```
> #Graduate Project - GLM, Count Data and the Poisson Regression
>
> #For my graduate (doctoral) project, I will be measuring the signaling
(clicking) count
> #data of the treehopper species Umbonia crassicornis under various
conditions and
> #contexts and levels of perturbation.


>
> #GLM/MULTIPLE REGRESSION (also see Biostats 380-410)
> #The response variable (dependent) will be number of clicks (Y). The first
> #independent variable will be time (X1), in seconds. The second independent
variable
> #will be the factor Plant ID (X2), with levels (plants) A, B, C, D, F, H, J
and K.
> #The third independent variable will be factor sex (X3), with levels M and
F. The fourth
> #independent variable will be factor level of perturbation (X4), with
factors light,
> #medium and heavy. The final independent variable will be date of recording
(X5).
>
> data1=read.table("Preliminary Click Data.txt", header=T)

> data1
   click_number  time plant_ID sex stimulus_strength       date
1            69 169.1        A   F              light  1/28/2013
2            27 107.0        A   F              light  1/28/2013
3            24  52.9        A   F              light  1/28/2013
4             0   8.2        B   F              light  1/31/2013
5            71  30.7        B   F              light  1/31/2013
6           202  91.0        B   F             medium  1/31/2013
7            17   6.4        B   F             medium  1/31/2013
8            68  33.1        B   F             medium  1/31/2013
9            56  26.4        B   F              heavy  1/31/2013
10           61  28.9        B   F              heavy  1/31/2013
11            0 101.3        B   M              light  3/27/2013
12            0 164.0        B   M             medium  3/27/2013
13            0 445.9        B   M              heavy  3/27/2013
14          169 186.0        B   F             medium   1/4/2014
15           44 144.0        B   F             medium   1/4/2014
16          140 323.0        B   F             medium   1/4/2014
17           35 160.0        B   F              heavy   1/4/2014
18           16 238.0        B   F              heavy   1/4/2014
19          330 505.0        B   F              heavy   1/4/2014
20            5 121.2        B   F              light   1/8/2014
21            0  38.0        C   F              light   4/3/2013
22            0  66.0        C   F             medium   4/3/2013
23            1  54.4        C   F             medium   4/3/2013
24          120 138.5        C   F             medium   4/3/2013
25          124 184.0        C   F             medium  4/12/2013
26            3 128.0        C   M             medium  4/12/2013
27          487 230.0        C   F              heavy  4/12/2013
28          159 226.0        C   F              heavy  4/12/2013
29          938 516.5        C   F              heavy  4/29/2013
30          437 314.0        C   F              heavy  4/29/2013
31           13 314.2        C   M             medium  8/21/2013
32            0  48.2        D   F              light   1/4/2014
33            0  74.0        D   F             medium   1/4/2014
34            1 140.5        D   F              heavy   1/4/2014
35            1 124.0        F   M              heavy  1/28/2013
36            2  83.8        F   M              heavy  1/28/2013
37            0  41.0        F   M              heavy  1/28/2013
```

```
38            0  63.5        F   F             light 11/25/2013
39            2 149.0        F   F            medium 11/25/2013
40            0  66.0        F   F             heavy 11/25/2013
41            3 206.0        F   F             heavy 11/25/2013
42            3 150.0        F   F             heavy 11/25/2013
43            9  62.0        H   F            medium   2/5/2013
44            3  37.5        H   F             heavy   2/5/2013
45          165  68.3        J   F             light  2/11/2013
46          397 200.5        J   F            medium  2/11/2013
47          287 235.0        J   F            medium  2/11/2013
48           67 163.0        J   F             light  2/25/2013
49            0  24.5        K   F             light   2/5/2013
50          134 117.0        K   F            medium   2/5/2013

> attach(data1)

> Y=click_number

> X1=time

> X2=factor(plant_ID)

> X3=factor(sex)

> X4=factor(stimulus_strength)

> X5=factor(date)

> FM=lm(Y~X1+X2+X3+X4+X5)

> FM

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5)

Coefficients:
 (Intercept)             X1            X2B            X2C            X2D
X2F            X2H
    -72.4390         0.7518        83.8924        28.5818        86.1066
220.9257        25.7973
         X2J            X2K            X3M       X4light      X4medium
X51/31/2013    X51/4/2014
    -13.0953        56.0132      -209.8348        29.9932        30.4808
10.6388      -99.3236
  X51/8/2014  X511/25/2013   X52/11/2013   X52/25/2013    X52/5/2013
X53/27/2013   X54/12/2013
   -127.5630     -254.3831      211.9661            NA            NA
NA       129.9826
 X54/29/2013    X54/3/2013   X58/21/2013
    419.1780      -12.0530            NA


> summary(FM)

Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5)

Residuals:
     Min       1Q   Median       3Q      Max
-174.382  -31.708   -1.554   27.793  227.964

Coefficients: (4 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -72.4390    79.8274  -0.907 0.371403
```

```
X1              0.7518     0.1975   3.807 0.000647 ***
X2B            83.8924   130.3884   0.643 0.524850
X2C            28.5818   151.7886   0.188 0.851909
X2D            86.1066   140.1656   0.614 0.543635
X2F           220.9257   137.6381   1.605 0.118944
X2H            25.7973    92.4716   0.279 0.782178
X2J           -13.0953   100.7522  -0.130 0.897454
X2K            56.0132    82.7885   0.677 0.503855
X3M          -209.8348   103.4823  -2.028 0.051550 .
X4light        29.9932    49.2562   0.609 0.547160
X4medium       30.4808    36.6221   0.832 0.411814
X51/31/2013    10.6388   130.0774   0.082 0.935358
X51/4/2014    -99.3236   120.7143  -0.823 0.417115
X51/8/2014   -127.5630   148.0443  -0.862 0.395713
X511/25/2013 -254.3831   124.0258  -2.051 0.049084 *
X52/11/2013   211.9661   103.7879   2.042 0.049997 *
X52/25/2013         NA         NA      NA       NA
X52/5/2013          NA         NA      NA       NA
X53/27/2013         NA         NA      NA       NA
X54/12/2013   129.9826   128.1100   1.015 0.318402
X54/29/2013   419.1780   146.5413   2.860 0.007631 **
X54/3/2013    -12.0530   151.5254  -0.080 0.937128
X58/21/2013         NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 86.78 on 30 degrees of freedom
Multiple R-squared:  0.8428,   Adjusted R-squared:  0.7433
F-statistic: 8.467 on 19 and 30 DF,  p-value: 1.846e-07


> anova(FM)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 558277  558277 74.1398 1.321e-09 ***
X2         7 161535   23076  3.0646 0.0147654 *
X3         1 102165  102165 13.5676 0.0009041 ***
X4         2  23557   11778  1.5642 0.2258469
X5         8 365809   45726  6.0725 0.0001157 ***
Residuals 30 225902    7530
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> #After creating the linear model, the summary() function give the p-values
for each
> #independent variable as well as an overall p-value, F stat and regression
coefficient
> #Shown by the low p-value and high regression coefficent, we can see that
there is an
> #overall relationship between Y and at least one of the X's. Therefore, we
reject the
> #null and accept the alternative hypothesis that at least one beta is not
equal to 0
>
> #Now lets compare our full model to reduced models to see which X's show an
association
> #with Y.
>
> step(FM, direction="backward")
Start:  AIC=460.79
Y ~ X1 + X2 + X3 + X4 + X5
```

```
       Df Sum of Sq    RSS    AIC
- X4    2      5401 231303 457.97
- X2    3     21539 247441 459.35
<none>              225902 460.79
- X3    1     30961 256863 465.21
- X1    1    109132 335034 478.50
- X5    8    365809 591711 492.94

Step:  AIC=457.97
Y ~ X1 + X2 + X3 + X5

       Df Sum of Sq    RSS    AIC
- X2    3     17102 248405 455.54
<none>              231303 457.97
- X3    1     27637 258940 461.62
- X1    1    118409 349712 476.64
- X5    8    383965 615268 490.89

Step:  AIC=455.54
Y ~ X1 + X3 + X5

       Df Sum of Sq    RSS    AIC
<none>              248405 455.54
- X3    1     13211 261616 456.13
- X1    1    157836 406242 478.13
- X5   12    493474 741879 486.25

Call:
lm(formula = Y ~ X1 + X3 + X5)

Coefficients:
 (Intercept)            X1           X3M   X51/31/2013    X51/4/2014
X51/8/2014  X511/25/2013
    -10.2490        0.7203      -77.2301       54.9848      -53.6392      -
72.0503      -79.5560
 X52/11/2013   X52/25/2013    X52/5/2013   X53/27/2013   X54/12/2013
X54/29/2013    X54/3/2013
    172.2881      -40.1585        3.3515      -83.2780       84.5106
398.6480      -12.9646
 X58/21/2013
   -125.8365


> #shown by the step() function and AIC values, we can see that the optimal
model to use
> #is a Reduced Model with only X1, X3 and X5 as independent variables. In
other words,
> #the variables of plant ID and stimulus strength show no relationship with
click count,
> #and the betas of these X's are not significantly different from 0
(hopefully this wont
> #be the case when i have usable data!) Since our reduced model is optimal,
lets assign
> #it to the variable RM.
>
> RM=lm(Y~X1+X3+X5)

> RM

Call:
lm(formula = Y ~ X1 + X3 + X5)

Coefficients:
```

```
 (Intercept)            X1           X3M   X51/31/2013    X51/4/2014
X51/8/2014  X511/25/2013
    -10.2490        0.7203      -77.2301       54.9848      -53.6392      -
72.0503      -79.5560
 X52/11/2013   X52/25/2013    X52/5/2013   X53/27/2013   X54/12/2013
X54/29/2013    X54/3/2013
    172.2881      -40.1585        3.3515      -83.2780       84.5106
398.6480      -12.9646
 X58/21/2013
   -125.8365


> #Now lets compare the FM and the RM with the anova() function
>
> anova(FM,RM)
Analysis of Variance Table

Model 1: Y ~ X1 + X2 + X3 + X4 + X5
Model 2: Y ~ X1 + X3 + X5
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     30 225902
2     35 248405 -5    -22504 0.5977 0.7019

> #Shown by the high p-value for Model 2, we cannot reject the null. This
means that
> #the more parsimonious (and thus optimal) model is the reduced model which
excludes
> #X2 (plant ID) and X4 (stimulus strength).
>
> #Now to individually compare our 3 favored variables to Y, lets make
individual linear
> #models for them
> LM1=lm(Y~X1)

> LM3=lm(Y~X3)

> LM5=lm(Y~X5)

> #Here's how to make dataframes of the values of Yhat and e for the 3
favored variables (RM).
>
> Yhat1=fitted(LM1)          #---------> Yhat = line of best fit

> Yhat3=fitted(LM3)

> Yhat5=fitted(LM5)

> e1=residuals(LM1)          #-------> error

> e3=residuals(LM3)

> e5=residuals(LM5)

> RESULTS1=data.frame(X1,Y,Yhat1,e1)

> RESULTS1
      X1   Y       Yhat1           e1
1  169.1  69 116.0599844  -47.0599844
2  107.0  27  60.7440568  -33.7440568
3   52.9  24  12.5541747   11.4458253
4    8.2   0 -27.2626041   27.2626041
5   30.7  71  -7.2206013   78.2206013
6   91.0 202  46.4919660  155.5080340
7    6.4  17 -28.8659643   45.8659643
```

```
8   33.1  68  -5.0827877   73.0827877
9   26.4  56 -11.0508507   67.0508507
10  28.9  61  -8.8239616   69.8239616
11 101.3   0  55.6667495  -55.6667495
12 164.0   0 111.5171304 -111.5171304
13 445.9   0 362.6211558 -362.6211558
14 186.0 169 131.1137553   37.8862447
15 144.0  44  93.7020169  -49.7020169
16 323.0 140 253.1472831 -113.1472831
17 160.0  35 107.9541077  -72.9541077
18 238.0  16 177.4330505 -161.4330505
19 505.0 330 415.2648163  -85.2648163
20 121.2   5  73.3927874  -68.3927874
21  38.0   0  -0.7180849    0.7180849
22  66.0   0  24.2230741  -24.2230741
23  54.4   1  13.8903082  -12.8903082
24 138.5 120  88.8028607   31.1971393
25 184.0 124 129.3322440   -5.3322440
26 128.0   3  79.4499260  -76.4499260
27 230.0 487 170.3070051  316.6929949
28 226.0 159 166.7439824   -7.7439824
29 516.5 938 425.5085066  512.4914934
30 314.0 437 245.1304820  191.8695180
31 314.2  13 245.3086331 -232.3086331
32  48.2   0   8.3676230   -8.3676230
33  74.0   0  31.3491195  -31.3491195
34 140.5   1  90.5843720  -89.5843720
35 124.0   1  75.8869033  -74.8869033
36  83.8   2  40.0785251  -38.0785251
37  41.0   0   1.9541821   -1.9541821
38  63.5   0  21.9961849  -21.9961849
39 149.0   2  98.1557953  -96.1557953
40  66.0   0  24.2230741  -24.2230741
41 206.0   3 148.9288689 -145.9288689
42 150.0   3  99.0465509  -96.0465509
43  62.0   9  20.6600514  -11.6600514
44  37.5   3  -1.1634627    4.1634627
45  68.3 165  26.2718121  138.7281879
46 200.5 397 144.0297126  252.9702874
47 235.0 287 174.7607835  112.2392165
48 163.0  67 110.6263747  -43.6263747
49  24.5   0 -12.7432865   12.7432865
50 117.0 134  69.6516136   64.3483864
```

```
> RESULTS3=data.frame(X3,Y,Yhat3,e3)
```

```
> RESULTS3
   X3   Y    Yhat3          e3
1   F  69 111.2143  -42.214286
2   F  27 111.2143  -84.214286
3   F  24 111.2143  -87.214286
4   F   0 111.2143 -111.214286
5   F  71 111.2143  -40.214286
6   F 202 111.2143   90.785714
7   F  17 111.2143  -94.214286
8   F  68 111.2143  -43.214286
9   F  56 111.2143  -55.214286
10  F  61 111.2143  -50.214286
11  M   0   2.3750   -2.375000
12  M   0   2.3750   -2.375000
13  M   0   2.3750   -2.375000
14  F 169 111.2143   57.785714
15  F  44 111.2143  -67.214286
16  F 140 111.2143   28.785714
```

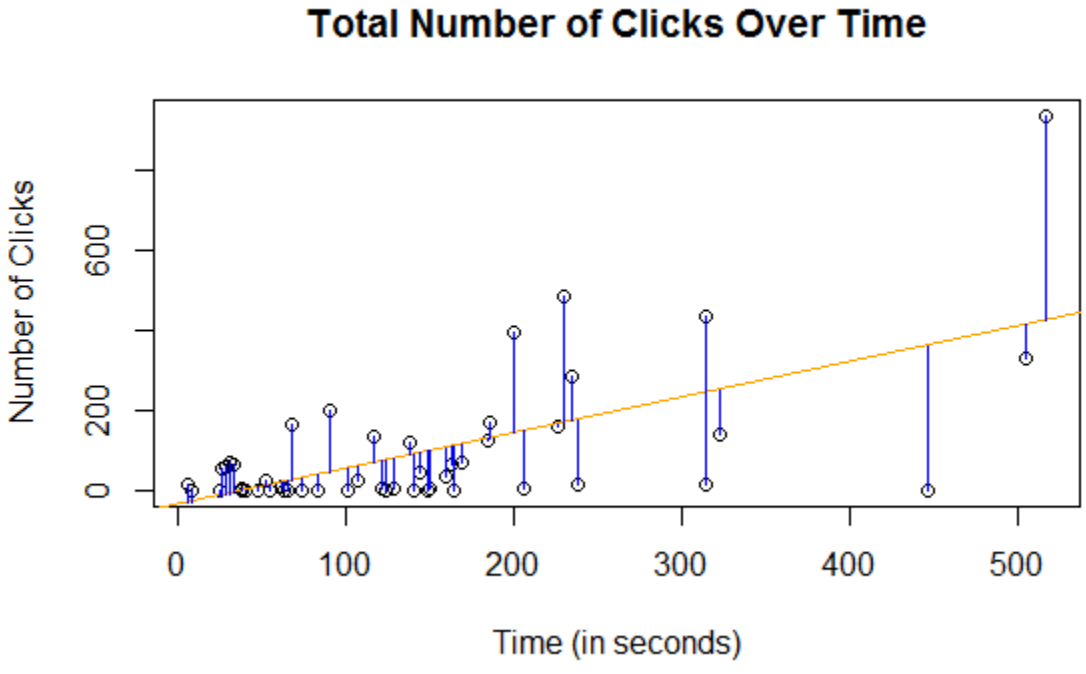
```
17  F  35 111.2143  -76.214286
18  F  16 111.2143  -95.214286
19  F 330 111.2143  218.785714
20  F   5 111.2143 -106.214286
21  F   0 111.2143 -111.214286
22  F   0 111.2143 -111.214286
23  F   1 111.2143 -110.214286
24  F 120 111.2143    8.785714
25  F 124 111.2143   12.785714
26  M   3   2.3750    0.625000
27  F 487 111.2143  375.785714
28  F 159 111.2143   47.785714
29  F 938 111.2143  826.785714
30  F 437 111.2143  325.785714
31  M  13   2.3750   10.625000
32  F   0 111.2143 -111.214286
33  F   0 111.2143 -111.214286
34  F   1 111.2143 -110.214286
35  M   1   2.3750   -1.375000
36  M   2   2.3750   -0.375000
37  M   0   2.3750   -2.375000
38  F   0 111.2143 -111.214286
39  F   2 111.2143 -109.214286
40  F   0 111.2143 -111.214286
41  F   3 111.2143 -108.214286
42  F   3 111.2143 -108.214286
43  F   9 111.2143 -102.214286
44  F   3 111.2143 -108.214286
45  F 165 111.2143   53.785714
46  F 397 111.2143  285.785714
47  F 287 111.2143  175.785714
48  F  67 111.2143  -44.214286
49  F   0 111.2143 -111.214286
50  F 134 111.2143   22.785714
```

```
> RESULTS5=data.frame(X5,Y,Yhat5,e5)
```

```
> RESULTS5
           X5   Y          Yhat5             e5
1   1/28/2013  69  2.050000e+01  4.850000e+01
2   1/28/2013  27  2.050000e+01  6.500000e+00
3   1/28/2013  24  2.050000e+01  3.500000e+00
4   1/31/2013   0  6.785714e+01 -6.785714e+01
5   1/31/2013  71  6.785714e+01  3.142857e+00
6   1/31/2013 202  6.785714e+01  1.341429e+02
7   1/31/2013  17  6.785714e+01 -5.085714e+01
8   1/31/2013  68  6.785714e+01  1.428571e-01
9   1/31/2013  56  6.785714e+01 -1.185714e+01
10  1/31/2013  61  6.785714e+01 -6.857143e+00
11  3/27/2013   0 -5.888623e-13  5.888623e-13
12  3/27/2013   0  2.957634e-13 -2.957634e-13
13  3/27/2013   0  2.886580e-13 -2.886580e-13
14   1/4/2014 169  8.166667e+01  8.733333e+01
15   1/4/2014  44  8.166667e+01 -3.766667e+01
16   1/4/2014 140  8.166667e+01  5.833333e+01
17   1/4/2014  35  8.166667e+01 -4.666667e+01
18   1/4/2014  16  8.166667e+01 -6.566667e+01
19   1/4/2014 330  8.166667e+01  2.483333e+02
20   1/8/2014   5  5.000000e+00 -2.309264e-14
21   4/3/2013   0  3.025000e+01 -3.025000e+01
22   4/3/2013   0  3.025000e+01 -3.025000e+01
23   4/3/2013   1  3.025000e+01 -2.925000e+01
24   4/3/2013 120  3.025000e+01  8.975000e+01
25  4/12/2013 124  1.932500e+02 -6.925000e+01
```

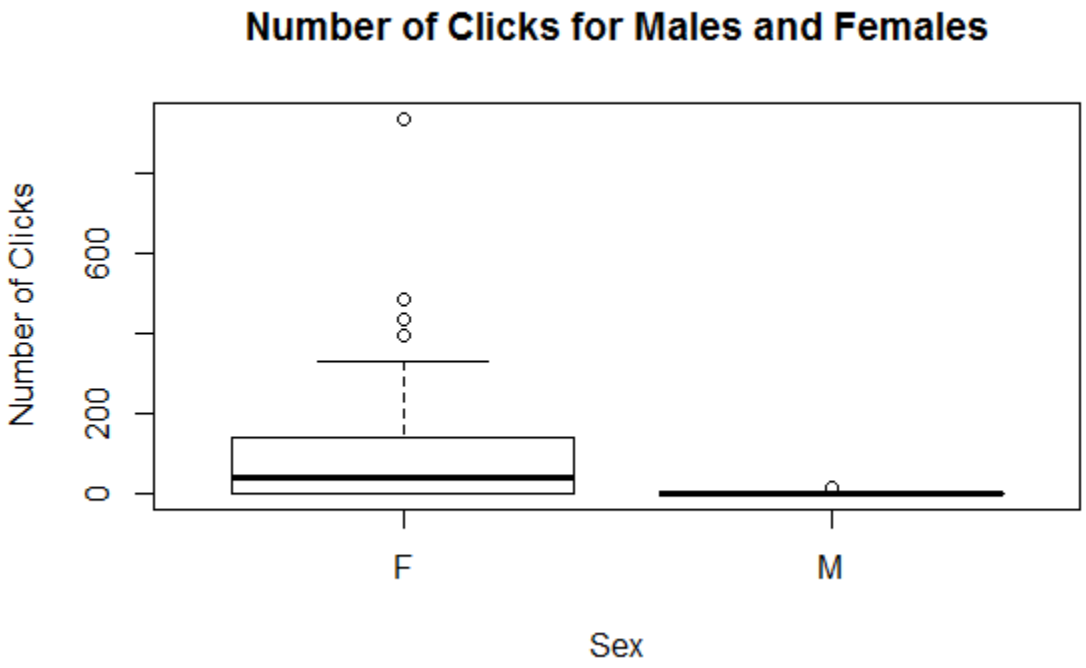
```
26  4/12/2013   3  1.932500e+02 -1.902500e+02
27  4/12/2013 487  1.932500e+02  2.937500e+02
28  4/12/2013 159  1.932500e+02 -3.425000e+01
29  4/29/2013 938  6.875000e+02  2.505000e+02
30  4/29/2013 437  6.875000e+02 -2.505000e+02
31  8/21/2013  13  1.300000e+01  2.664535e-15
32   1/4/2014   0  8.166667e+01 -8.166667e+01
33   1/4/2014   0  8.166667e+01 -8.166667e+01
34   1/4/2014   1  8.166667e+01 -8.066667e+01
35  1/28/2013   1  2.050000e+01 -1.950000e+01
36  1/28/2013   2  2.050000e+01 -1.850000e+01
37  1/28/2013   0  2.050000e+01 -2.050000e+01
38 11/25/2013   0  1.600000e+00 -1.600000e+00
39 11/25/2013   2  1.600000e+00  4.000000e-01
40 11/25/2013   0  1.600000e+00 -1.600000e+00
41 11/25/2013   3  1.600000e+00  1.400000e+00
42 11/25/2013   3  1.600000e+00  1.400000e+00
43   2/5/2013   9  3.650000e+01 -2.750000e+01
44   2/5/2013   3  3.650000e+01 -3.350000e+01
45  2/11/2013 165  2.830000e+02 -1.180000e+02
46  2/11/2013 397  2.830000e+02  1.140000e+02
47  2/11/2013 287  2.830000e+02  4.000000e+00
48  2/25/2013  67  6.700000e+01  8.881784e-16
49   2/5/2013   0  3.650000e+01 -3.650000e+01
50   2/5/2013 134  3.650000e+01  9.750000e+01
```

```
> #Heres how to plot the regression plots for the 3 favored variables (RM)
compared to Y
> plot(X1,Y, main="Total Number of Clicks Over Time", xlab="Time (in
seconds)", ylab="Number of Clicks")

> abline(LM1,col="orange")

> segments(X1,Yhat1,X1,Y,col="blue")
```



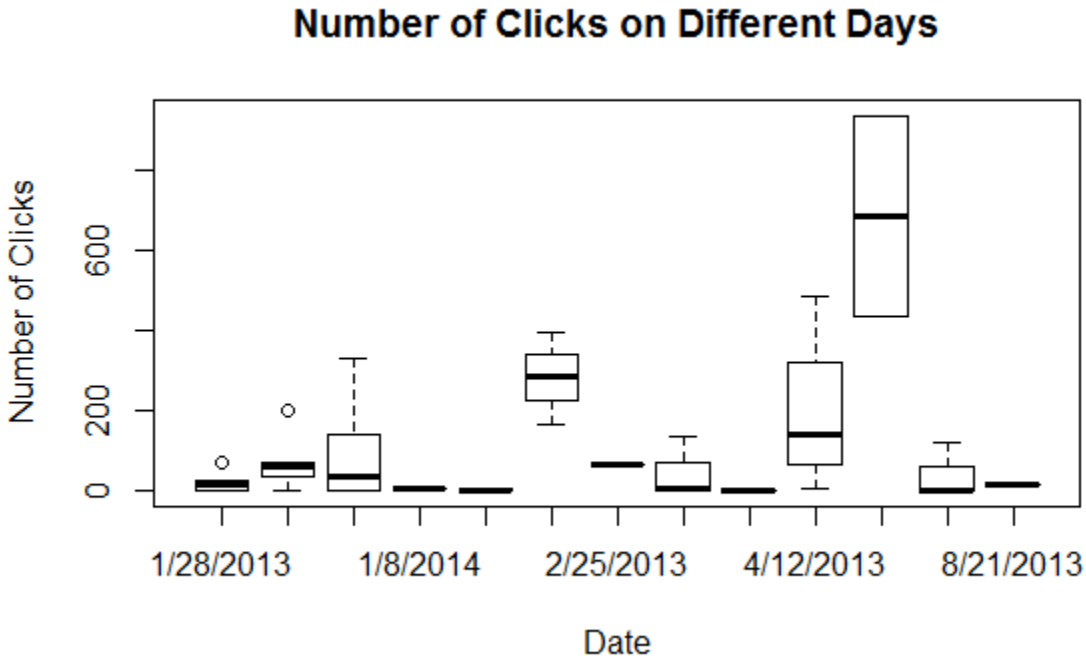

```
> #Shown by Graph #1, the number of clicks was greater with longer
recordings, as expected
```

```
> plot(X3,Y, main="Number of Clicks for Males and Females", xlab="Sex",
ylab="Number of Clicks")
```




```
> #Graph #2 shows that females clicked more than males
```

```
> plot(X5,Y, main="Number of Clicks on Different Days", xlab="Date",
ylab="Number of Clicks")
```



```
> #Shown by Graph #3, the number of clicks varied on different days
>
>
> #-------------------------------------------------------------------------
---------


>
> #GLM 040 - Poisson Regression for Count Data
>
> #Poisson Regression and related Negative Binomial Regression, are
generalized linear
> #models typically involving count data as the response variable. As such,
non-negative
> #values are not expected (nor allowed), and responses are considered to be
integers.
> #The distribution of response values, as in a histogram for each set of
independent
> #variables indicated by index i, may look quite different for small versus
large expected
> #values (means). For the Poisson distribution, the distribution around
small expected
> #values are highly asymmetric, whereas for larger values symmetric. In
addition, variance
> #increases with expected value.
>
> #Modeling with a Poisson Regression will be more accurate for this type of
count data.
> #Lets reassign both the full and reduced model here:
>
> library(nlme)

> library(MASS)

> FM1=glm(Y~X1*X2*X3*X4*X5, family=poisson)
Error in glm.fit(x = c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,  :
  NA/NaN/Inf in 'x'
In addition: Warning message:
step size truncated due to divergence
```

#As these procedures are beyond the scope of this class, next semester's seminar should

#clarify these approaches so I can use them with my real data

#to be continued....