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/TaggingDataRelEv.csv",
                   colClasses=c("RelTime"="character"))
  fishrel$RelTime = str_pad(fishrel$RelTime, width=4, side="left", pad="0")
  fishrel$RelTime = str_pad(fishrel$RelTime, width=6, side="right", pad="0")

  fishrel$datetime = as.POSIXct(paste(fishrel$RelDate, fishrel$RelTime), format="%Y-%m-%d %H%M%S", 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

  dat$grp = NA
   dat$grp[dat$F.time<as.POSIXct("2016-03-05")] <- 1
   dat$grp[dat$F.time<as.POSIXct("2016-03-05")] <- 2
```

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
```

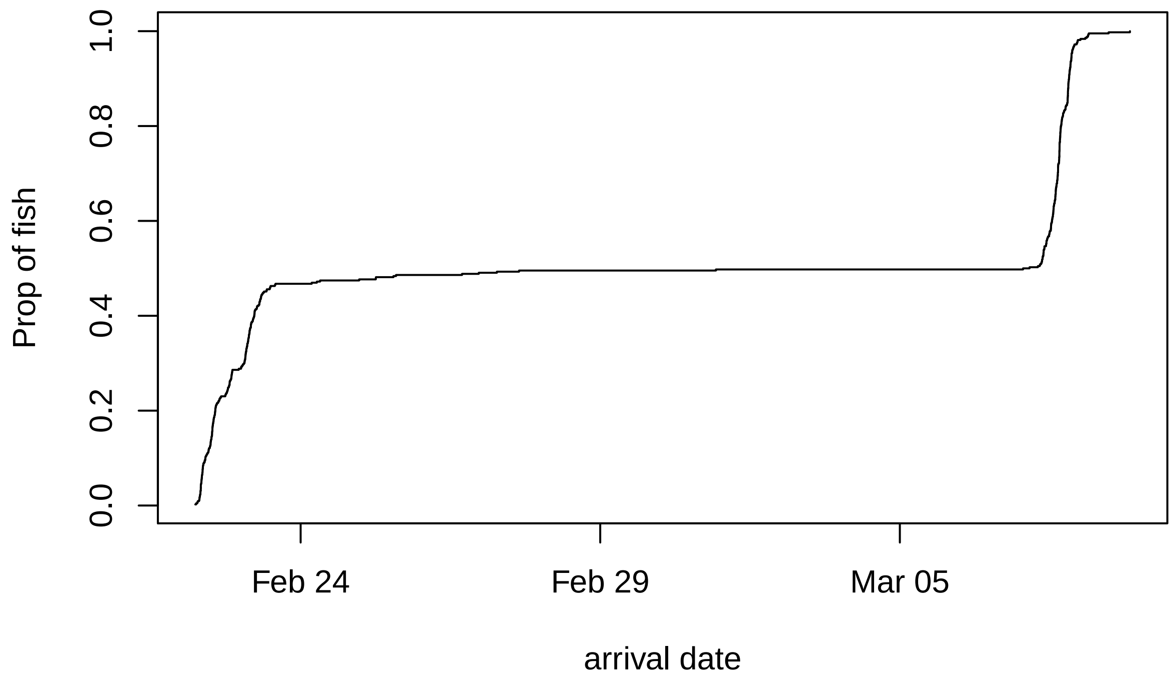
Store distance between Tisdale Weir release site and top of receiver array

- 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 = 40.3666981133873"
```

```
## [1] "sd = 27.0882547388836"
```

```
## [1] "median = 35.480412409438"
```

```
##   RelEv  mean_hr     sd_hr median_hr
## 1     1 44.43287 37.513737  35.36590
## 2     2 36.26252  4.868016  36.47698
```

```
##   RelEv Rel.hrDayfac  mean_hr     sd_hr median_hr
## 1     1            6 43.51184 46.437722  34.92753
## 2     1           11 43.44157 46.753742  34.37964
## 3     1           17 49.18773 34.779144  40.31850
## 4     1           23 41.85917 12.894065  37.05642
## 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
```

Calculate ground speed during initial transit

```
 fl.df2$transitspd.kmpd = 55.1/fl.df2$Delay.hr

mean(fl.df2$transitspd.kmpd, na.rm=T)
```

```
## [1] 1.491759
```

```
median(fl.df2$transitspd.kmpd)
```

```
## [1] 1.55297
```

```
sd(fl.df2$transitspd.kmpd)
```

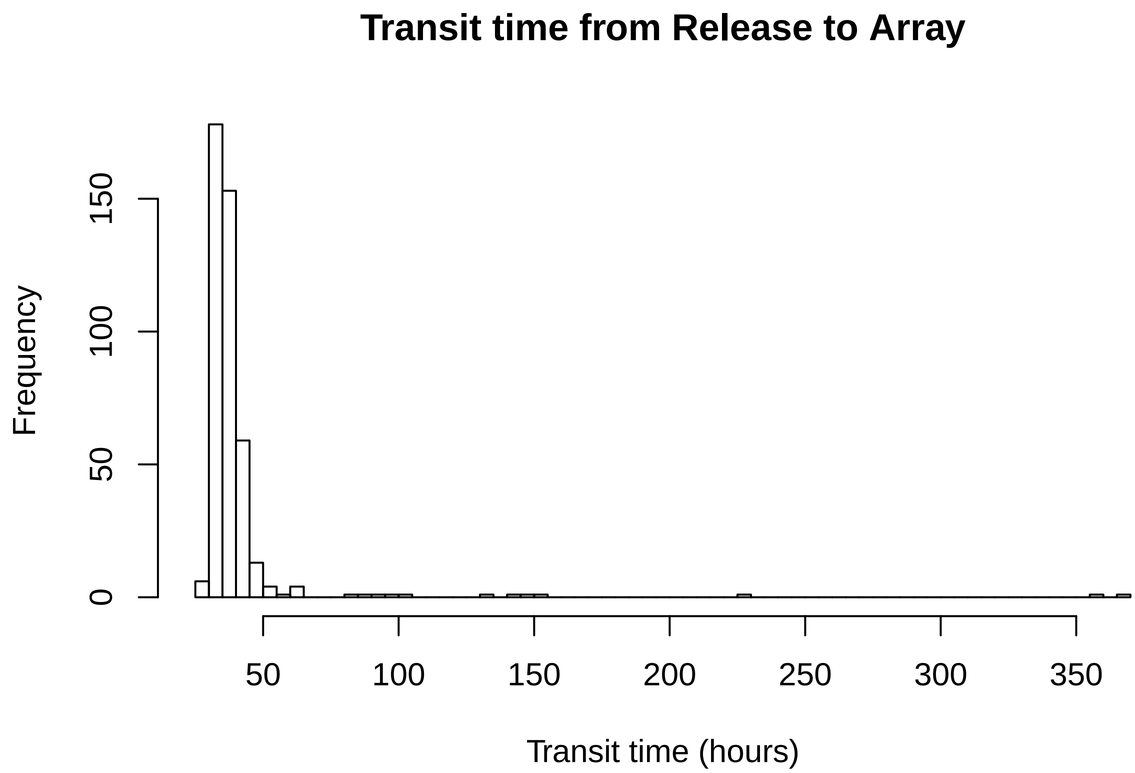
```
## [1] 0.2556208
```

```
summarize(group_by(fl.df2, RelEv), mean(transitspd.kmpd), median(transitspd.kmpd), sd(transitspd.kmpd))
```

```
## # A tibble: 2 x 4
##   RelEv mean(transitspd.kmpd) median(transitspd.kmpd) sd(transitspd.kmpd)
##   <int>                 <dbl>                   <dbl>               <dbl>
## 1     1              1.439468                1.557998           0.2977202
## 2     2              1.544539                1.510542           0.1912145
```

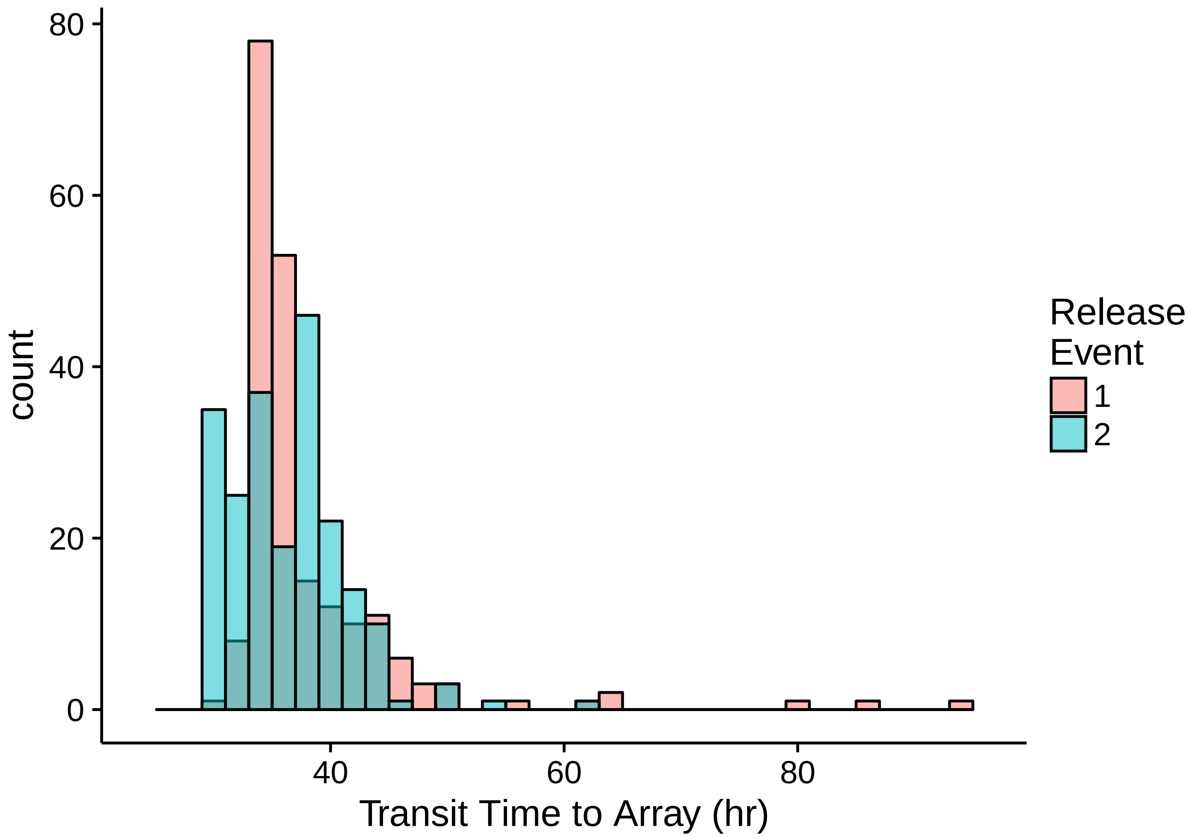
Plots to visualize initial transit time and passage time

```
 hist(fl.df2$Delay.hr,
      main="Transit time from Release to Array",
      xlab="Transit time (hours)", ylab="Frequency", breaks=50)
```



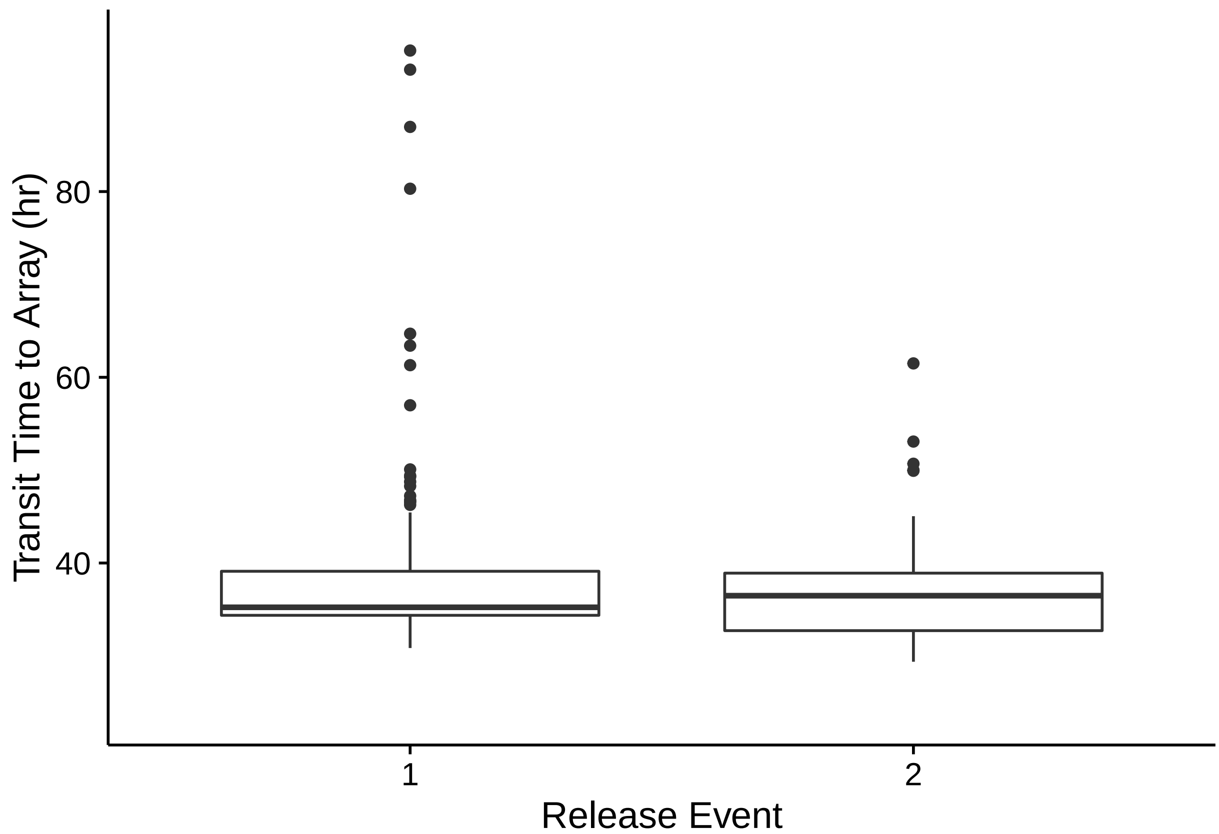

```
ggplot(data=fl.df2, aes(x=Delay.hr, group=factor(RelEv), fill=factor(RelEv))) +
  geom_histogram(color="black", alpha=0.5, position="identity", binwidth=2) +
  xlim(c(24,96)) + xlab("Transit Time to Array (hr)") +
  scale_fill_discrete(name="Release\nEvent")
```

```
## Warning: Removed 8 rows containing non-finite values (stat_bin).
```




```
ggplot(data=fl.df2, aes(y=Delay.hr, x=factor(RelEv))) +
  geom_boxplot() +
  ylim(c(24,96)) + ylab("Transit Time to Array (hr)") + xlab("Release Event")
```

```
## Warning: Removed 8 rows containing non-finite values (stat_boxplot).
```



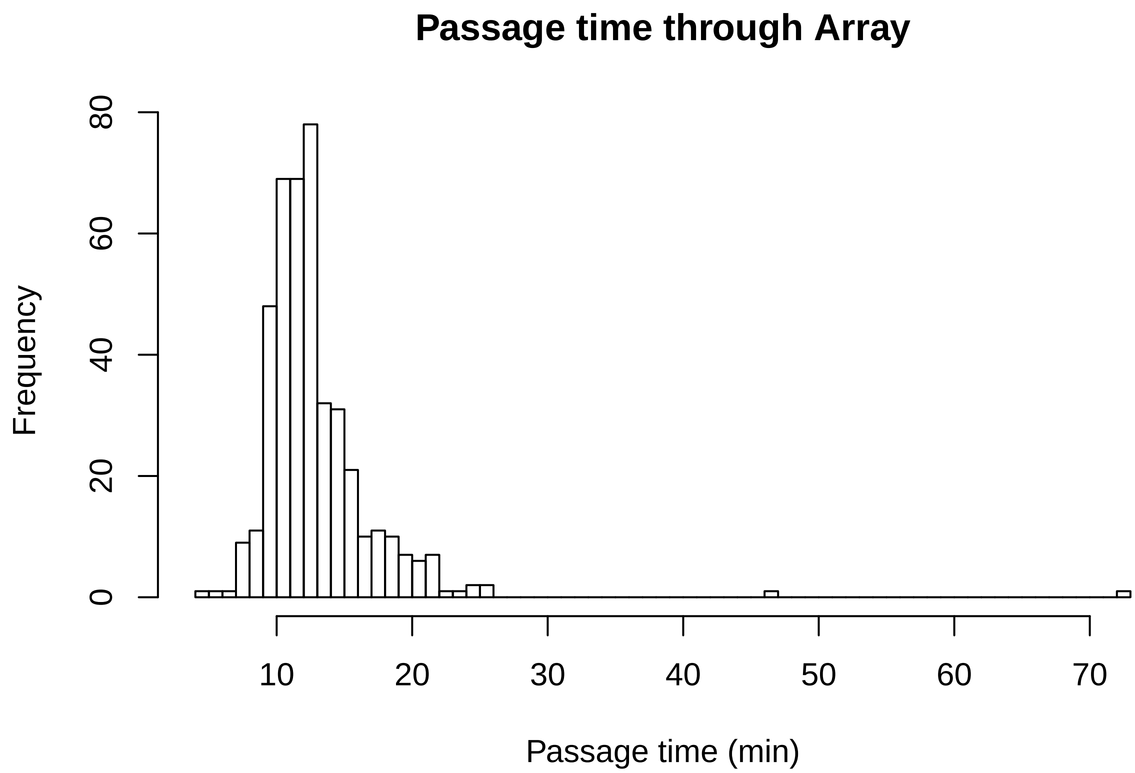

```
wilcox.test(Delay.hr ~ RelEv, data=fl.df2)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Delay.hr by RelEv
## W = 26804, p-value = 0.00417
## alternative hypothesis: true location shift is not equal to 0
```

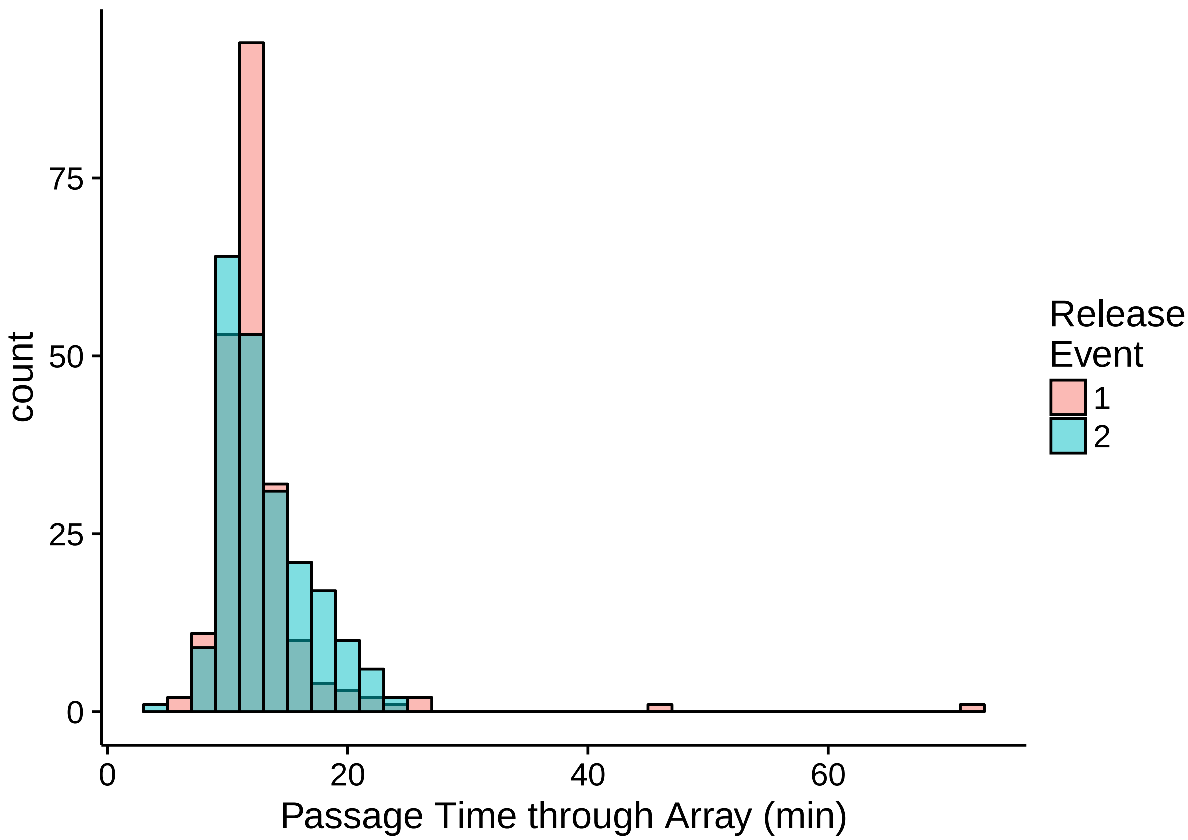
```
summarize(group_by(fl.df2, RelEv), meanDelay = mean(Delay.hr), sdDelay=sd(Delay.hr))
```

```
## # A tibble: 2 x 3
##   RelEv meanDelay   sdDelay
##   <int>     <dbl>     <dbl>
## 1     1  44.43287 37.513737
## 2     2  36.26252  4.868016
```

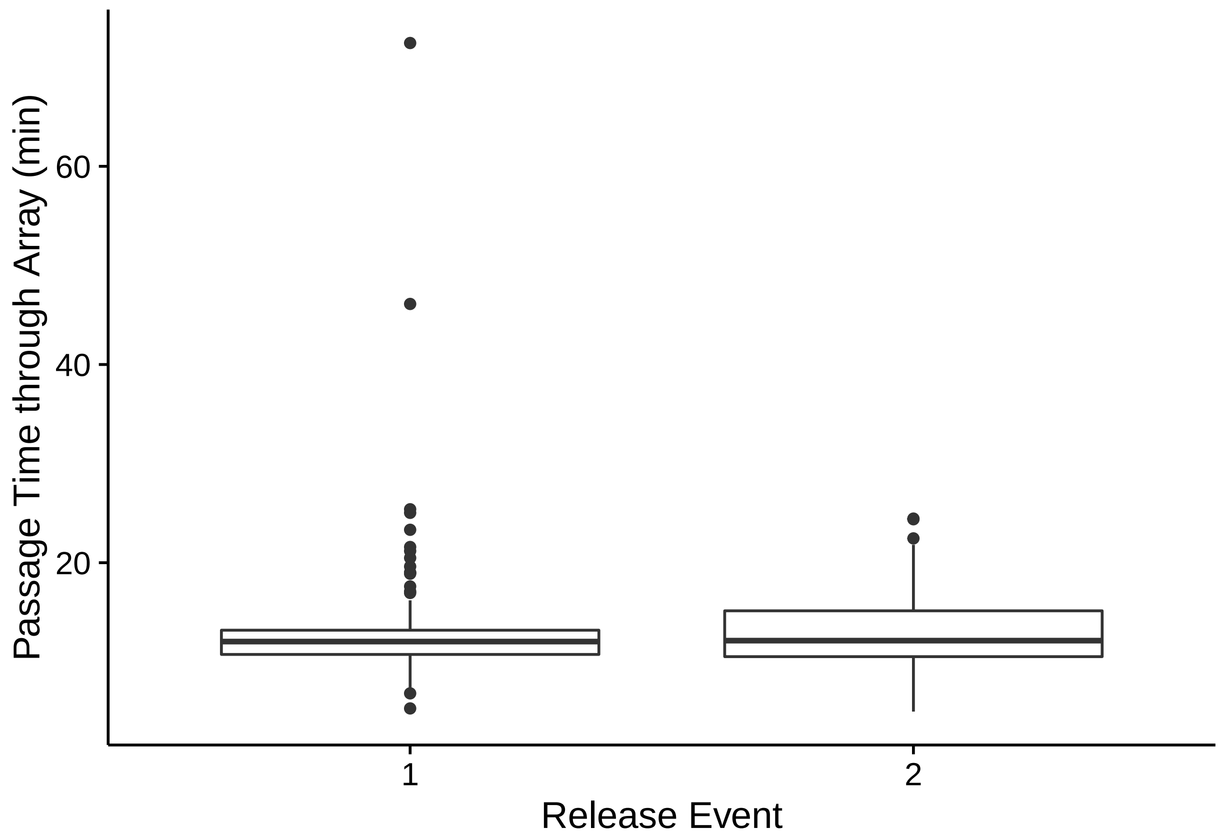
```
hist(fl.df2$passtime.min,
     main="Passage time through Array",
     xlab="Passage time (min)",ylab="Frequency", breaks=50)
```




```
ggplot(data=fl.df2, aes(x=passtime.min, group=factor(RelEv), fill=factor(RelEv))) +
  geom_histogram(color="black", alpha=0.5, position="identity", binwidth=2) +
  xlab("Passage Time through Array (min)") +
  scale_fill_discrete(name="Release\nEvent")
```



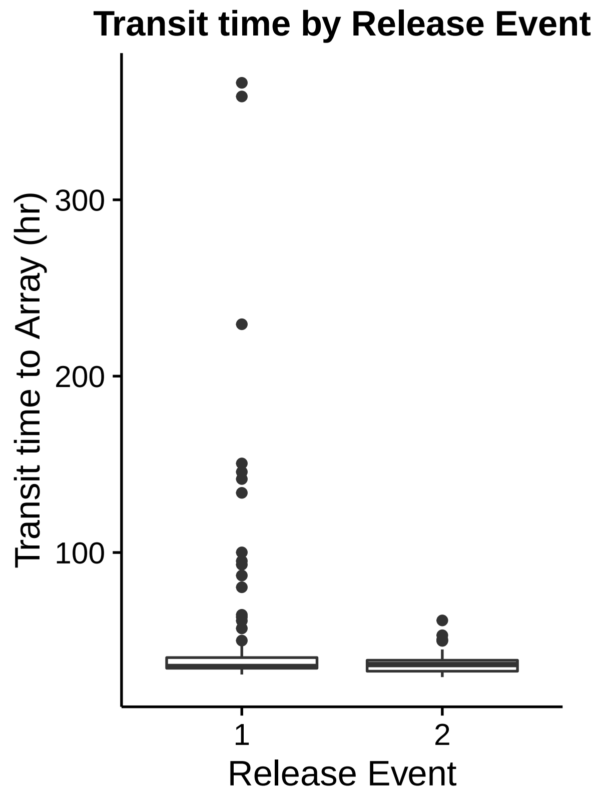

```
ggplot(data=fl.df2, aes(y=passtime.min, x=factor(RelEv))) +
  geom_boxplot() +
  ylab("Passage Time through Array (min)") + xlab("Release Event")
```



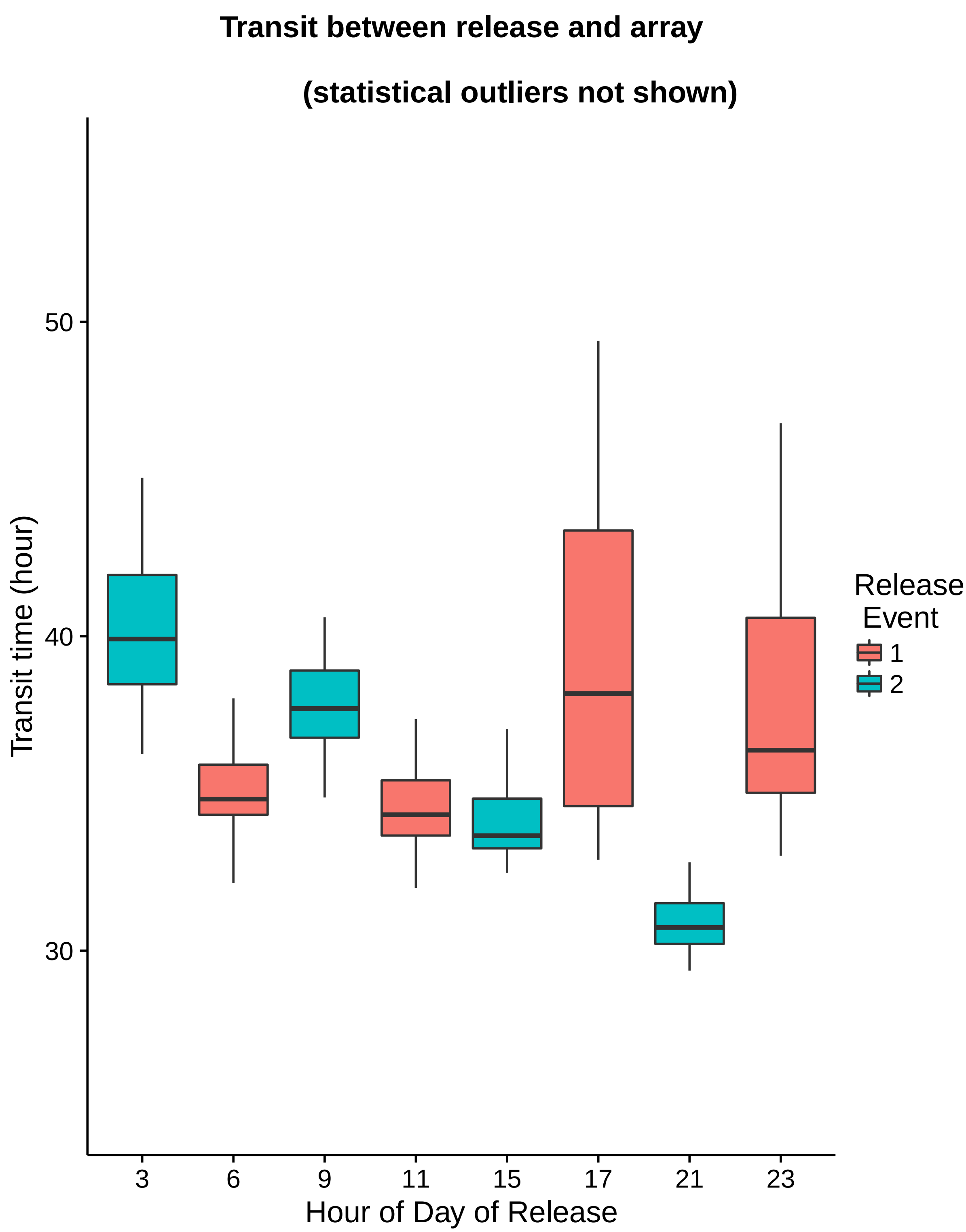

```r
wilcox.test(passtime.min ~ RelEv, data=fl.df2)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  passtime.min by RelEv
## W = 21583, p-value = 0.2355
## alternative hypothesis: true location shift is not equal to 0
```

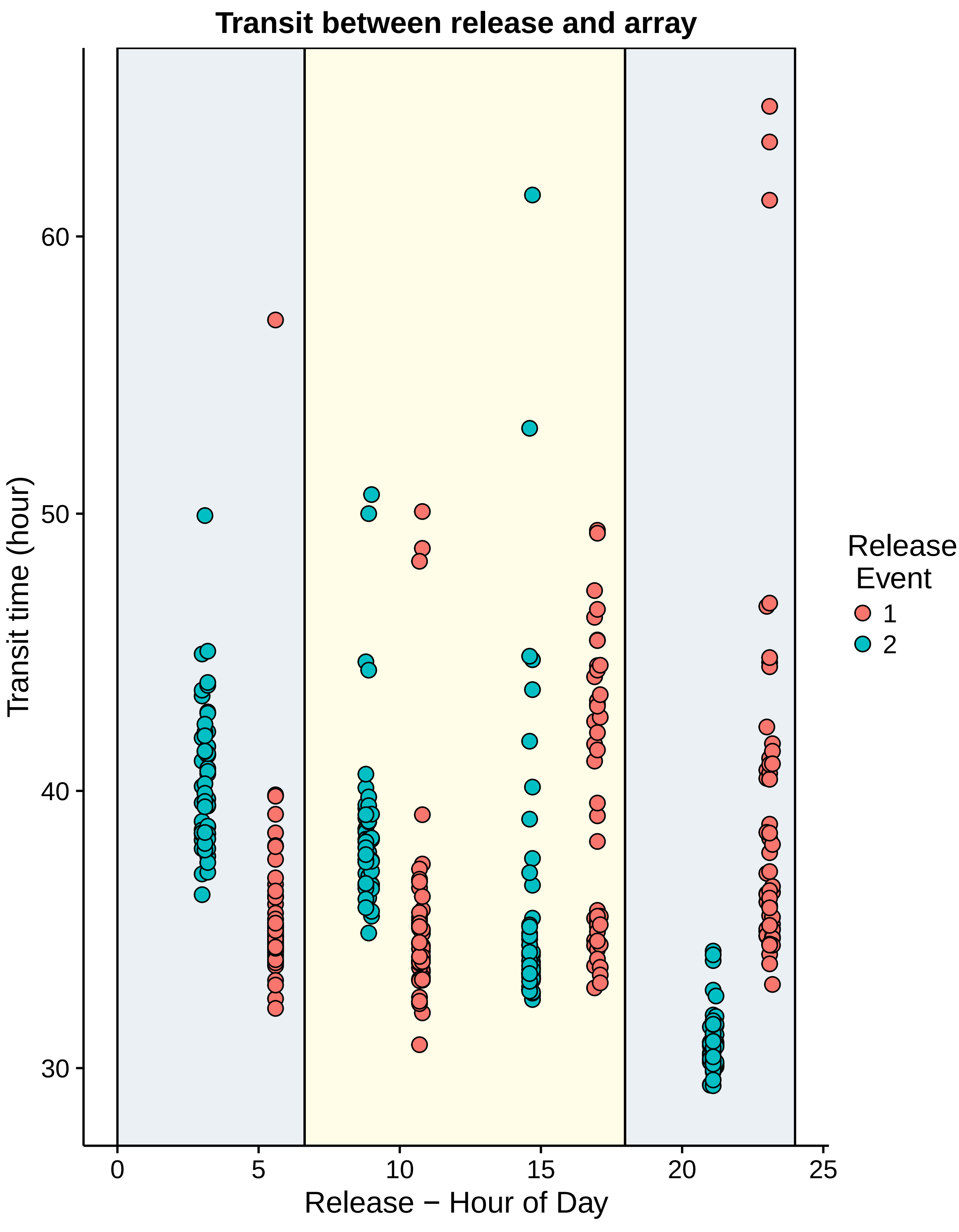
Differences by Release Event or Release Hour

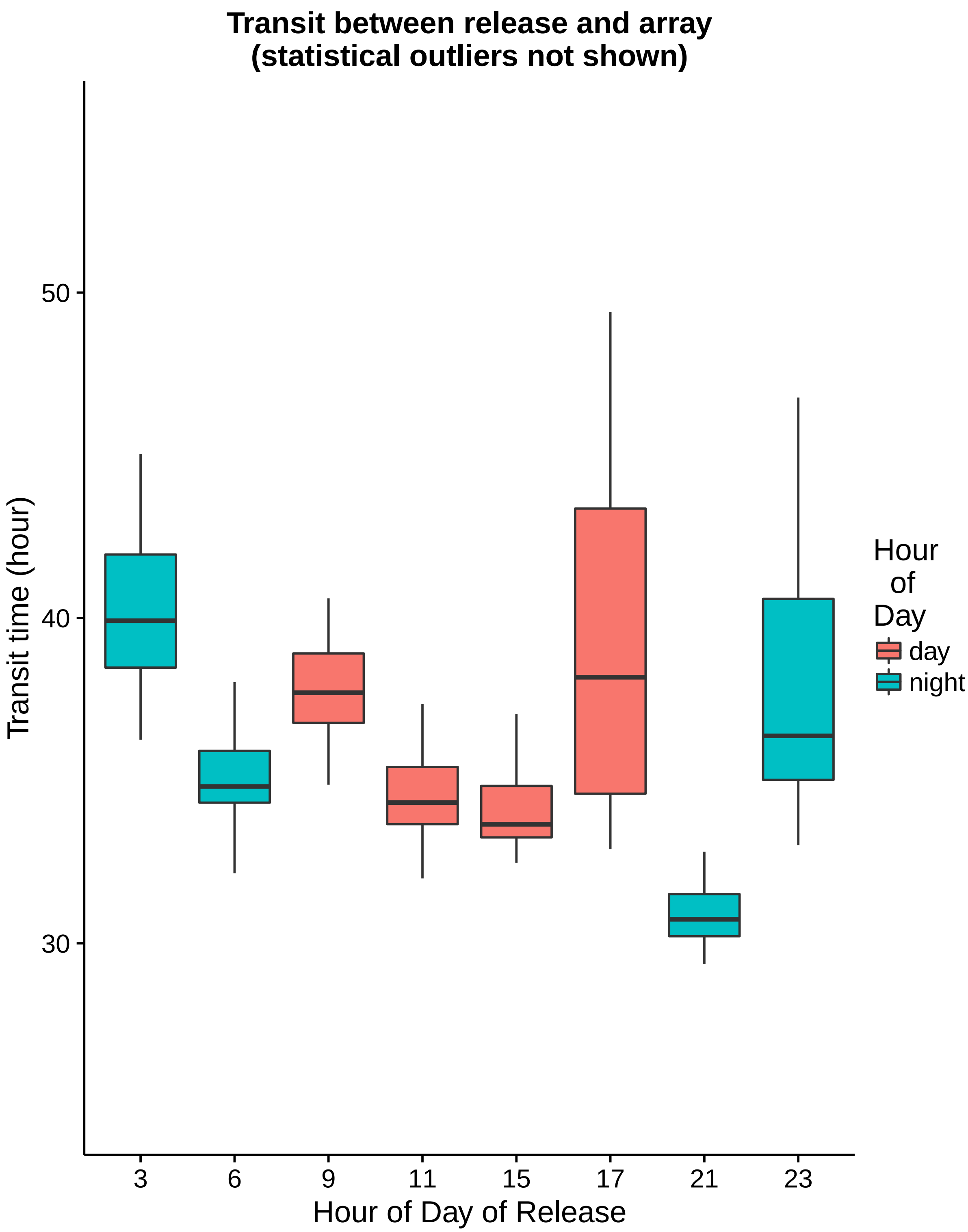


- note: it would be nice to annotate boxes with respective sample sizes



Transit between release and array
(statistical outliers not shown)
Transit time (hour)
50
40
30
3
6
9
11
15
17
21
23
Hour of Day of Release
Release Event
1
2


Transit between release and array
Transit time (hour)
60
50
40
30
0
5
10
15
20
25
Release − Hour of Day
Release
Event
1
2



- 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)

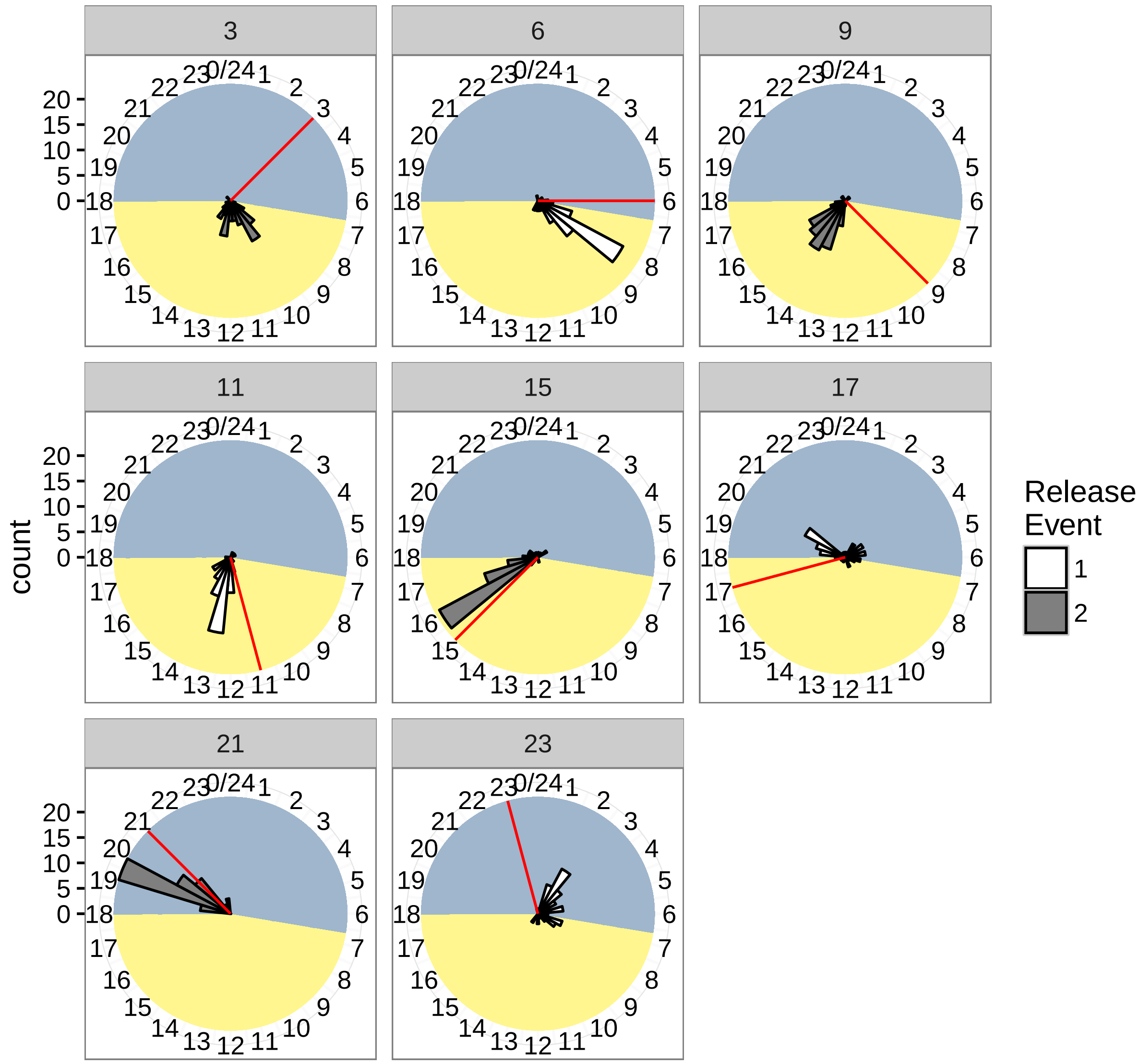
```
## [1] "mean = 13.7783149990447"
```

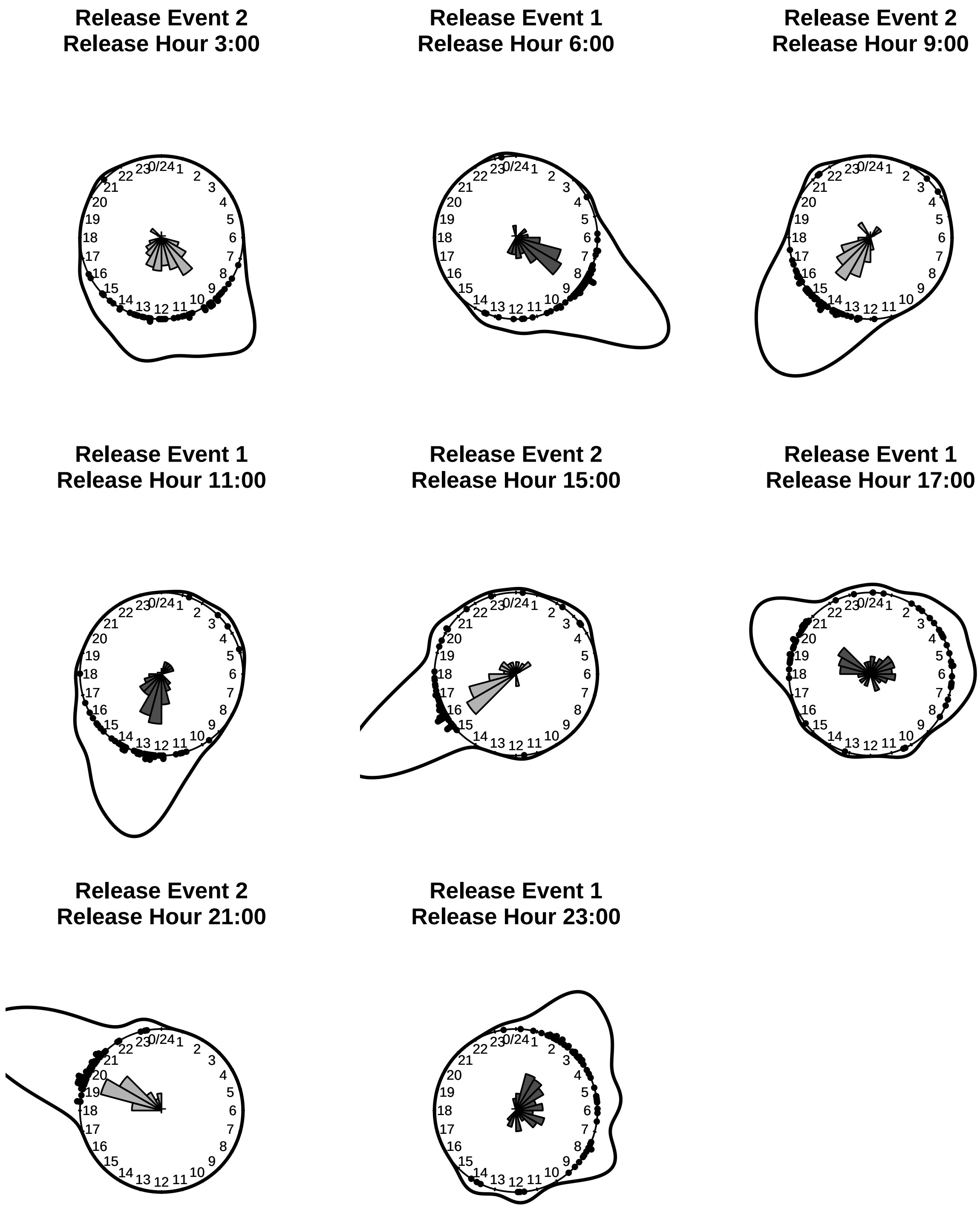
```
## [1] "sd = 1.6187296058185"
```

```
## [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     3.85000  4.539209 0.9462162
```

Circular plots: same data, two visualizations





- 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

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.df2$Delay.hr, fl.df2$Rel.hrDay)
```

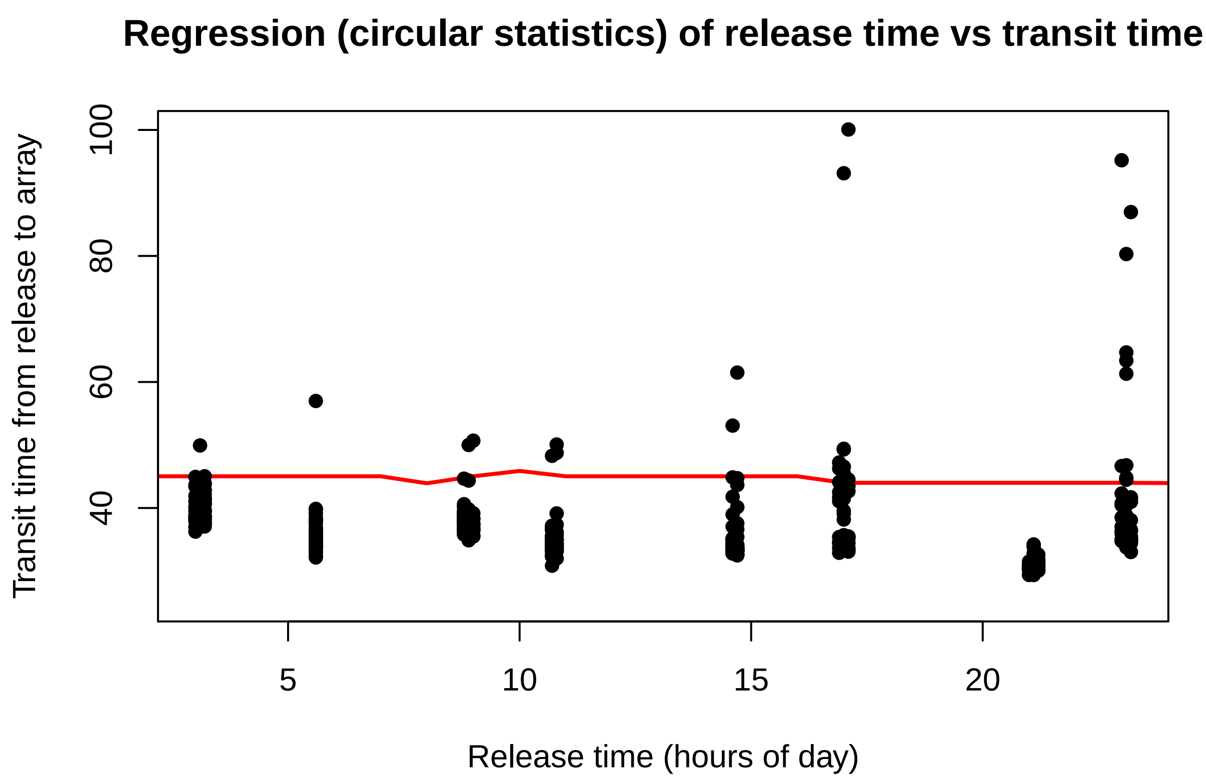
```
## [1] 0.3924005
```

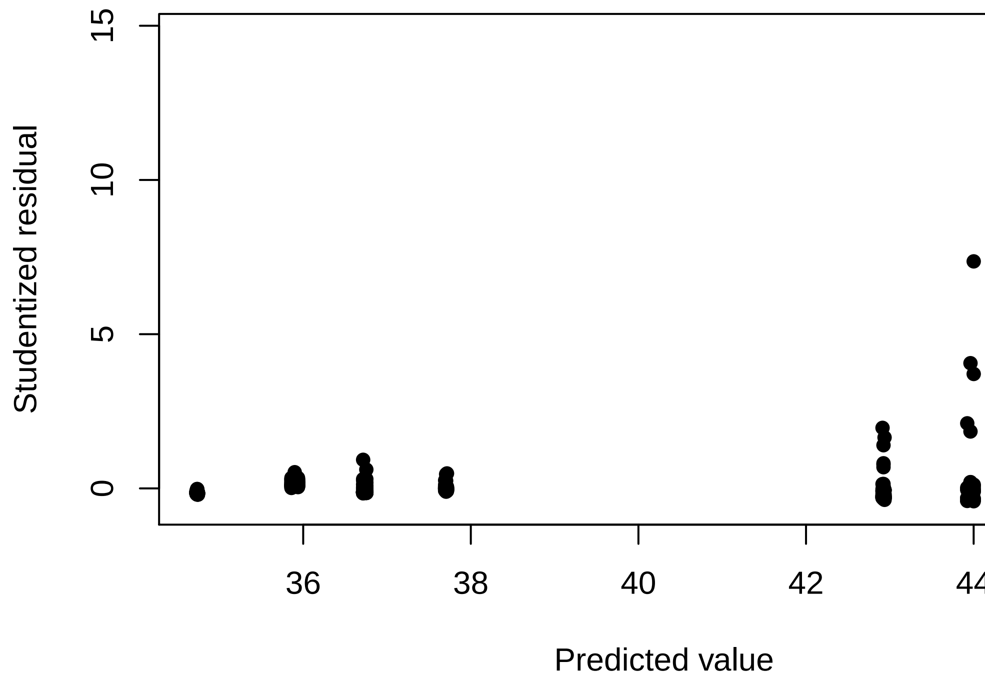
```r
# Next step: regression of a linear response on a circular predictor
# basic cosine model: x = a + b1*cos(2pi/24*Rel.hr*Day) + b2*sin(2pi/12*Rel.hrDay) + e
omega = 2*pi/24
cosrelhr = cos(omega*fl.df2$Rel.hrDay)
sinrelhr = sin(omega*fl.df2$Rel.hrDay)

delaymod = lm(fl.df2$Delay.hr ~ cosrelhr + sinrelhr + factor(fl.df2$RelEv))
summary(delaymod)
```

```
##
## Call:
## lm(formula = fl.df2$Delay.hr ~ cosrelhr + sinrelhr + factor(fl.df2$RelEv))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -15.04  -8.69  -3.47   0.77 321.29
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             44.3845     1.8286  24.273  < 2e-16 ***
## cosrelhr                -1.3123     1.8390  -0.714  0.47586
## sinrelhr                 0.7889     1.8302   0.431  0.66667
## factor(fl.df2$RelEv)2  -8.1548     2.5902  -3.148  0.00176 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.85 on 426 degrees of freedom
## Multiple R-squared:  0.02446,    Adjusted R-squared:  0.01759
## F-statistic: 3.561 on 3 and 426 DF,  p-value: 0.01434
```

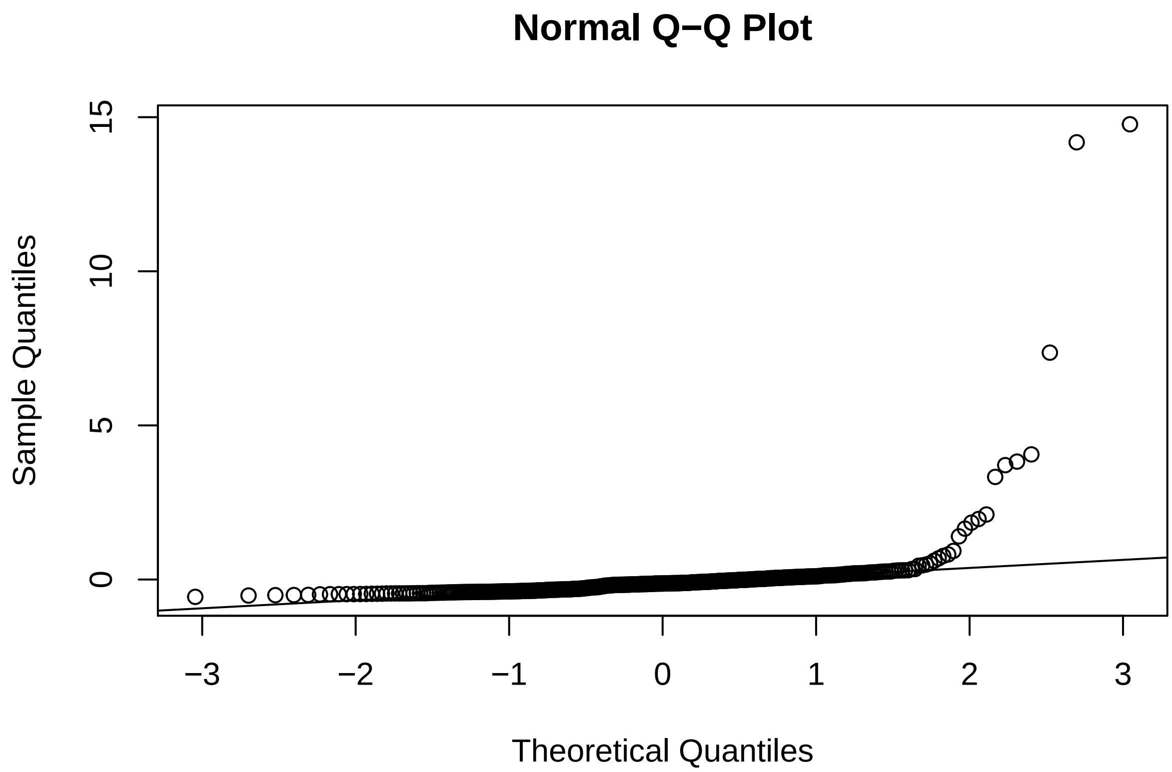
```
plot(fl.df2$Rel.hrDay, fl.df2$Delay.hr,
     xlab="Release time (hours of day)",
     ylab="Transit time from release to array",
     main="Regression (circular statistics) of release time vs transit time",
     pch=16, ylim=c(25,100),
     lines(predict(delaymod), lwd=2, col="red") )
```





Diagnostics of basic cosine model:

```
##
##  Shapiro-Wilk normality test
##
## data:  delayresid
## W = 0.24609, p-value < 2.2e-16
```



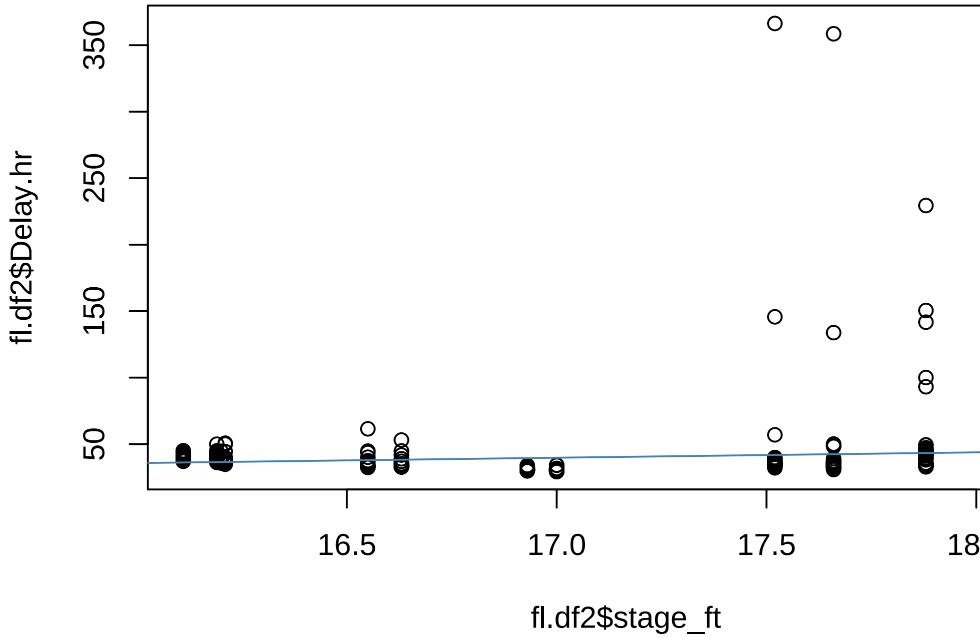

```
##
##  Bartlett test of homogeneity of variances
##
## data:  delayresid and fl.df2$Rel.hrDay
## Bartlett's K-squared = 1143.5, df = 19, p-value < 2.2e-16
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  delayresid and fl.df2$Rel.hrDay
## Fligner-Killeen:med chi-squared = 98.12, df = 19, p-value =
## 1.171e-12
```

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

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.df2$Delay.hr ~ fl.df2$stage_ft)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11.58  -8.21  -5.00   1.23 324.46
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -28.703     30.995  -0.926   0.3549
## fl.df2$stage_ft    4.028      1.806   2.230   0.0262 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.96 on 428 degrees of freedom
## Multiple R-squared:  0.01149,    Adjusted R-squared:  0.00918
## F-statistic: 4.975 on 1 and 428 DF,  p-value: 0.02624
```

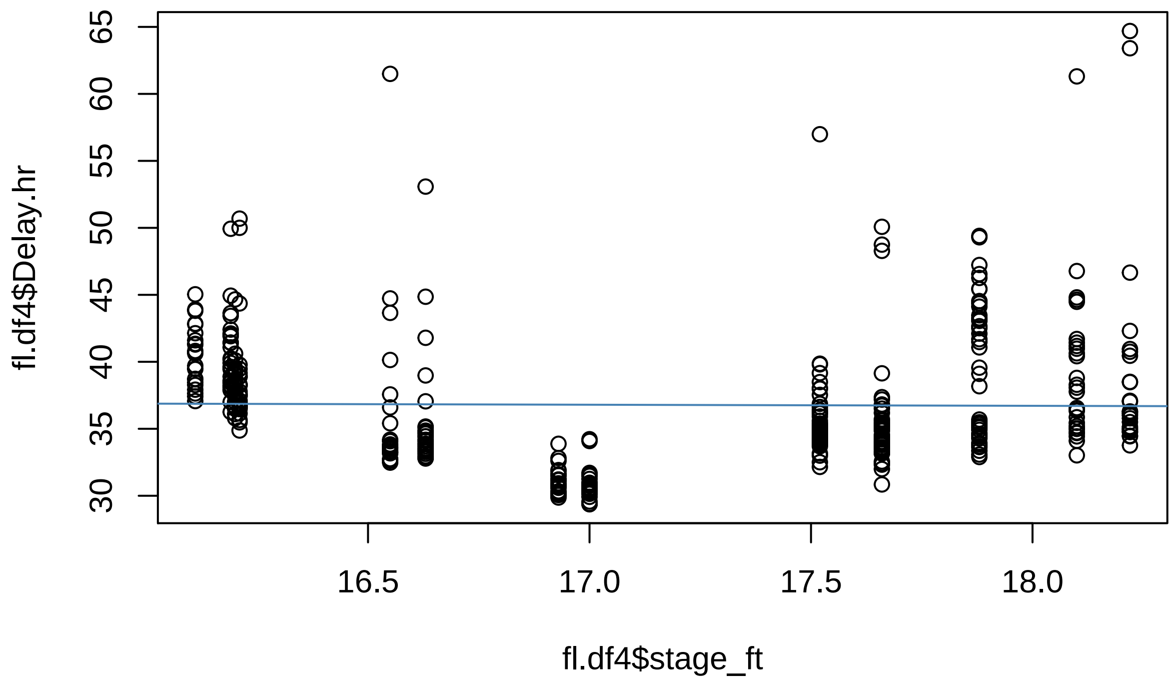
At higher stages fish actually took longer. Is this driven by the heavy outliers?

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

```
fl.df4 = fl.df2[fl.df2$Delay.hr<72,]

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

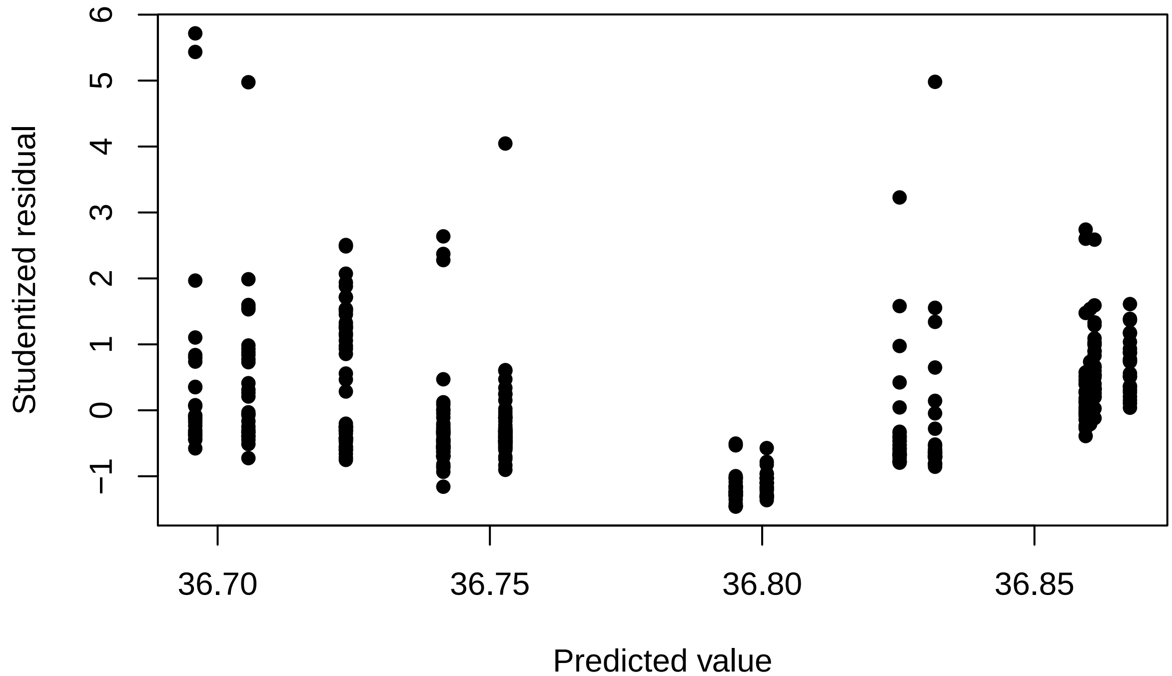
```r
stgmod = lm(fl.df4$Delay.hr ~ fl.df4$stage_ft)
abline(stgmod, col="steelblue")
```



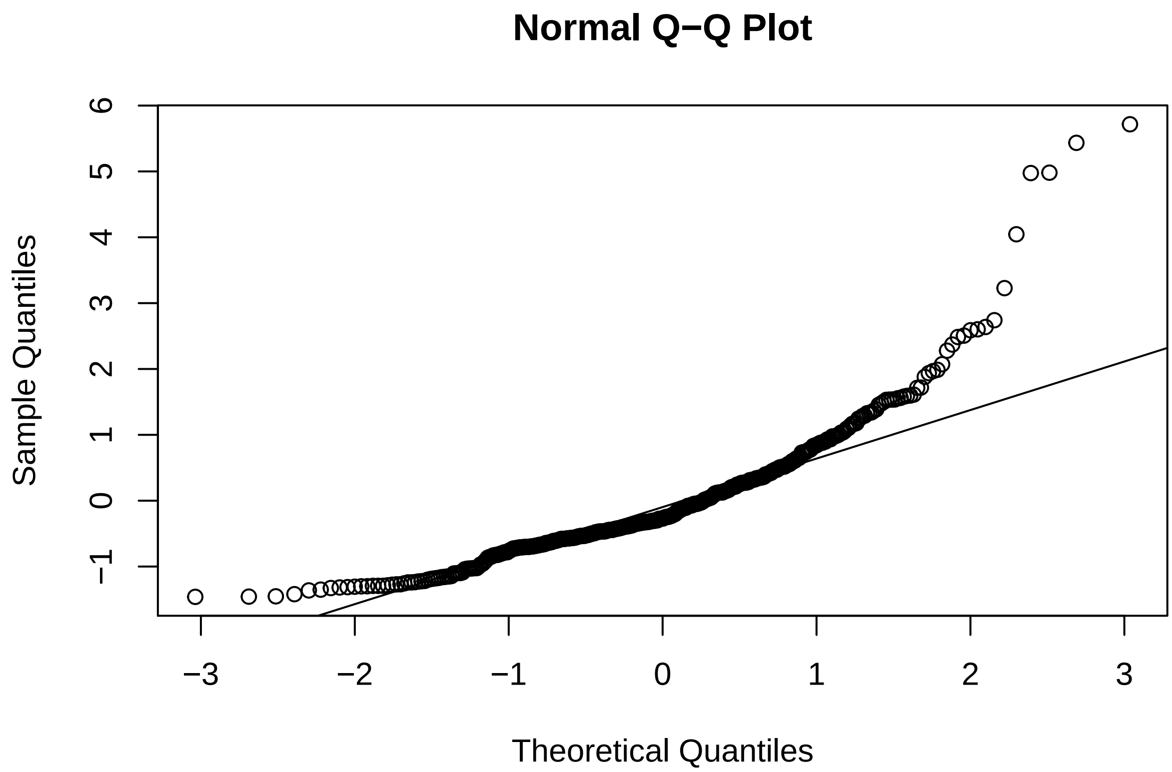

```r
summary(stgmod)
```

```
##
## Call:
## lm(formula = fl.df4$Delay.hr ~ fl.df4$stage_ft)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.430 -3.026 -1.377  2.035 27.997
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     38.17809    5.94535   6.422 3.69e-10 ***
## fl.df4$stage_ft -0.08135    0.34678  -0.235    0.815
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.101 on 416 degrees of freedom
## Multiple R-squared:  0.0001323,  Adjusted R-squared:  -0.002271
## F-statistic: 0.05503 on 1 and 416 DF,  p-value: 0.8146
```

Diagnostics




```
##
##  Shapiro-Wilk normality test
##
## data:  stgresid
## W = 0.85647, p-value < 2.2e-16
```



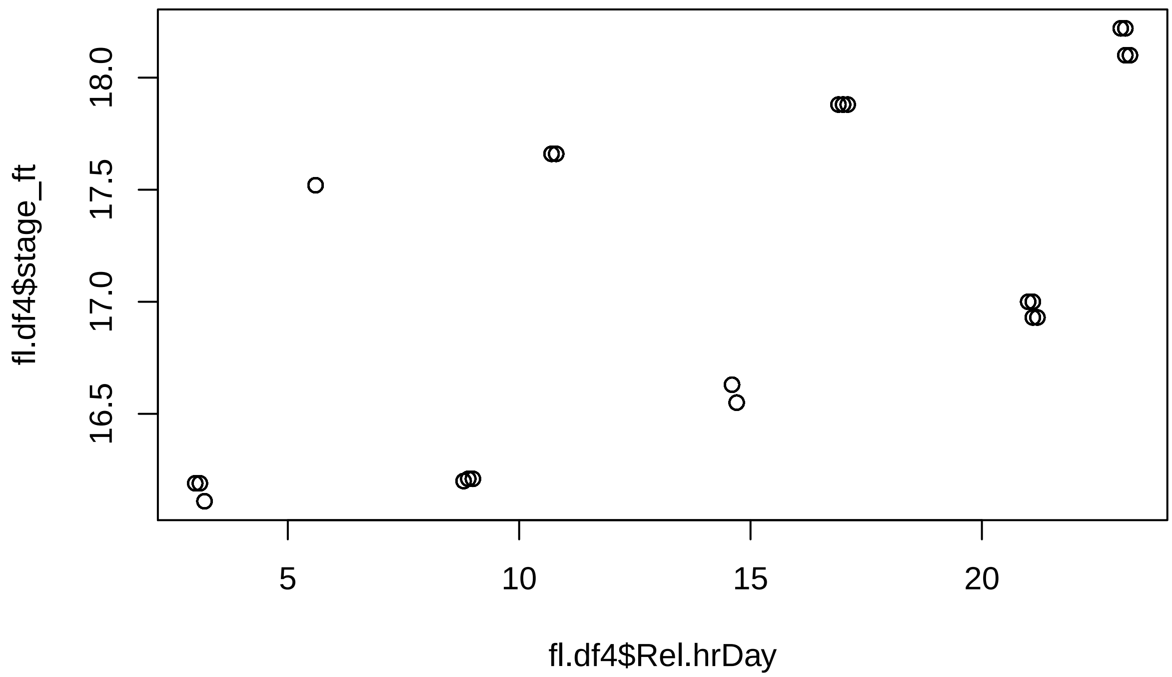

```
##
##  Bartlett test of homogeneity of variances
##
## data:  stgresid and fl.df4$stage_ft
## Bartlett's K-squared = 199.32, df = 12, p-value < 2.2e-16
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  stgresid and fl.df4$stage_ft
## Fligner-Killeen:med chi-squared = 106.91, df = 12, p-value <
## 2.2e-16
```

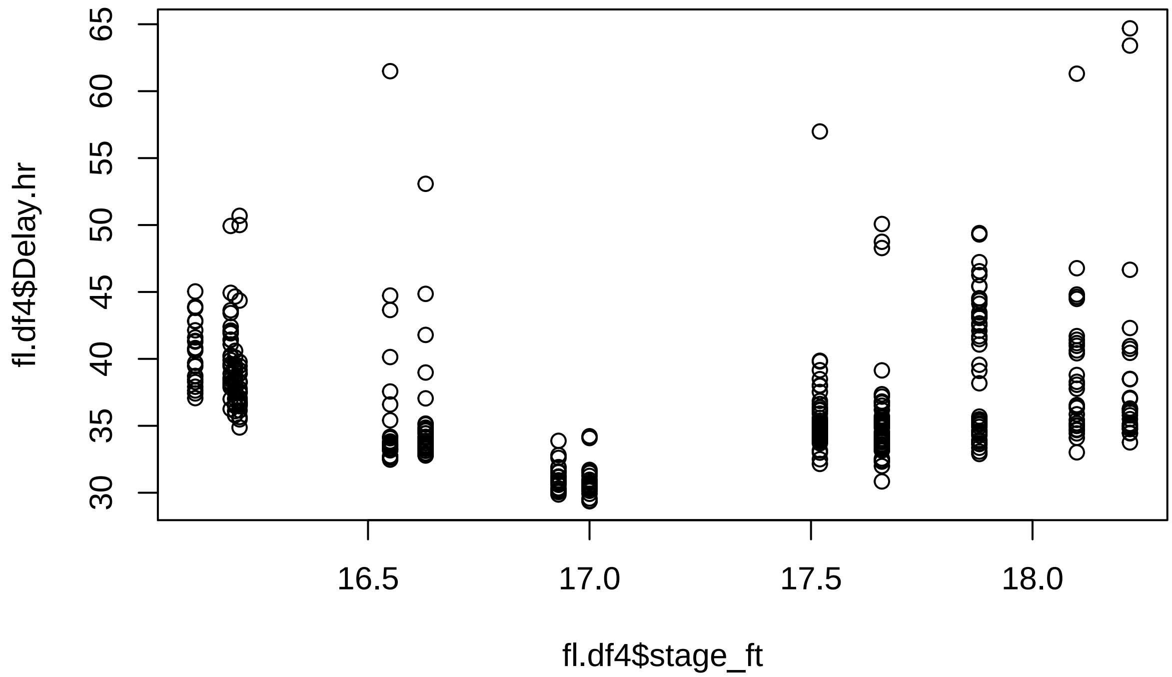
The fit is very non-significant (p=0.83), but the diagnistic plots still don't look good.

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

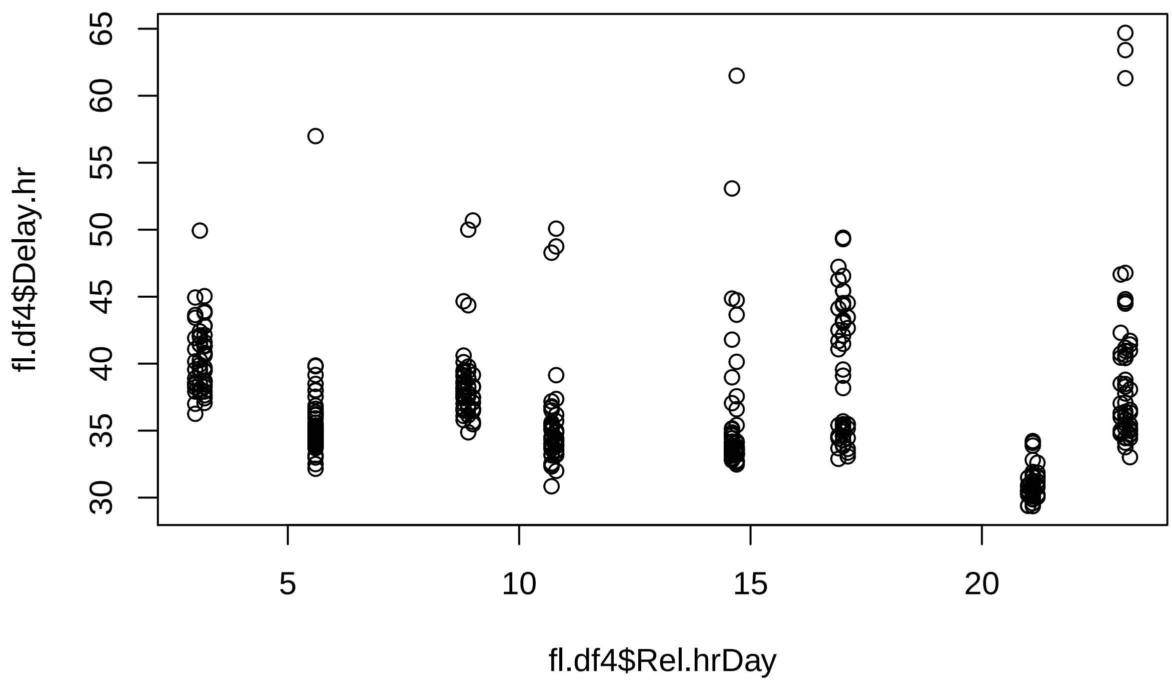
```
plot(fl.df4$stage_ft ~ fl.df4$Rel.hrDay)
```




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




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



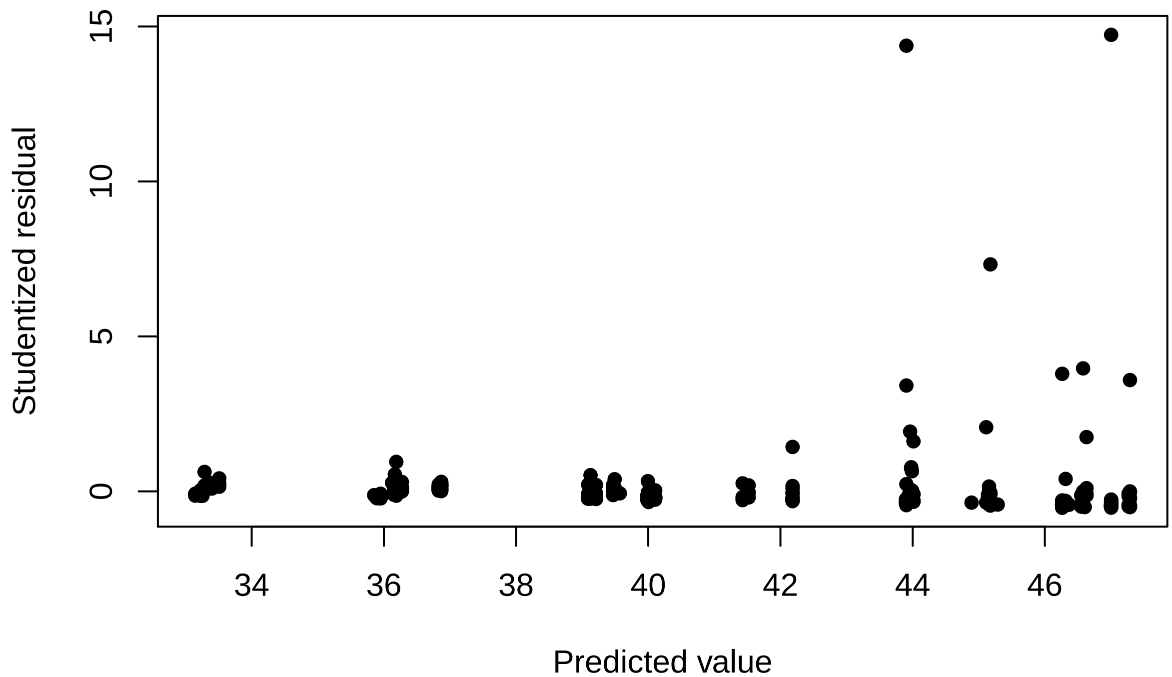
Messier Circular Statistics:

Extended cosine model (additional sin & cos parameters):

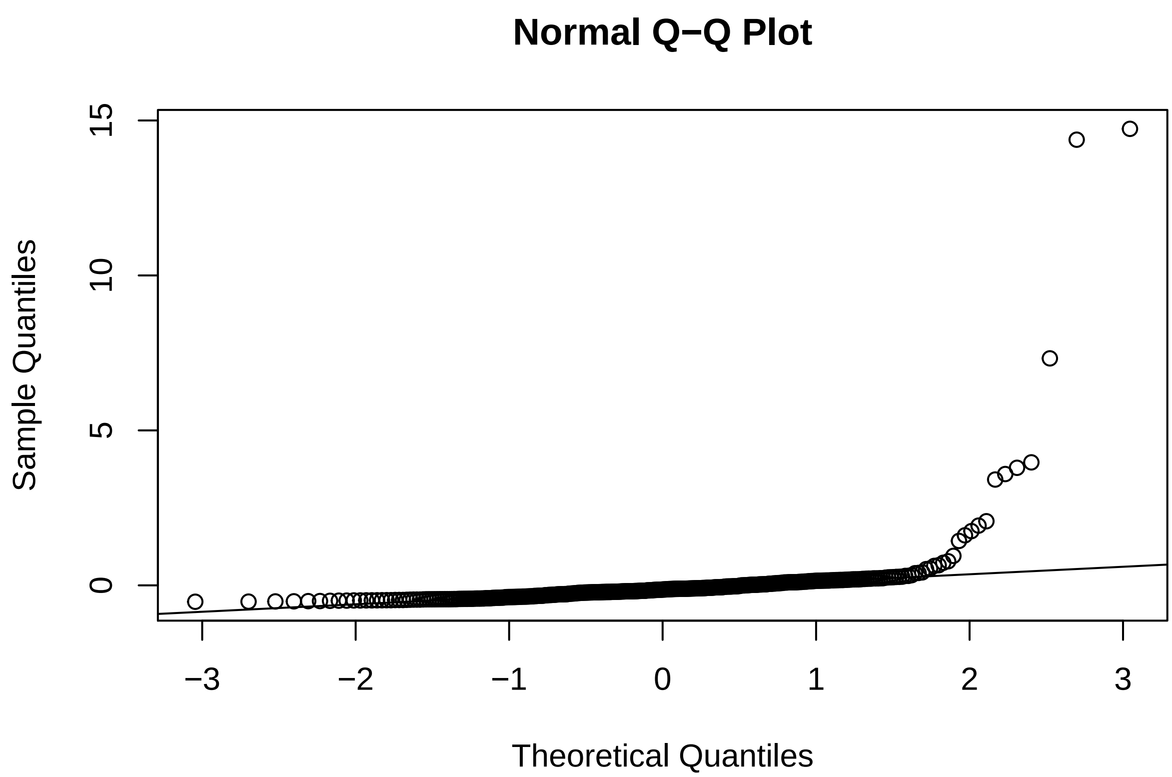
```
##
## Call:
## lm(formula = fl.df2$Delay.hr ~ cosrelhr + sinrelhr + cos2var +
##     sin2var + factor(fl.df2$RelEv))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14.11  -7.73  -3.62   0.97 319.32
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           44.52485    1.83885  24.213  < 2e-16 ***
## cosrelhr               2.03271    1.92160   1.058  0.29074
## sinrelhr              -0.02533    1.82974  -0.014  0.98896
## cos2var               -1.98769    2.29391  -0.867  0.38670
## sin2var                1.66925    1.80930   0.923  0.35674
## factor(fl.df2$RelEv)2 -8.14527    2.60016  -3.133  0.00185 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.83 on 424 degrees of freedom
## Multiple R-squared:  0.03036,    Adjusted R-squared:  0.01893
## F-statistic: 2.655 on 5 and 424 DF,  p-value: 0.02228
```

Diagnostics of extended cosine model:




```
##
##  Shapiro-Wilk normality test
##
## data:  delay2resid
## W = 0.24733, p-value < 2.2e-16
```




```
##
##  Bartlett test of homogeneity of variances
##
## data:  delay2resid and fl.df2$Rel.hrDay
## Bartlett's K-squared = 1115.2, df = 19, p-value < 2.2e-16
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  delay2resid and fl.df2$Rel.hrDay
## Fligner-Killeen:med chi-squared = 95.694, df = 19, p-value =
## 3.199e-12
```

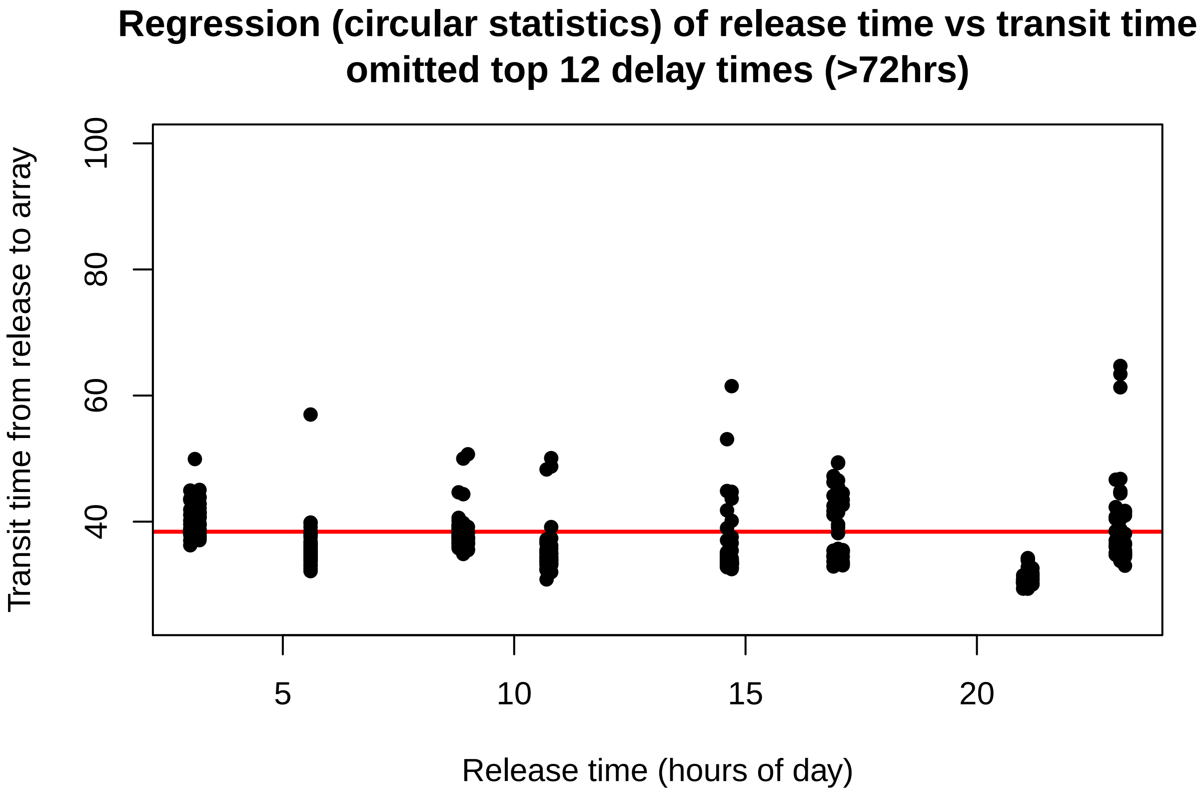
Still doesn't meet assumptions well. Could be due to outliers?

Extended cosine model without outliers

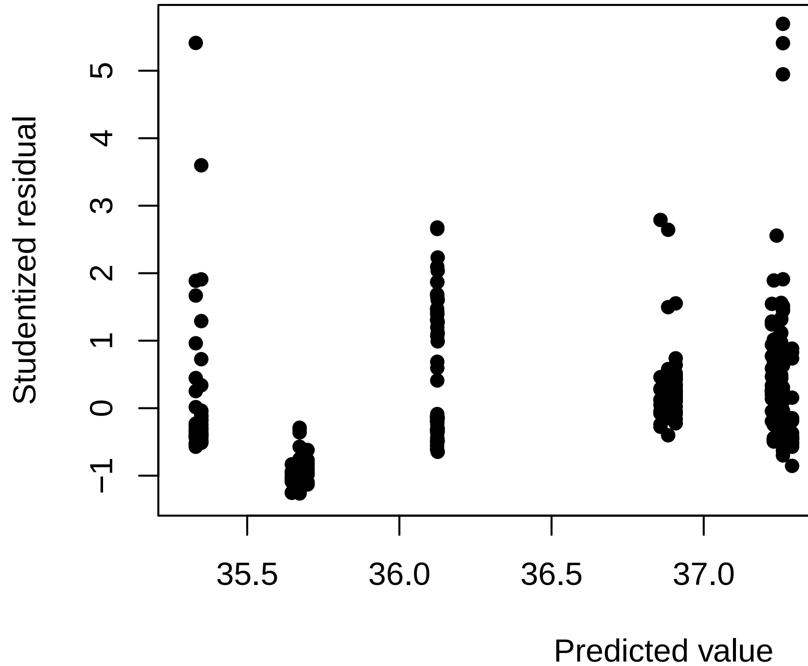
```
##
## Call:
## lm(formula = fl.df4$Delay.hr ~ cosvar + sinvar + cos2var + sin2var +
##     factor(fl.df4$RelEv))
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
## -6.274 -3.257 -1.682  2.120 27.276
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             37.4693     0.3414 109.766  < 2e-16 ***
## cosvar                   0.3270     0.3360   0.973   0.3310
## sinvar                   1.3419     0.3408   3.938 9.64e-05 ***
## cos2var                  0.8998     0.4100   2.195   0.0287 *
## sin2var                  1.8908     0.3263   5.795 1.36e-08 ***
## factor(fl.df4$RelEv)2  -1.2205     0.4757  -2.566   0.0106 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.84 on 412 degrees of freedom
## Multiple R-squared:  0.1084, Adjusted R-squared:  0.0976
## F-statistic: 10.02 on 5 and 412 DF,  p-value: 4.6e-09
```

```
##
## Call:
## lm(formula = fl.df4$Delay.hr ~ sinvar + cosvar + factor(fl.df4$RelEv))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.552 -3.335 -1.263  1.706 27.433
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)             37.2690     0.3522 105.816  < 2e-16 ***
## sinvar                   1.1155     0.3483   3.203  0.00147 **
## cosvar                   0.2589     0.3481   0.744  0.45742
## factor(fl.df4$RelEv)2  -1.0168     0.4916  -2.068  0.03925 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.022 on 414 degrees of freedom
## Multiple R-squared:  0.0357, Adjusted R-squared:  0.02871
## F-statistic: 5.108 on 3 and 414 DF,  p-value: 0.001761
```

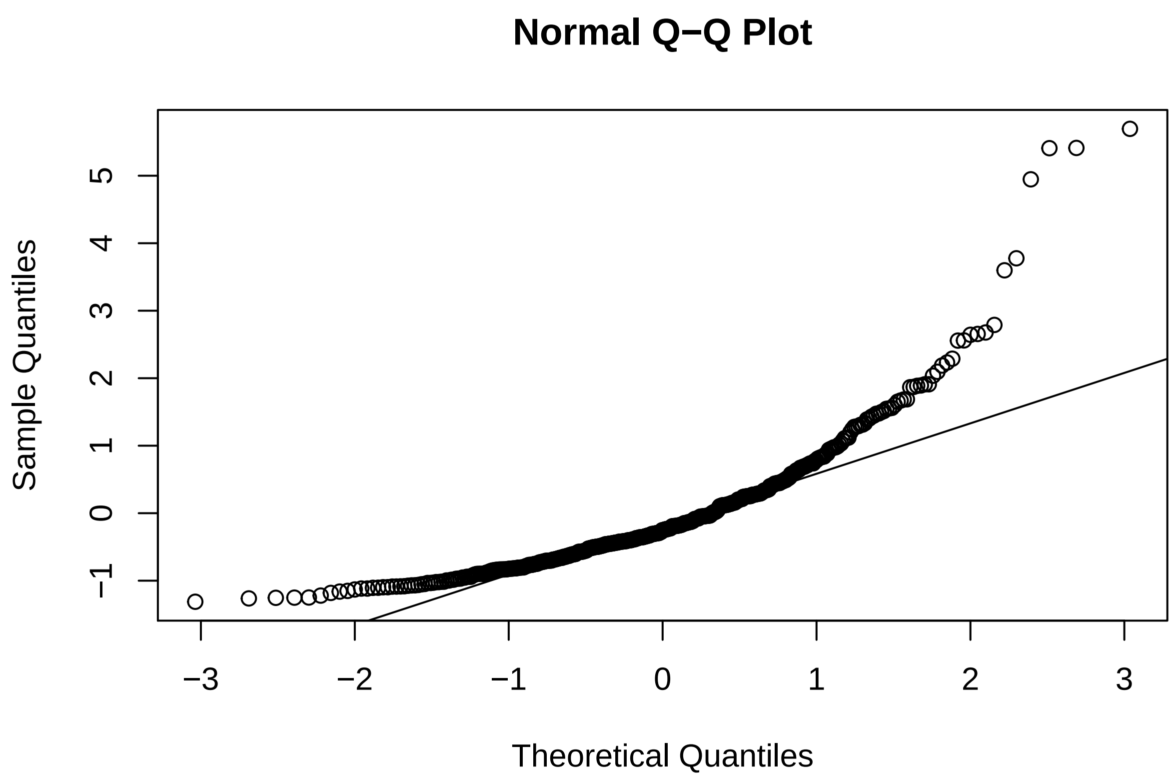


Regression (circular statistics) of release time vs transit time
omitted top 12 delay times (>72hrs)
Transit time from release to array
100
80
60
40
5
10
15
20
Release time (hours of day)



Diagnostics of extended cosine model, no outliers:

```
##
##  Shapiro-Wilk normality test
##
## data:  delay4resid
## W = 0.82682, p-value < 2.2e-16
```




```
##
##  Bartlett test of homogeneity of variances
##
## data:  delay4resid and fl.df4$Rel.hrDay
## Bartlett's K-squared = 226.84, df = 19, p-value < 2.2e-16
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  delay4resid and fl.df4$Rel.hrDay
## Fligner-Killeen:med chi-squared = 87.951, df = 19, p-value =
## 7.636e-11
```

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. No insight.

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 0.068226670
## 2        6 0.3442692 0.730643811
## 3        9 0.9718112 0.331144484
## 4       11 1.1539002 0.248541089
## 5       15 2.3837405 0.017137685
## 6       17 1.0828236 0.278886734
## 7       21 2.2563366 0.024049560
## 8       23 3.1685938 0.001531783
```

- 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.

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.