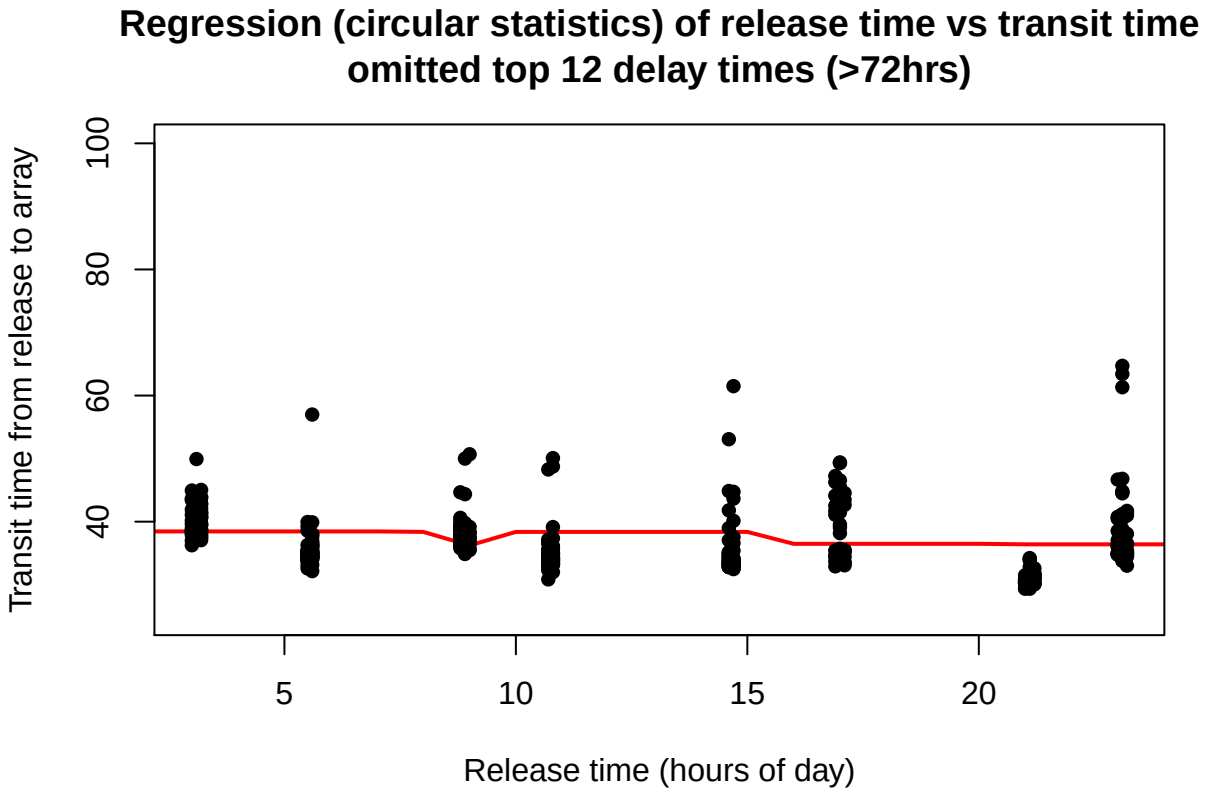
Still doesn't meet assumptions well. Could be due to outliers?

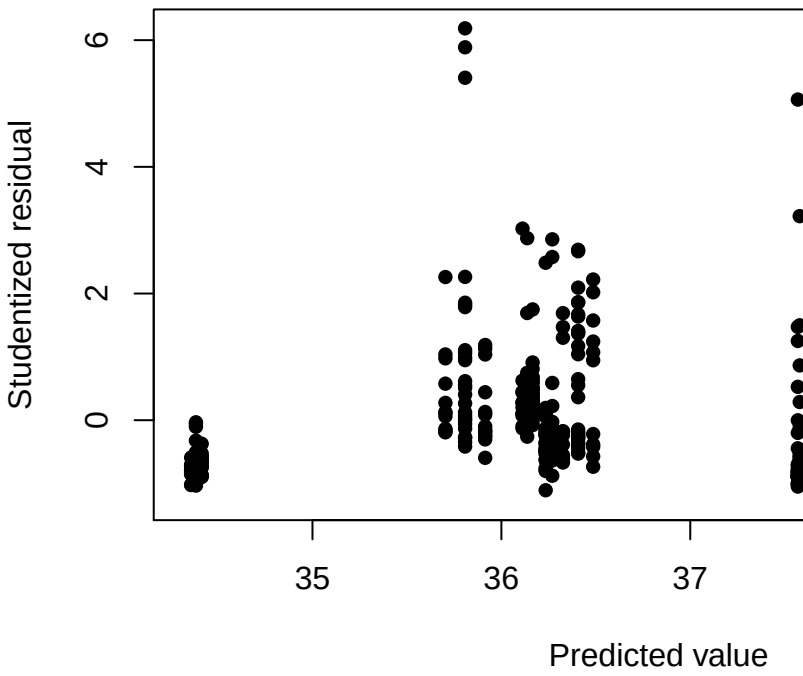
Extended cosine model without outliers

```
##
## Call:
## lm(formula = fl.df4$Delay.hr ~ cosvar + sinvar + cos2var + sin2var)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.666 -3.240 -1.376  1.741 28.030
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.8285     0.2394 153.820  < 2e-16 ***
## cosvar        0.2998     0.3386   0.885   0.3765
## sinvar        1.3599     0.3427   3.968 8.55e-05 ***
## cos2var       0.7858     0.4138   1.899   0.0583 .
## sin2var       1.8134     0.3276   5.535 5.53e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.872 on 412 degrees of freedom
## Multiple R-squared:  0.09389,    Adjusted R-squared:  0.08509
## F-statistic: 10.67 on 4 and 412 DF,  p-value: 3.073e-08
```

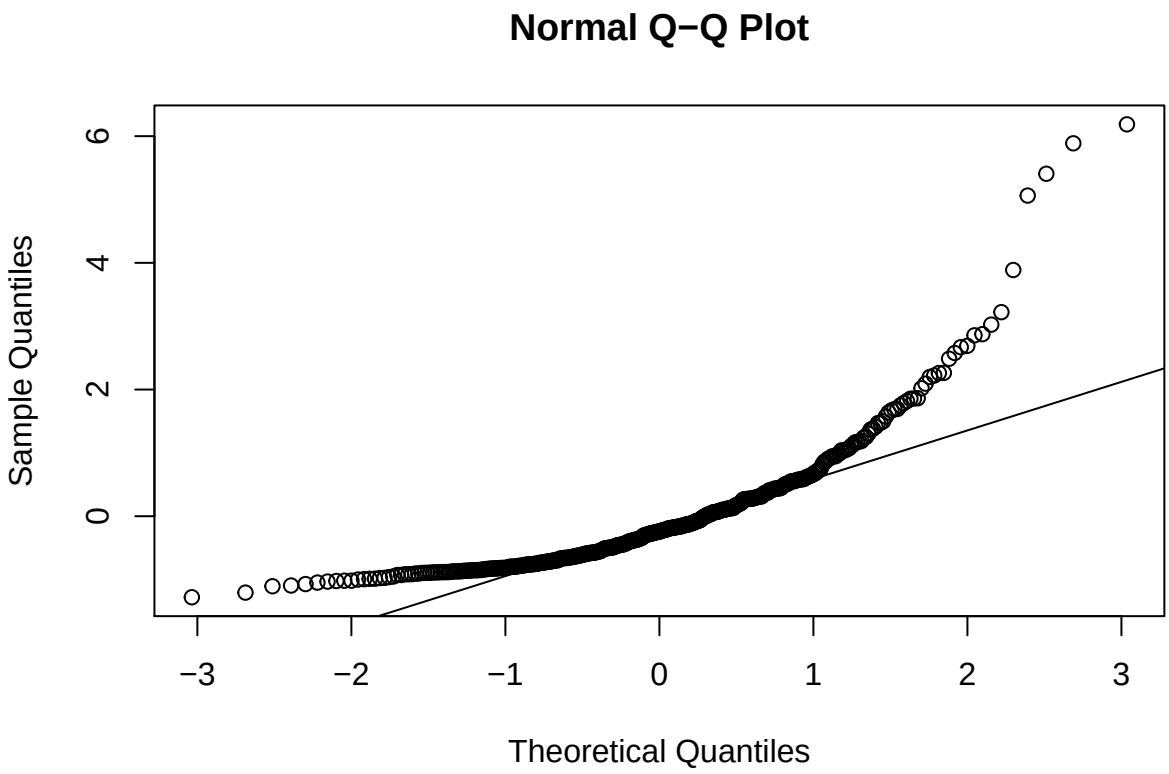
```
##
## Call:
## lm(formula = fl.df4$Delay.hr ~ sinvar + sin2var)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.222 -3.389 -1.185  1.648 28.902
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.8148     0.2400 153.376  < 2e-16 ***
## sinvar        1.2406     0.3389   3.661 0.000284 ***
## sin2var       1.5800     0.3068   5.150 4.04e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.886 on 414 degrees of freedom
## Multiple R-squared:  0.08411,    Adjusted R-squared:  0.07969
## F-statistic: 19.01 on 2 and 414 DF,  p-value: 1.262e-08
```


Regression (circular statistics) of release time vs transit time
omitted top 12 delay times (>72hrs)
Transit time from release to array
40
60
80
100
5
10
15
20
Release time (hours of day)



Diagnostics of extended cosine model, no outliers:

```
##
##  Shapiro-Wilk normality test
##
## data:  delay4resid
## W = 0.78791, p-value < 2.2e-16
```




```
##
##  Bartlett test of homogeneity of variances
##
## data:  delay4resid and fl.df4$Rel.hrDay
## Bartlett's K-squared = 245.79, df = 20, p-value < 2.2e-16
```

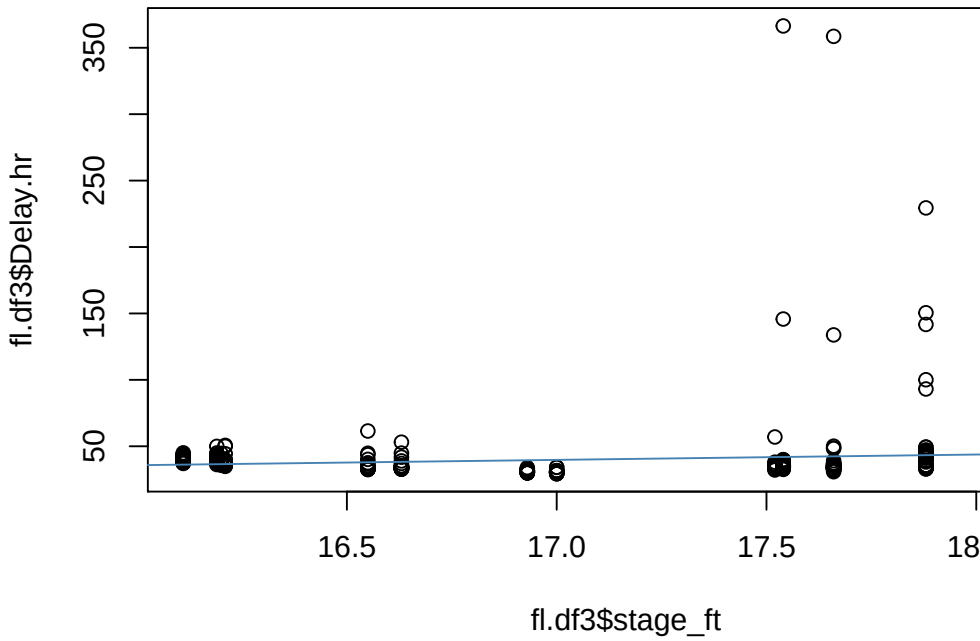
```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  delay4resid and fl.df4$Rel.hrDay
## Fligner-Killeen:med chi-squared = 89.185, df = 20, p-value =
## 1.029e-10
```

STILL doesn't meet assumptions well. =/

So, if we can't use the linear modeling approach, what CAN we use to answer this question? Moving on for now.

Is there a relationshp between transit time and river stage?

- this code tries to use linear regression, but both the delay times and the stage measurements are dreadfully non-normal. Need to find another test, if we'd like to use a statistical test. Again, leaving this incomplete



until I know if it will be of interest.

```
##
## Call:
## lm(formula = fl.df3$Delay.hr ~ fl.df3$stage_ft)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11.57  -8.20  -5.03   1.24 324.48
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -28.676     30.824  -0.930   0.3527
## fl.df3$stage_ft    4.025      1.795   2.242   0.0255 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27 on 427 degrees of freedom
## Multiple R-squared:  0.01164,    Adjusted R-squared:  0.009321
## F-statistic: 5.027 on 1 and 427 DF,  p-value: 0.02547
```

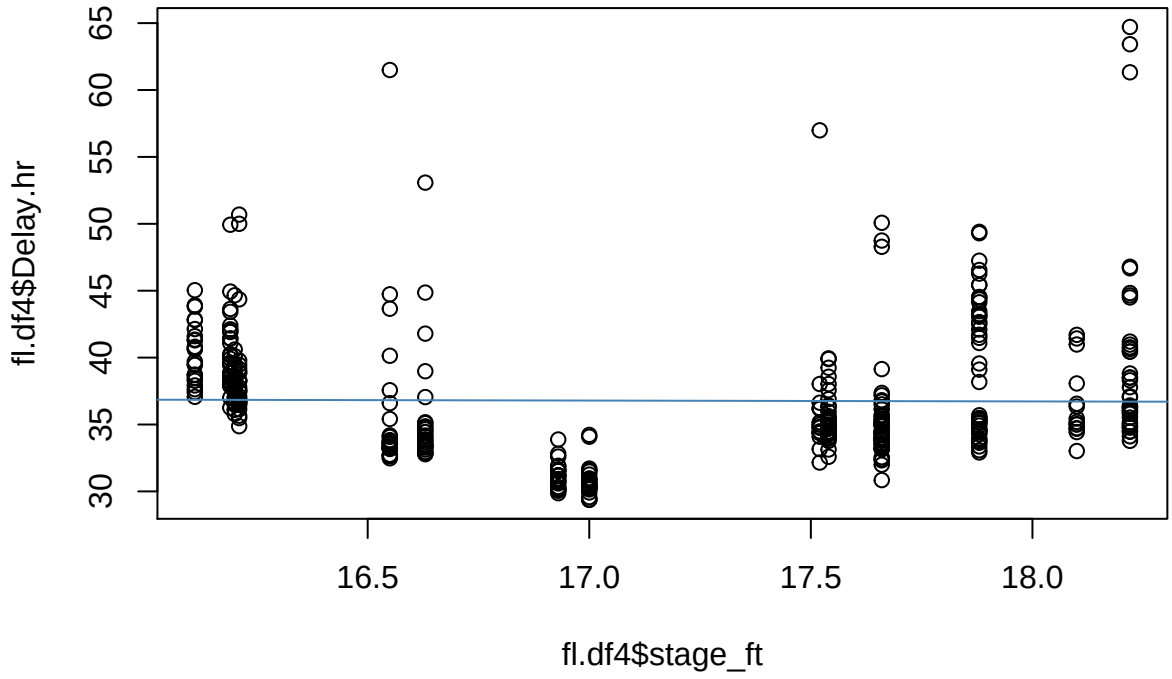
At higher stages fish actually took longer. Is this driven by the heavy outliers?

- Use the reduced dataset from above without the longest 12 travel times (>72)

```
fl.df4 = fl.df3[fl.df3$Delay.hr<72,]

plot(fl.df4$Delay.hr ~ fl.df4$stage_ft)
```

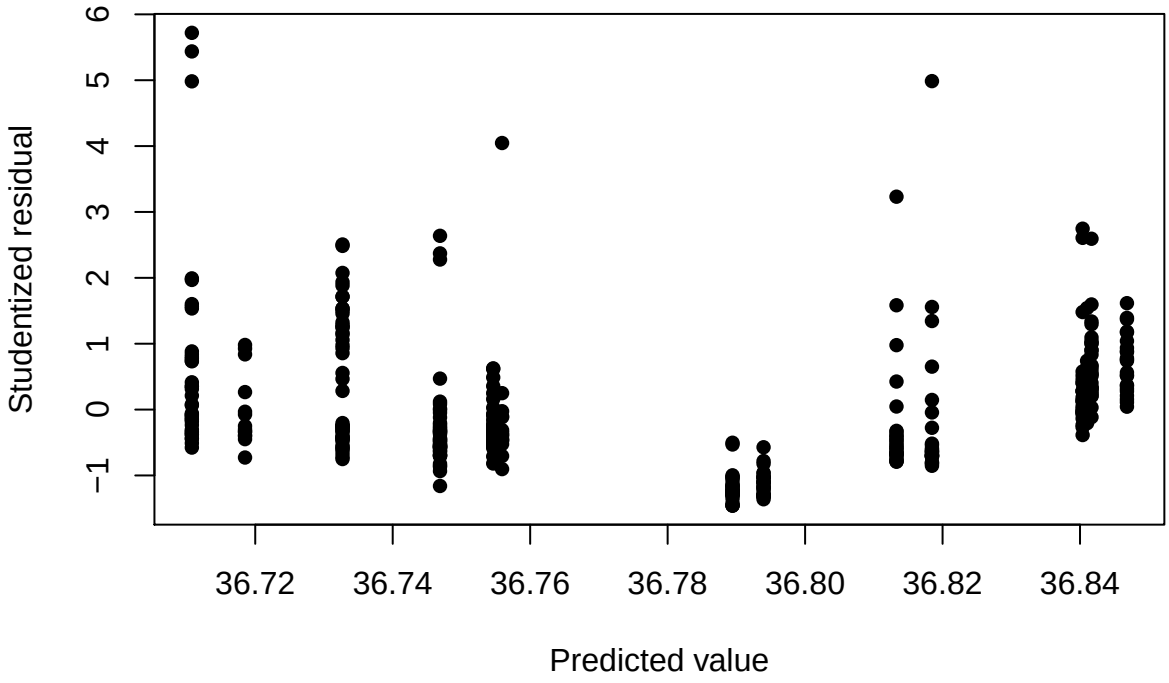
```r
stgmod = lm(fl.df4$Delay.hr ~ fl.df4$stage_ft)
abline(stgmod, col="steelblue")
```



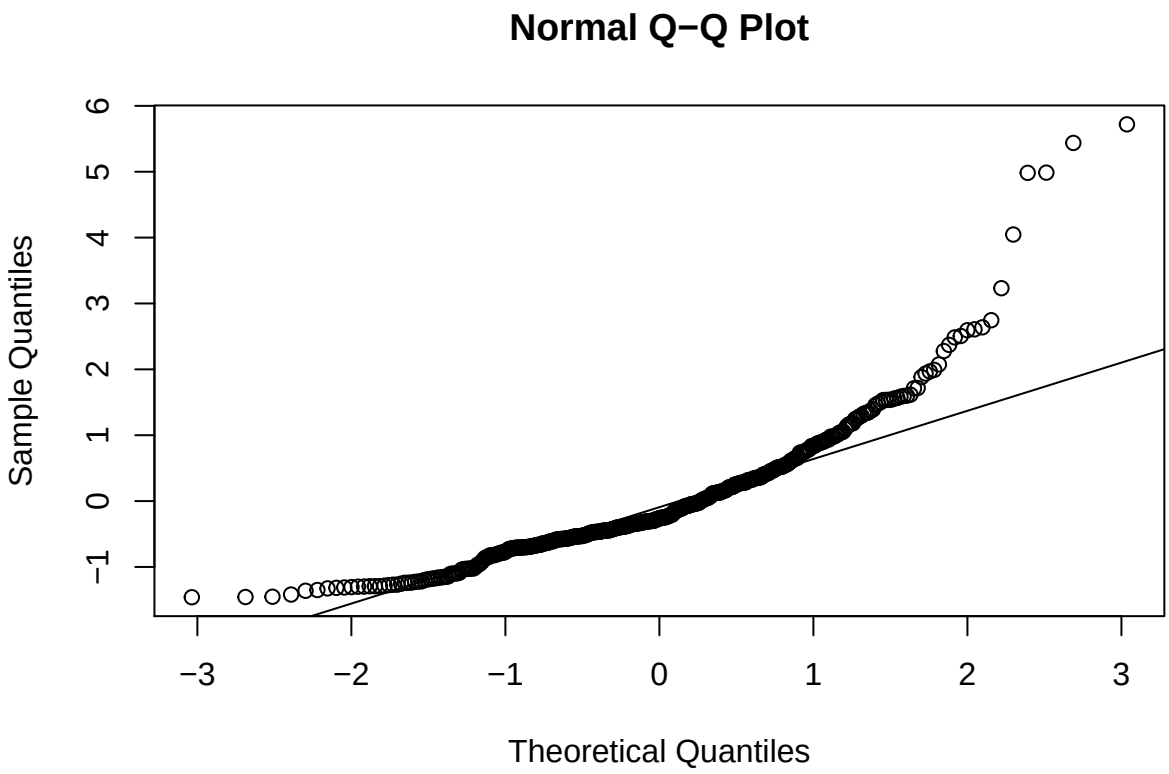

```r
summary(stgmod)
```

```
##
## Call:
## lm(formula = fl.df4$Delay.hr ~ fl.df4$stage_ft)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.424 -2.985 -1.318  2.040 27.999
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     37.88560    5.90480   6.416 3.82e-10 ***
## fl.df4$stage_ft -0.06448    0.34434  -0.187    0.852
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.099 on 415 degrees of freedom
## Multiple R-squared:  8.449e-05,  Adjusted R-squared:  -0.002325
## F-statistic: 0.03506 on 1 and 415 DF,  p-value: 0.8516
```

Diagnostics




```
##
##  Shapiro-Wilk normality test
##
## data:  stgresid
## W = 0.85557, p-value < 2.2e-16
```



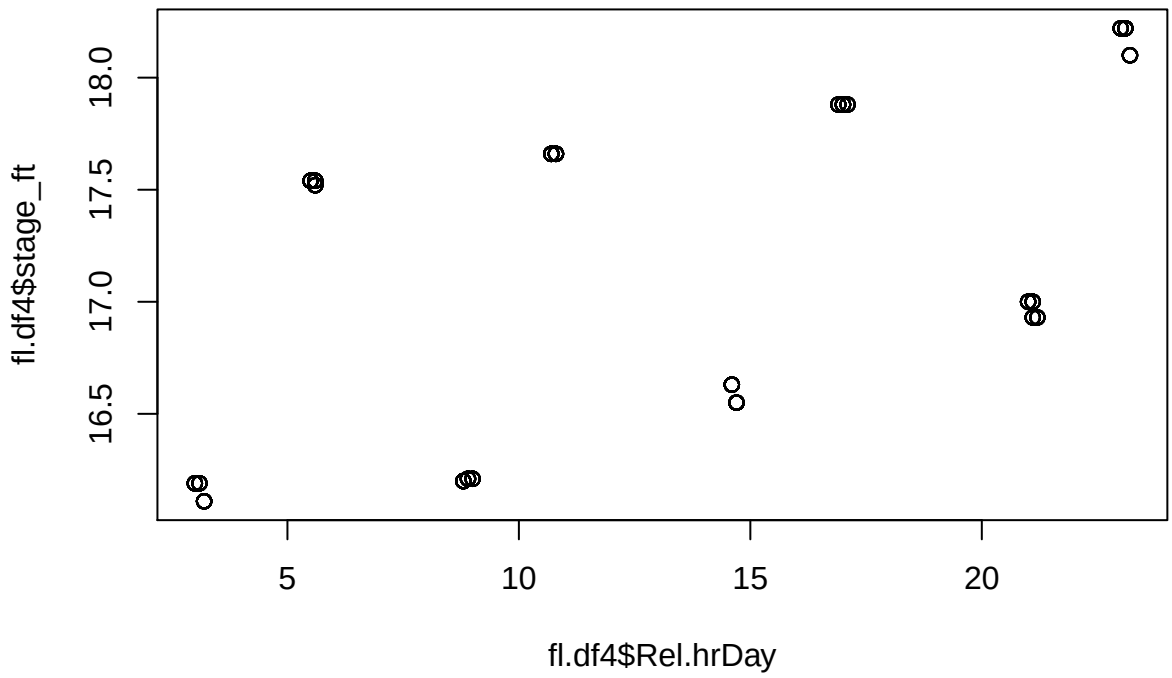

```
##
##  Bartlett test of homogeneity of variances
##
## data:  stgresid and fl.df4$stage_ft
## Bartlett's K-squared = 241.44, df = 13, p-value < 2.2e-16
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  stgresid and fl.df4$stage_ft
## Fligner-Killeen:med chi-squared = 109.65, df = 13, p-value <
## 2.2e-16
```

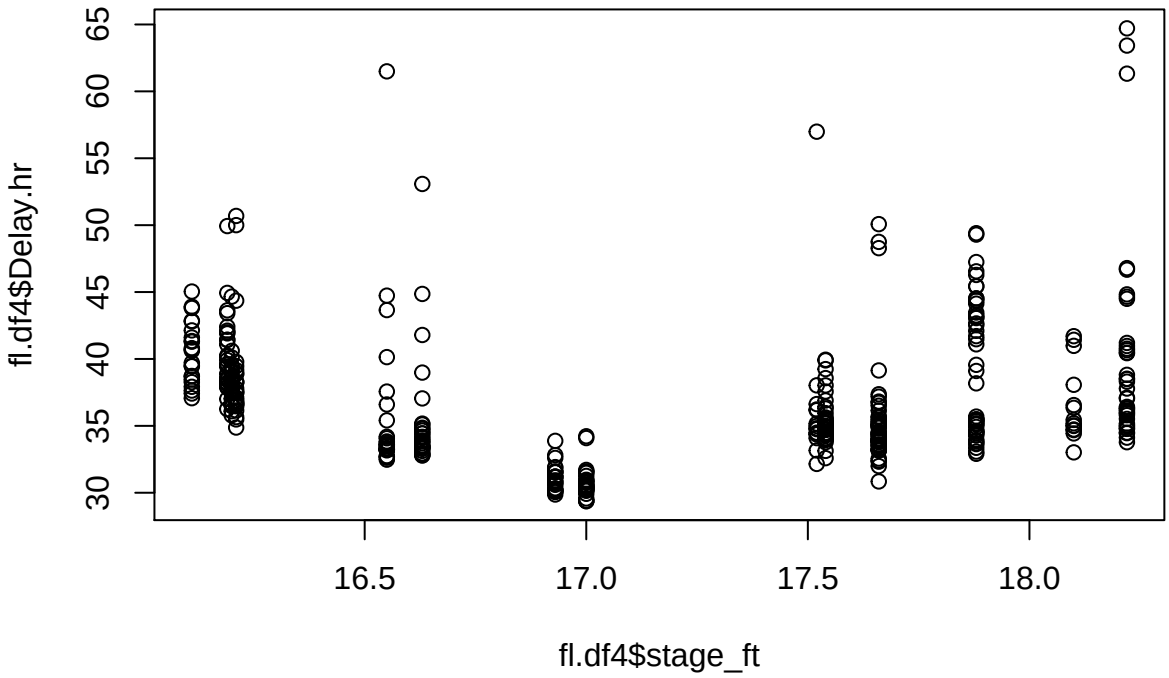
The fit is very non-significant (p=0.83), but the diagnistic plots still don't look good.

One final set of plots to simply look at the relationships in data:

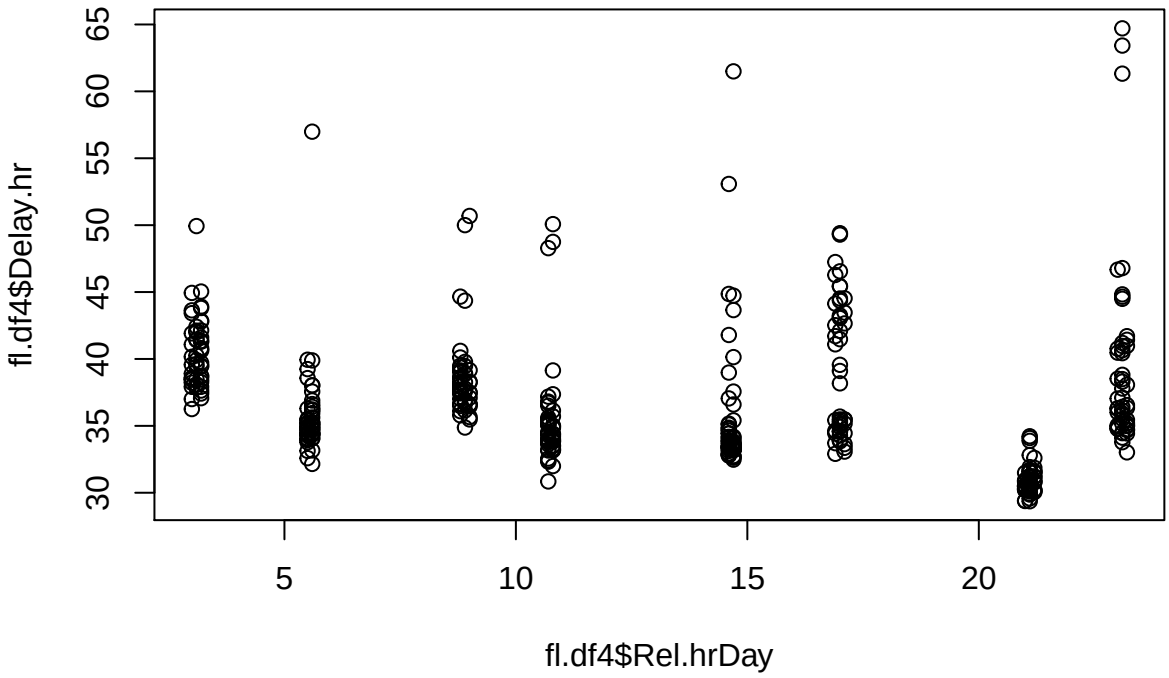
```
plot(fl.df4$stage_ft ~ fl.df4$Rel.hrDay)
```




```
plot(fl.df4$Delay.hr ~ fl.df4$stage_ft)
```




```
plot(fl.df4$Delay.hr ~ fl.df4$Rel.hrDay)
```



Hm.

From looking at the plots and the models that violate assumptions, the big picture that emerges is that the fish take ~40 hours to transit the 55.1 km between the release and the array, and this doesn't change too dramatically by the time of day they enter the river. But there does seem to be some sort of a trend between stage and transit time. The spread of fishes also changes with time of day, although there is no clear directional trend. So, time of day is less important than discharge, and the influence of time of day may be small enough to disregard or categorical and therefore we might be able to justify combining the fish from release 1 and release 2 to analyze together. We will increase our variability overall, but it might be something we can control for in a mixed-effects model down the line.