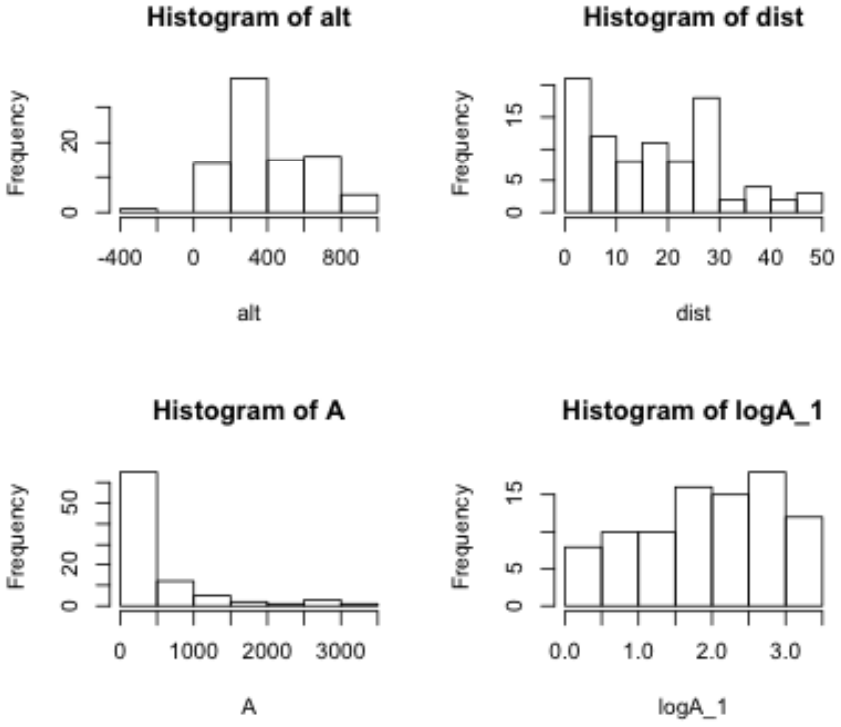
Preliminary Analysis of Medfly Data (with R code)

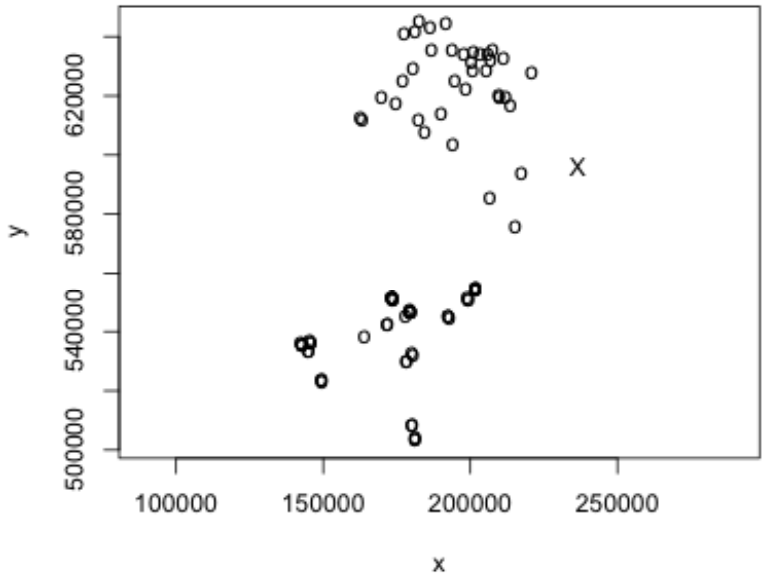
```
> # read in tab-delimited file with missing values
> m <- read.table(file="medfly.data.tab.txt", sep="\t", header=TRUE  )

> m[1:5,]
  trap        x        y south      alt dist loc_host    A  W
1   72 211353.7 632340.9     1 726.6240 29.5        1 2696 16
2   92 142408.4 535646.2     1 240.7184 30.0        1    7 18
3   97 142721.0 535212.8     1 240.4684 30.0        1  169 19
4  910 142770.8 535113.3     1 244.8809 30.0        0   32 19
5  911 142834.7 535013.8     1 246.4829 30.0        0    1  9
>
> attach(m)    # allows us to refer to the variables as x and not m$x

> logA_1 = log10(A + 1)  # dependent variable

> # exploratory analysis

> # marginal distributions of continuous variables
> par(mfrow=c(2,2)) # arrange plots in 2 x 2 window
> hist(alt)
> hist(dist)
> hist(A)
> hist(logA_1)
> par(mfrow=c(1,1)) # return to 1 x 1 window
```




```
> # is the negative value for alt an error?
```

locate this trap on a map with aspect (y/x ratio) = 1

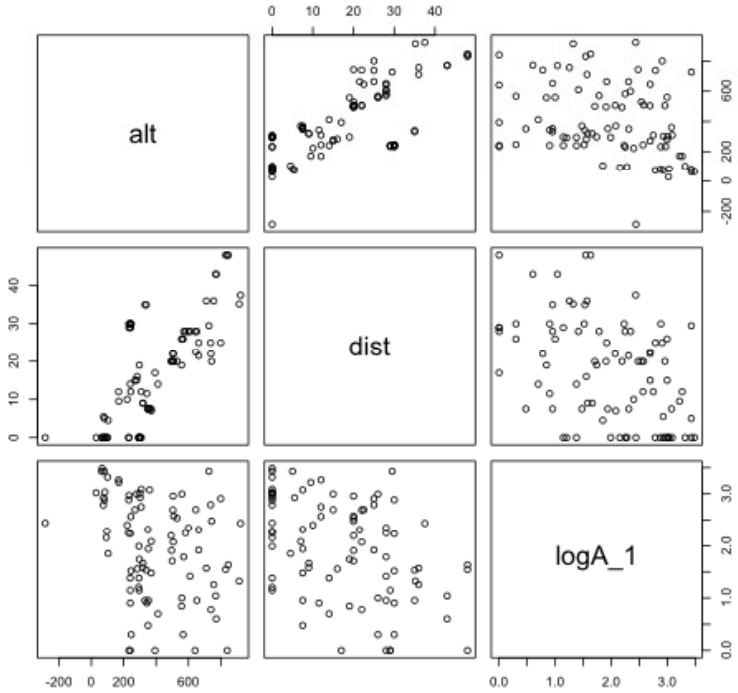
```
> plot(x, y, asp=1, pch=" ")       # set up the axes, but no points
> points(x[alt >= 0], y[alt >= 0], pch="o") # o if alt >= 0
> points(x[alt < 0], y[alt < 0], pch="X")   # X if alt < 0
```



> # the point is by the Dead Sea

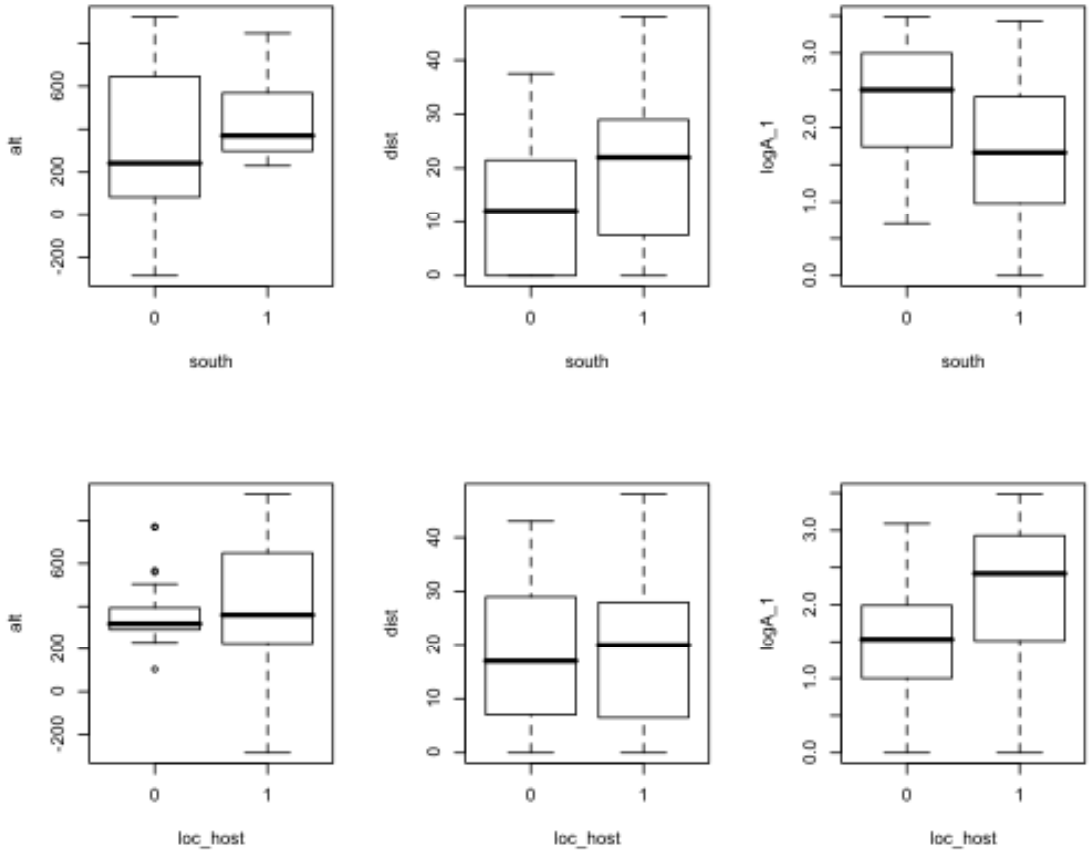
> # joint distribution of continuous variables: scatterplot matrix

```
> pairs(cbind(alt, dist, logA_1))
```



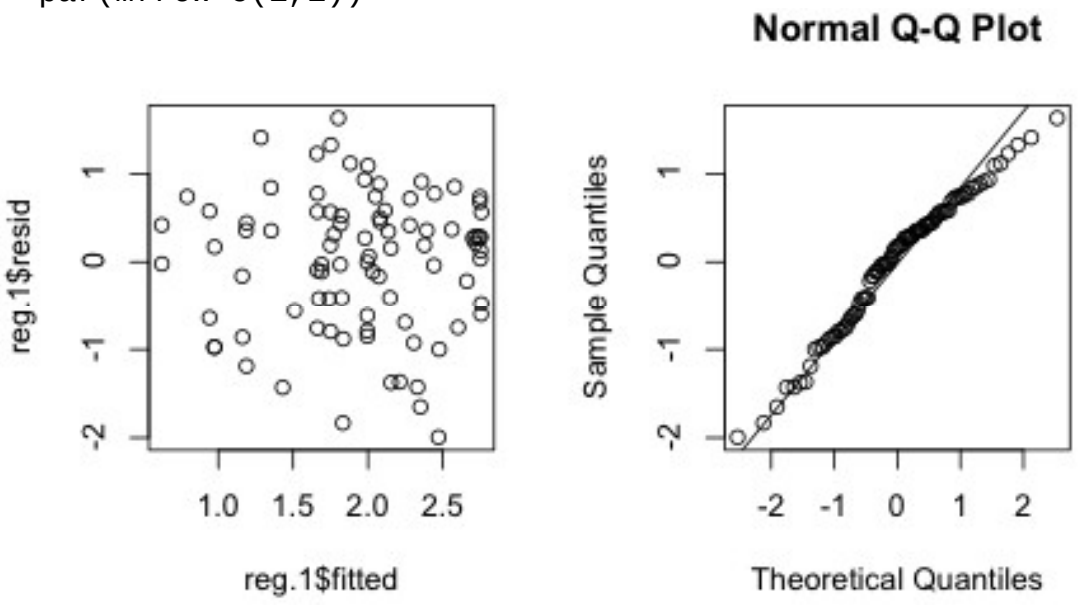
> # strong correlation between alt and dist

```
> # joint distribution of discrete variables
> table(south, loc_host)

     loc_host
south  0  1
    0  2 32
    1 27 28

> # should only use loc_host for traps in south; define appropriate
variable
> loc_host.south = loc_host * south

> # relation between continuous and categorical explanatory variables
> par(mfrow=c(2,3))
> boxplot(alt ~ south, xlab="south", ylab="alt")
> boxplot(dist ~ south, xlab="south", ylab="dist")
> boxplot(logA_1 ~ south, xlab="south", ylab="logA_1")
>
> boxplot(alt ~ loc_host, xlab="loc_host", ylab="alt")
> boxplot(dist ~ loc_host, xlab="loc_host", ylab="dist")
> boxplot(logA_1 ~ loc_host, xlab="loc_host", ylab="logA_1")
> par(mfrow=c(1,1))
```



> in south, dist tends to be greater and logA_1 lower
> presence of loc_host does not depend on dist
> when loc_host is present, logA_1 tends to be higher

```
> # fit linear regression model
> reg.1 = lm(logA_1 ~ south + loc_host.south + alt + dist)

> summary(reg.1)

Call:
lm(formula = logA_1 ~ south + loc_host.south + alt + dist)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9941 -0.6001  0.1705  0.5624  1.6287

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.7320340  0.1807736  15.113  < 2e-16 ***
south          -0.8103601  0.2066947  -3.921 0.000180 ***
loc_host.south  0.7214667  0.2210740   3.263 0.001593 **
alt             0.0002521  0.0005232   0.482 0.631178
dist           -0.0347156  0.0094032  -3.692 0.000395 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7955 on 84 degrees of freedom
Multiple R-squared: 0.3387,   Adjusted R-squared: 0.3072
F-statistic: 10.76 on 4 and 84 DF,  p-value: 4.359e-07

> # the hypothesis is "proved"!
>
> # diagnostic plots (resid vs pred, npp of resids)
> par(mfrow=c(1,2))
> plot(reg.1$fitted, reg.1$resid)
> qqnorm(reg.1$resid)
> qqline(reg.1$resid)
> par(mfrow=c(1,1))
```




```
> OK
```