Chapter 18

Regression - hockey sticks, broken sticks, piecewise, change points

Contents

A simple regression analysis assumes that the change in response is the same across the range of X values. In some cases, a model where the slope changes in different parts of the X space may be biologically more realistic.

This chapter examines two cases of fitting regression lines with breaks in the slope. In the first case, the location of the change in slope is known in advance; the second cases also estimates the location of the change, also known as the change point problem.

The examples in this chapter look at cases with a single change point – the extension to multiple change points (both known and unknown) is straightforward. Similarly, the change from linear to quadratic lines is also straightforward.

A related method, a spline fit to the data, where a flexible curve is fit between (evenly) spaced knot points that is a like a non-parametric curve fit is explored in a different chapter.

18.1 Hockey-stick, piecewise, or broken-stick regression

In this section, the location of the change point is known. The statistical model is:

$$Y = \beta_0 + \beta_1(X) + \beta_2(X - C)^+ + \epsilon$$

where β_0 is the intercept, β_1 is the slope before the change point C, and β_2 is the DIFFERENCE in slope after the change point. The slope after the change point is $\beta_1 + \beta_2$. The variable $(X - C)^+$ is a derived variable which takes the value of 0 for values of X less than C and the values $X - C$ for values of X greater than C. This is usually created using a Formula Editor based on the actual data.

The hypothesis of interest is $H : \beta_2 = 0$ which indicates no change in slope between $X < C$ and $X > C$.

Because the value of C is specified in advance, ordinary least-squares can be used to fit the model. Most computer packages can easily fit this model.

18.1.1 Example: Nenana River Ice Breakup Dates

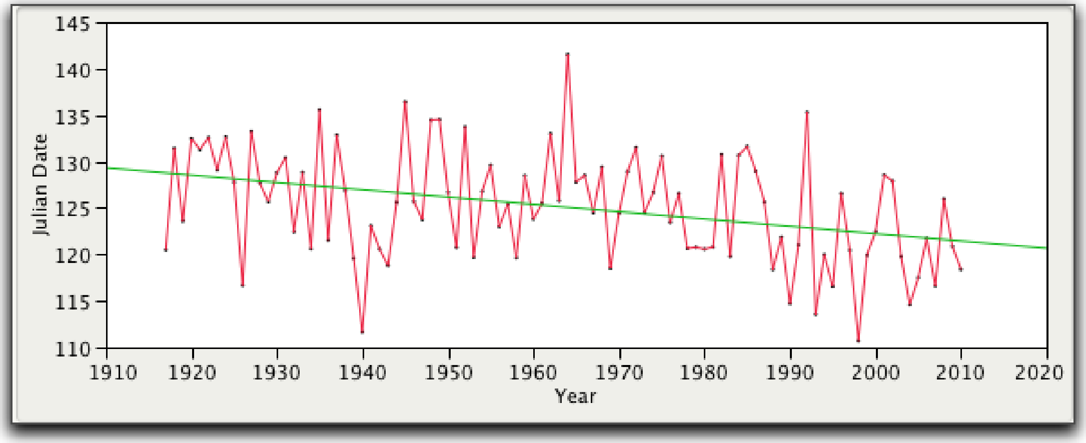
The Nenana river in the Interior of Alaska usually freezes over during October and November. The ice continues to grow throughout the winter accumulating an average maximum thickness of about 110 cm, depending upon winter weather conditions. The Nenana River Ice Classic competition began in 1917 when railroad engineers bet a total of 800 dollars, winner takes all, guessing the exact time (month, day, hour, minute) ice on the Nenana River would break up. Each year since then, Alaska residents have guessed at the timing of the river breakup. A tripod, connected to an on-shore clock with a string, is planted in two feet of river ice during river freeze-up in October or November. The following spring, the clock automatically stops when the tripod moves as the ice breaks up. The time on the clock is used as the river ice breakup time. Many factors influence the river ice breakup, such as air temperature, ice thickness, snow cover, wind, water temperature, and depth of water below the ice. Generally, the Nenana river ice breaks up in late April or early May (historically, April 20 to May 20). The time series of the Nenana river ice breakup dates can be used to investigate the effects of climate change in the region.

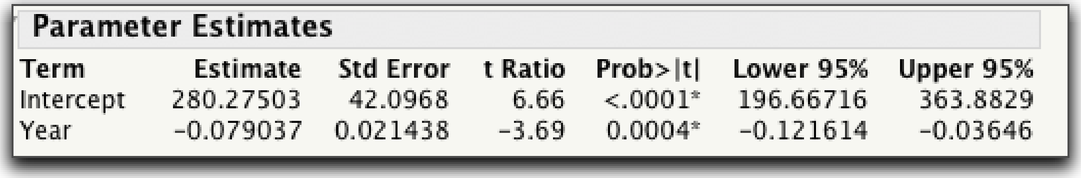
In 2010, the jackpot was almost $300,000 and the ice went out at 9:06 on 2010-04-29. In 2012, the jackpot was over $350,000 and the ice went out at 19:39 on 2012-04-23 - as reported at `http://www.cbc.ca/news/offbeat/story/2012/05/02/alaska-ice-contest.html`. The latest winner, Tommy Lee Waters, has also won twice before, but never has been a solo winner. Waters spent time drilling holes in the area to measure the thickness of the ice. Altogether he spent $5,000 on tickets for submitting guesses (he purchased every minute of the afternoon of 23 April) and spent an estimated 1,200 hours working out the math by hand. And, it was also his birthday! (What are the odds?) You too can use statistical methods to gain fame and fortune!

More details about the Ice Classic are available at `http://www.nenanaakiceclassic.com`.

The data are available in the *nenana.csv* data file available in the the Sample Program Library at `http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms`. It is also available in the *nenana.jmp* data file.

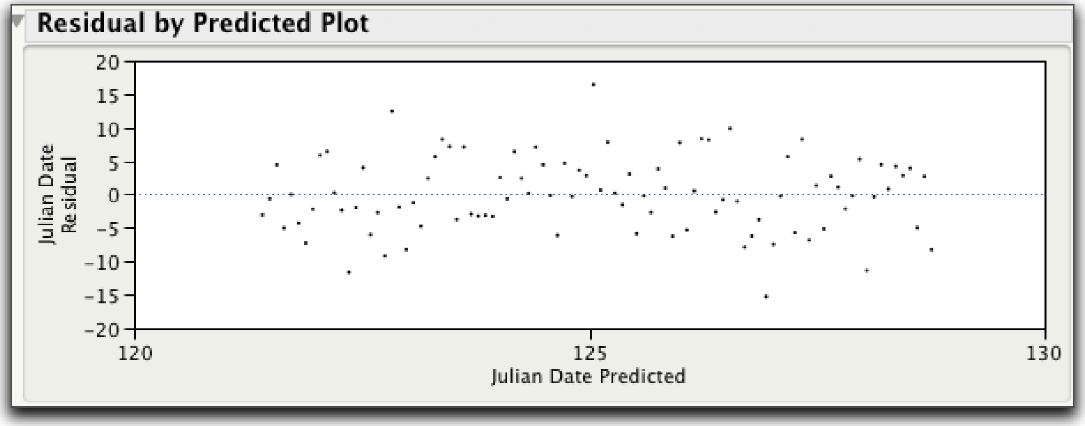
A simple regression line fit to the time of break up with year as the predictor show evidence of a decline over time (i.e. the time of breakup is tending to occur earlier) and there is no evidence of auto-correlation.

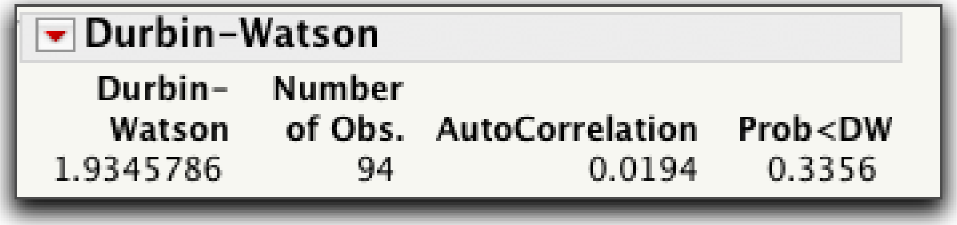




Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	280.27503	42.0968	6.66	<.0001*	196.66716	363.8829
Year	-0.079037	0.021438	-3.69	0.0004*	-0.121614	-0.03646





Durbin-Watson

Durbin-Watson	Number of Obs.	AutoCorrelation	Prob<DW
1.9345786	94	0.0194	0.3356

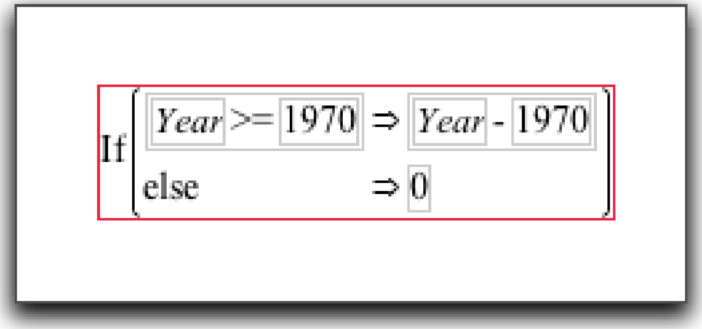
A closer inspection of the top graph gives the impression that until about 1970, the regression line was "flat" and only after 1970 did the time of breakup seem to decrease.

A broken stick model (separate slopes in the pre-1970 and the post-1970 eras) can be easily fit. The statistical model is:

$$JulianDate = \beta_0 + \beta_1(Year) + \beta_2(Year - 1970)^+ + \epsilon$$

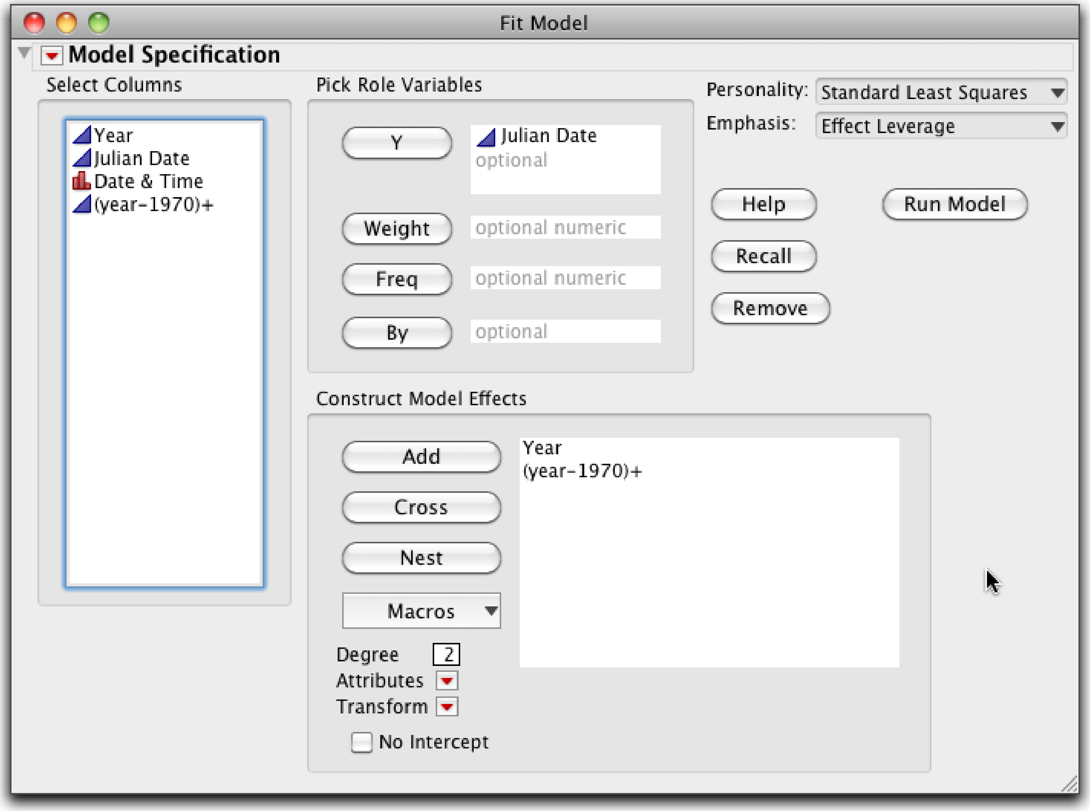
where $JulianDate$ is the date of breakup, $Year$ is the calendar year. The parameters to be estimated are β_0 the intercept, β_1 the change in the breakup prior to the change point, β_2 the change in slope after the change point which is assumed to happen in 1970. The term $(Year - 1970)^+$ takes the value 0 if the argument is negative (i.e. before 1970); and the value of the argument if it is positive.

To fit this model, we need to create a new variable that is zero for the pre-1970 period and equal to $(year - 1970)$ in the post 1970 period. This is easily created in *JMP* using the Formula Editor:



If
Year >= 1970 ⇒ Year - 1970
else ⇒ 0

The change point model with a known change point is then fit using standard multiple regression. In *JMP*. this is done using the *Analyze->Fit Model* platform:



Fit Model

Model Specification

Select Columns
Year
Julian Date
Date & Time
(year-1970)+

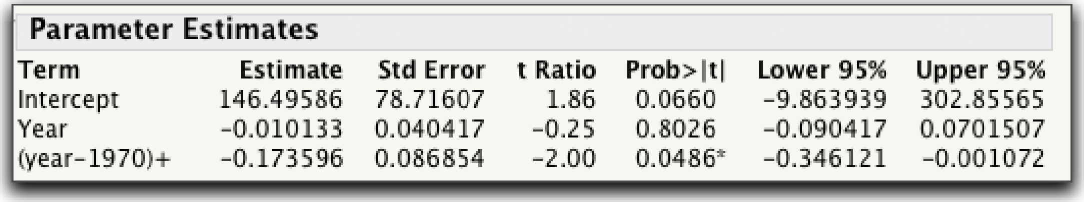
Pick Role Variables
Y: Julian Date
Weight: optional numeric
Freq: optional numeric
By: optional

Personality: Standard Least Squares
Emphasis: Effect Leverage

Construct Model Effects
Year
(year-1970)+

Degree 2
□ No Intercept

which gives the estimates:



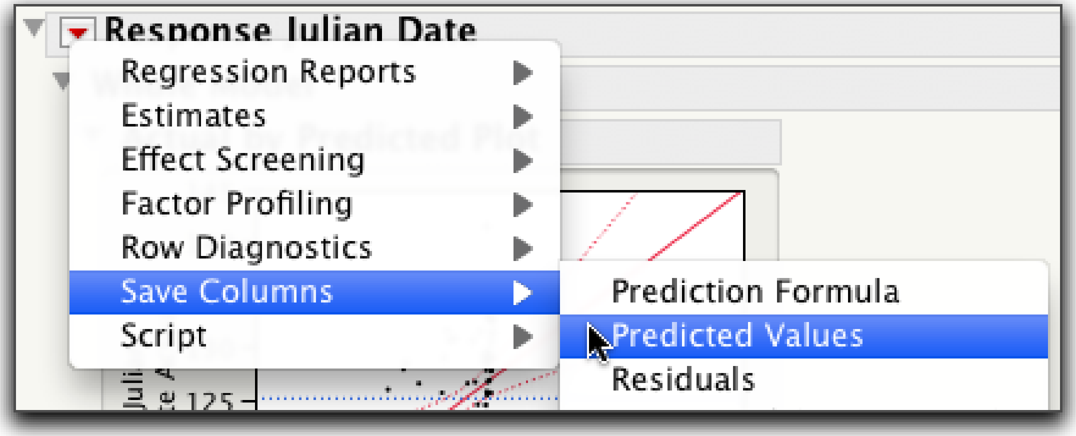
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Intercept	146.49586	78.71607	1.86	0.0660	-9.863939	302.85565
Year	-0.010133	0.040417	-0.25	0.8026	-0.090417	0.0701507
(year-1970)+	-0.173596	0.086854	-2.00	0.0486*	-0.346121	-0.001072

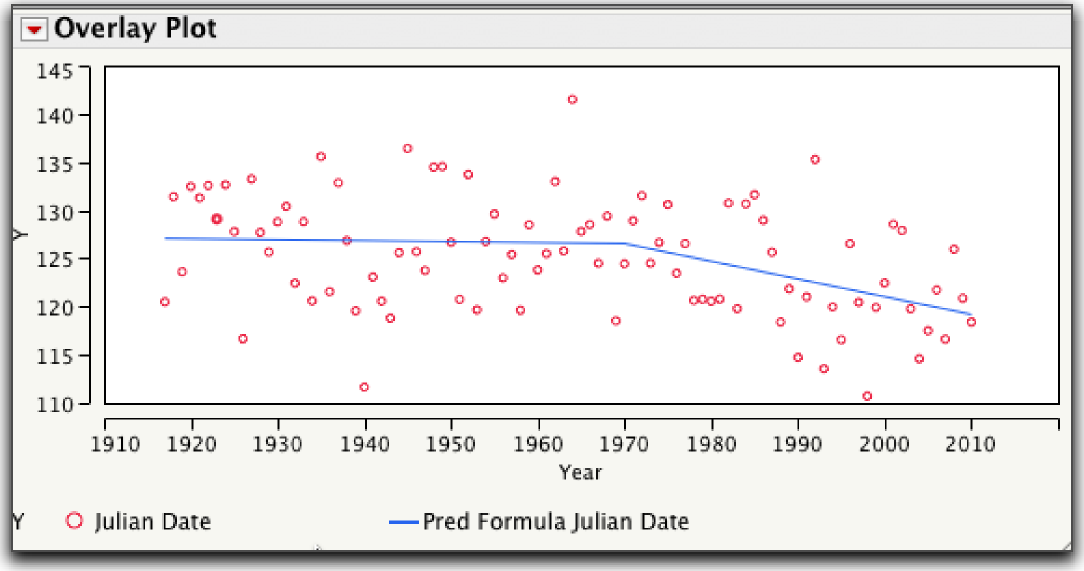
A test for differential slopes in the two eras is then equivalent to a test if $\beta_2 = 0$.

In this case the *p*-value for the β_2 coefficient (associated with the $(year - 1970)^+$ variable) is just under 0.05 providing some evidence of a different slope in the two eras.

A plot of the fitted model is obtained by saving the predicted values to the data table:



and then plotting the actual data and the fitted points on the same graph using the *Graph->Overlay* platform:



Confidence intervals for the MEAN response in a particular year (not likely of interest in this example) and for the individual responses in a particular year are generated in the usual way.

Note that the estimated slope for the pre-1970 era is not statistically different from 0. If you wanted to fit a model where the line was flat (i.e. the slope was 0) in the pre-1970 era, this is done by using only the $(year - 1970)^+$ variable. Many of the automatically generated plots look odd (e.g. all of the points appear to be replotted at 1970), the intercept has a different interpretation in the two models because $year = 0$ has a different definition in the two models, but if the fitted model is plotted against the original year variable everything works out properly. In this particular case, the two latter models give predicted lines that are almost identical. It is quite RARE that you would fit a line where the slope is known to be zero in practice, but see the next example for a case where it is sensible.

18.2 Searching for the change point

. In the previous section on segmented regression (also known as hockey-stick regression or broken-stick regression), the locations of the break are assumed to be known. In many cases, the location of the break is not known, and it is of interest to estimate the break point as well.

The problems of identifying changes at unknown times and of estimating the location of changes is known as "the change-point problem". Numerous methodological approaches have been implemented in examining change-point models. Maximum-likelihood estimation, Bayesian estimation, isotonic regression, piecewise regression, quasi-likelihood and non-parametric regression are among the methods which have been applied to resolving challenges in change- point problems. Grid-searching approaches have also been used to examine the change-point problem. A review of the literature especially as it applies to regression problem (as of 2008) is available at: `http://biostats.bepress.com/cgi/viewcontent.cgi?article=1075&context=cobra`.

The standard change-point problem in regression models consists of

- testing the null hypothesis that no change in regimes has taken place against the alternative that observations were generated by two (or possibly more) distinct regression equations, and
- estimating the two regimes that gave rise to the data.

There are two common models. First are models where the regression line is continuous at the break point, and models where the regression line can be discontinuous. In these notes, we only consider the continuous case.

This problem has a long history. A nice summary and treatment of the problem is available in

> Toms, J. D. and Lesperance, M L. (2003).
> Piecewise regression: A tool for identifying ecological thresholds.
> Ecology, 84, 2034-2041
> `http://dx.doi.org/10.1890/02-0472`.

The change point model starts with the broken-stick model seen earlier, i.e.

$$Y = \beta_0 + \beta_1(X) + \beta_2(X - C)^+ + \epsilon$$

where Y is the response variable, X is the covariate, and C is the change point, i.e. where the break occurs. This model is appropriate where there is an abrupt transition at the break point, but a smooth transition may be more realistic for some data. One drawback of this model is that convergence problems can occur in locating C when the data are sparse around the neighborhood of C.

Toms and Lesperance (2003) review the use of model with gentler transitions, e.g. the hyperbolic tangent model or the bent-cable model. The bent-cable regression model was recently developed by Chui, Lockhart and Routledge (2006, Bent-cable regression theory and application, Journal of the American Statistical Association, 101, 542-553). The bent-cable regression model fits a smooth transition between the two linear parts of the model. The latter is also applicable to regression models where the X variable is time and auto-correlation may be present[1].

The simple piece-wise linear model can be fit using the *Analyze->Modelling ->NonLinear* platform of *JMP*.

18.2.1 Change point model for the Nenana River Ice Breakup

Refer to the previous section about details on the Nenana River Ice Breakup contest. Rather than specifying a break point at 1970, we will fit the change point model to estimate the change point.

The data are available in the *nenana.csv* data table in the the Sample Program Library at `http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms`.

The statistical model is:

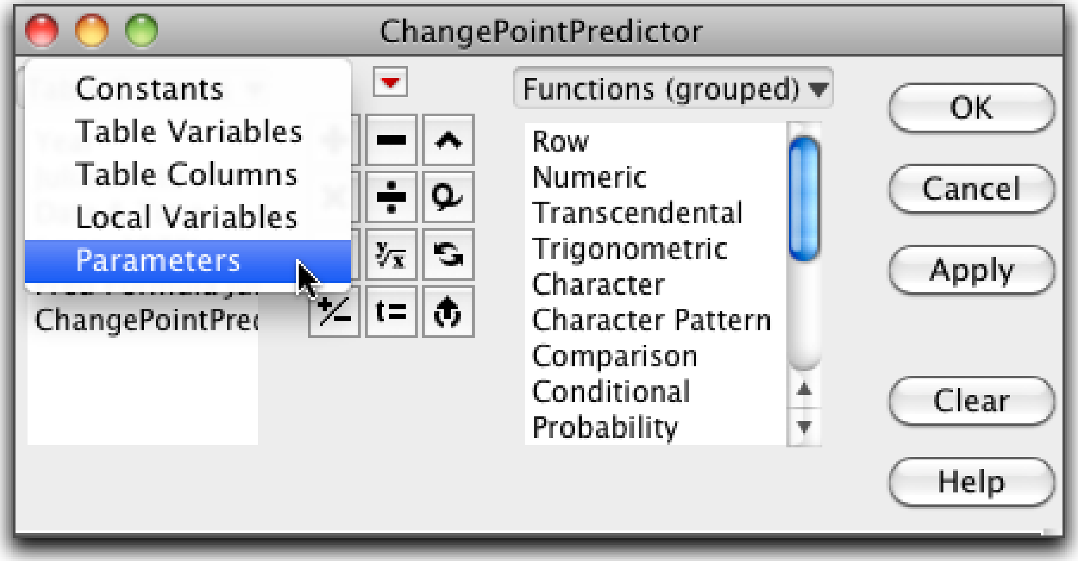
$$JulianDate = \beta_0 + \beta_1(Year) + \beta_2(Year - C)^+ + \epsilon$$

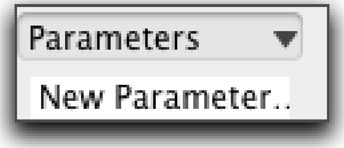
where $JulianDate$ is the date of breakup, $Year$ is the calendar year. The parameters to be estimated are β_0 the intercept, β_1 the change in the breakup prior to the change point, β_2 the change in slope after the breakup, and C the change point.

We first need to define the parameters of the model ($\beta_0, \beta_1, \beta_2, C$) and the predicted value in terms of the parameters of the model. We start by creating a new column in the data table *ChangePointPredictor* and start the Formula Editor.

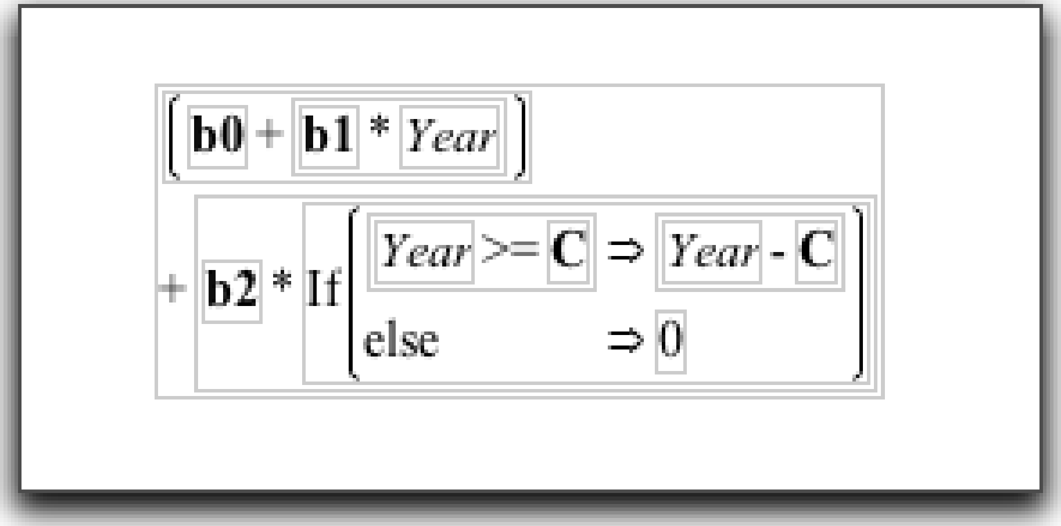
New parameters are defined (along with initial starting guesses), by using the drop-down menu in the top left of the formula editor:

[1] Chiu, G. S. and Lockhart, R. L. (2010). Bent-cable regression with auto-regressive noise. Canadian Journal of Statistics, 38, 386-407. `http://dx.doi.org/10.1002/cjs.10070`





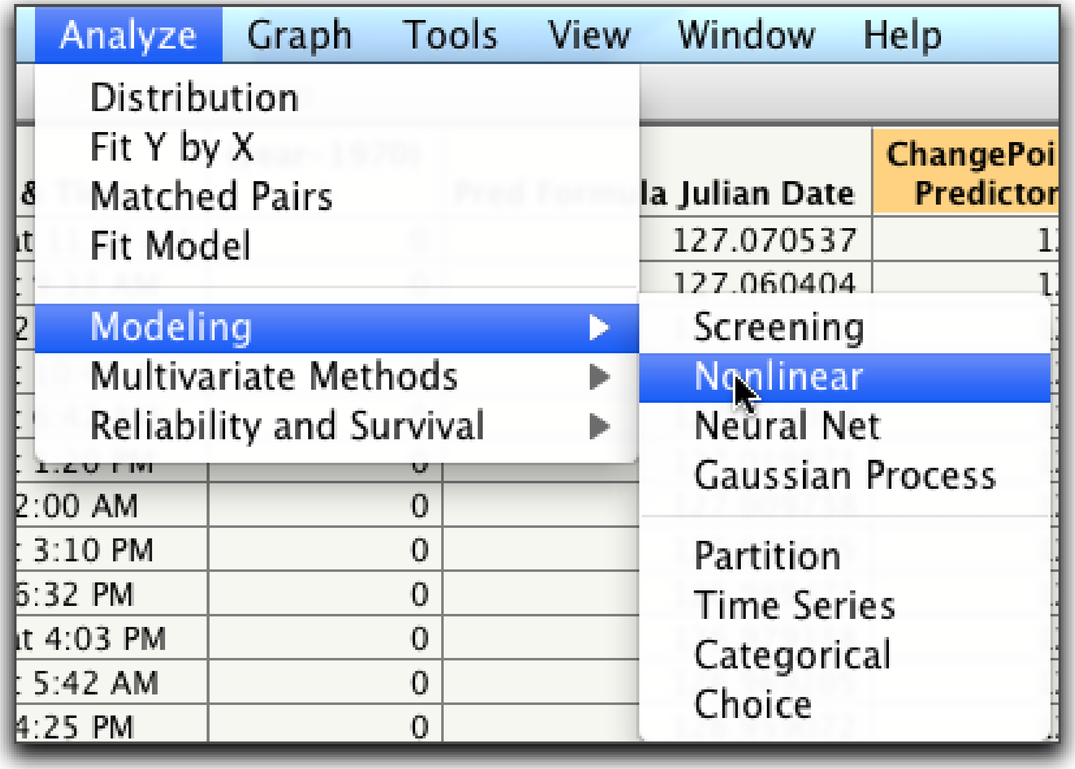
Click on the *New Parameters* item and create the four parameters and their initial values (based on the results from the previous example). The choice of initial values is not that crucial. Then create the predicted value in terms of the parameters and the columns in the data table:



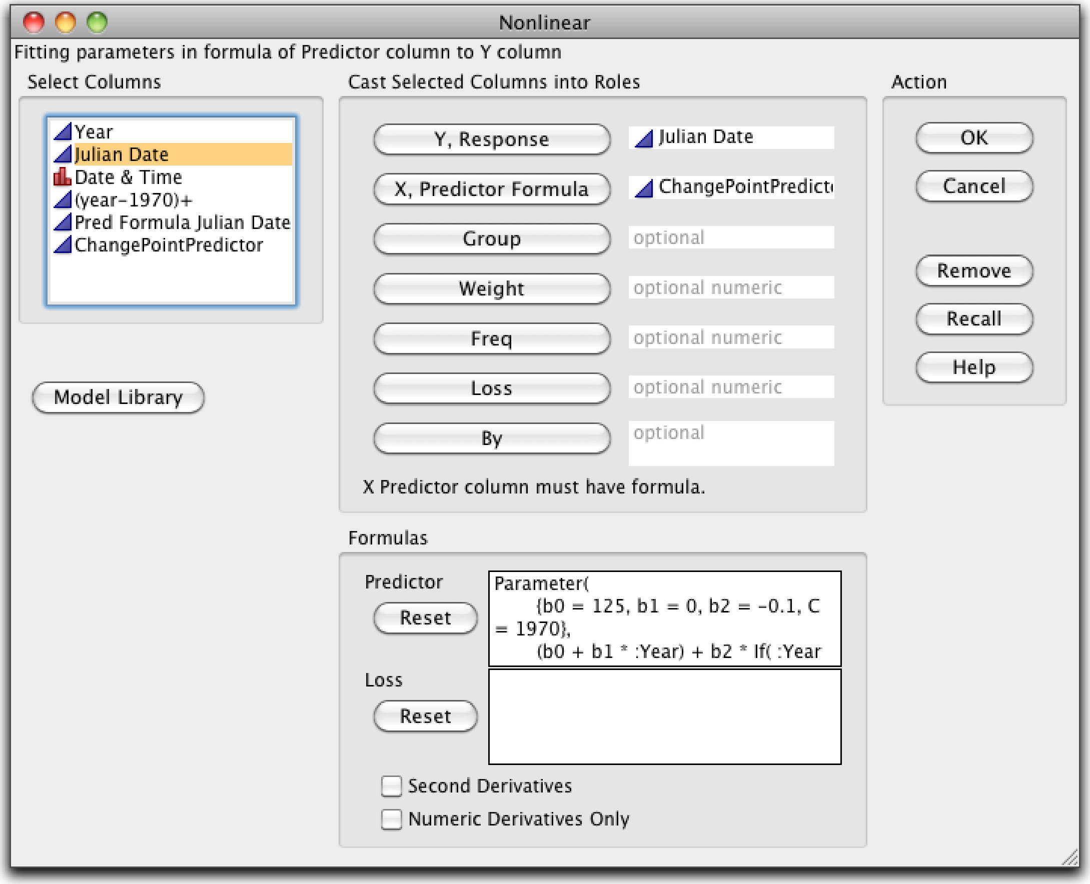
$$(b0 + b1 * Year) + b2 * If\begin{cases} Year >= C \Rightarrow Year - C \\ else \Rightarrow 0 \end{cases}$$

Notice the use of the *If* function to adjust for the break point. You can switch back and forth between the parameters, data table columns, etc. using the drop down menu in the top right of the formula editor. When you are finished, close the Formula Editor, and the data table will be updated with initial predictions based on the initial values specified.

Select the *Analyze->Modelling ->NonLinear* platform:



Specify the predicted value and Y variables appropriately:



Nonlinear

Fitting parameters in formula of Predictor column to Y column

Select Columns

- Year
- Julian Date
- Date & Time
- (year-1970)+
- Pred Formula Julian Date
- ChangePointPredictor

Cast Selected Columns into Roles

Y, Response	Julian Date
X, Predictor Formula	ChangePointPredict
Group	optional
Weight	optional numeric
Freq	optional numeric
Loss	optional numeric
By	optional

X Predictor column must have formula.

Formulas

Predictor

```
Parameter(
      {b0 = 125, b1 = 0, b2 = -0.1, C
= 1970},
      (b0 + b1 * :Year) + b2 * If( :Year
```

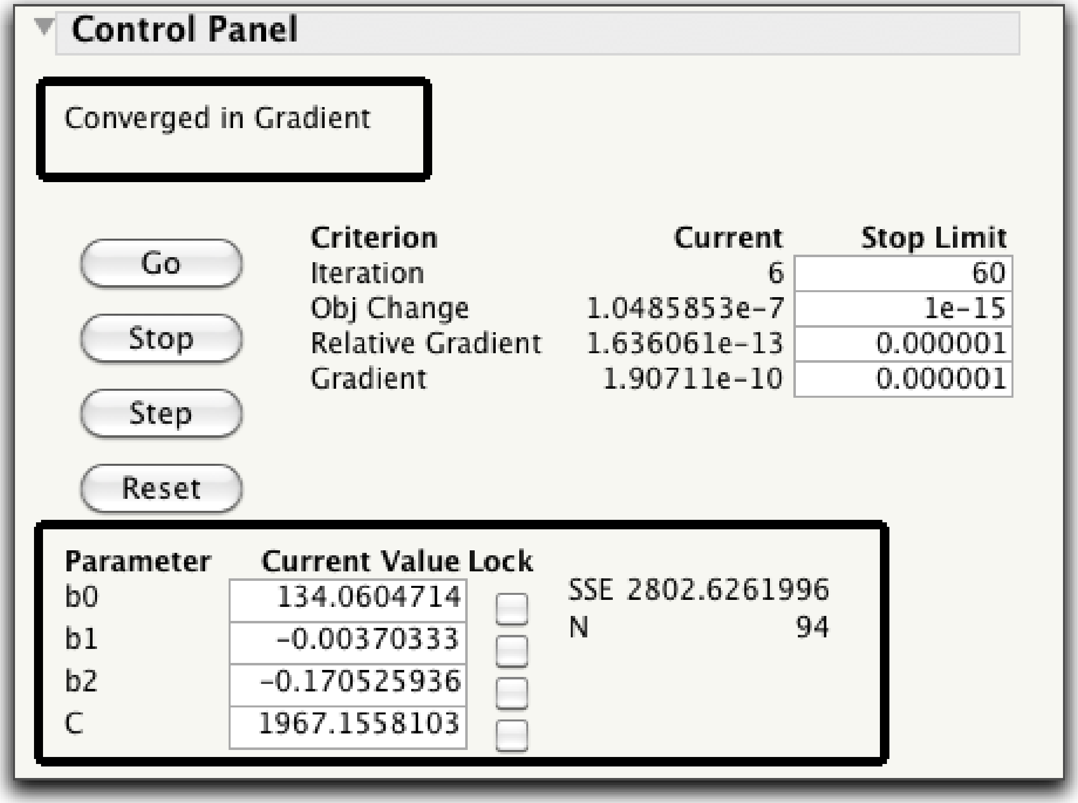
Loss

□ Second Derivatives

□ Numeric Derivatives Only

Notice that the formula for the predictions is displayed.

This brings up the *Analyze->Modelling ->NonLinear* platform control panel. The initial fit is displayed. Press the *Go* button to find the non-linear least squares fit.



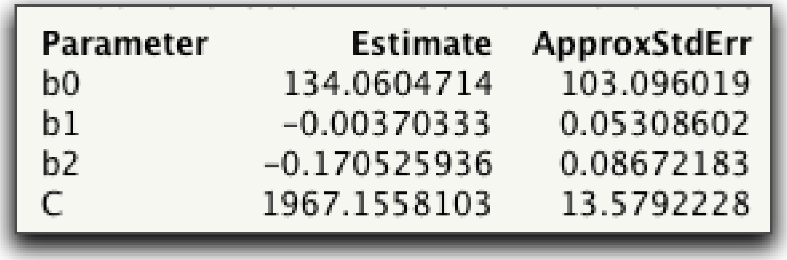
```
Control Panel

Converged in Gradient

Criterion                 Current      Stop Limit
Iteration                       6              60
Obj Change           1.0485853e-7           1e-15
Relative Gradient    1.636061e-13        0.000001
Gradient              1.90711e-10        0.000001

Parameter   Current Value Lock
b0            134.0604714   □     SSE 2802.6261996
b1            -0.00370333   □     N             94
b2           -0.170525936   □
C            1967.1558103   □
```

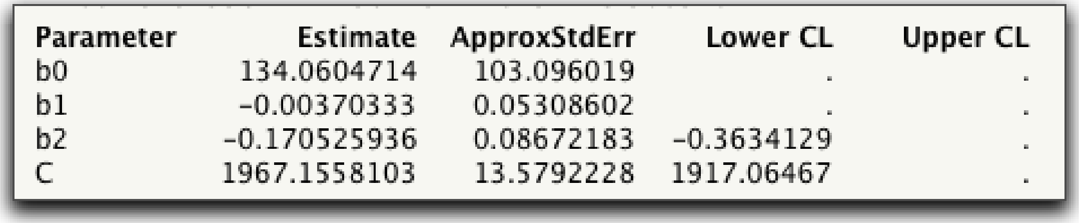
Approximate standard errors are also presented at the bottom of the output:



```
Parameter       Estimate  ApproxStdErr
b0           134.0604714    103.096019
b1           -0.00370333    0.05308602
b2          -0.170525936    0.08672183
C           1967.1558103    13.5792228
```

The non-linear least squares algorithm appears to have converged at the estimates listed in the table. The estimated change-point of 1967 is close to the value of 1970 “guess-timated” earlier.

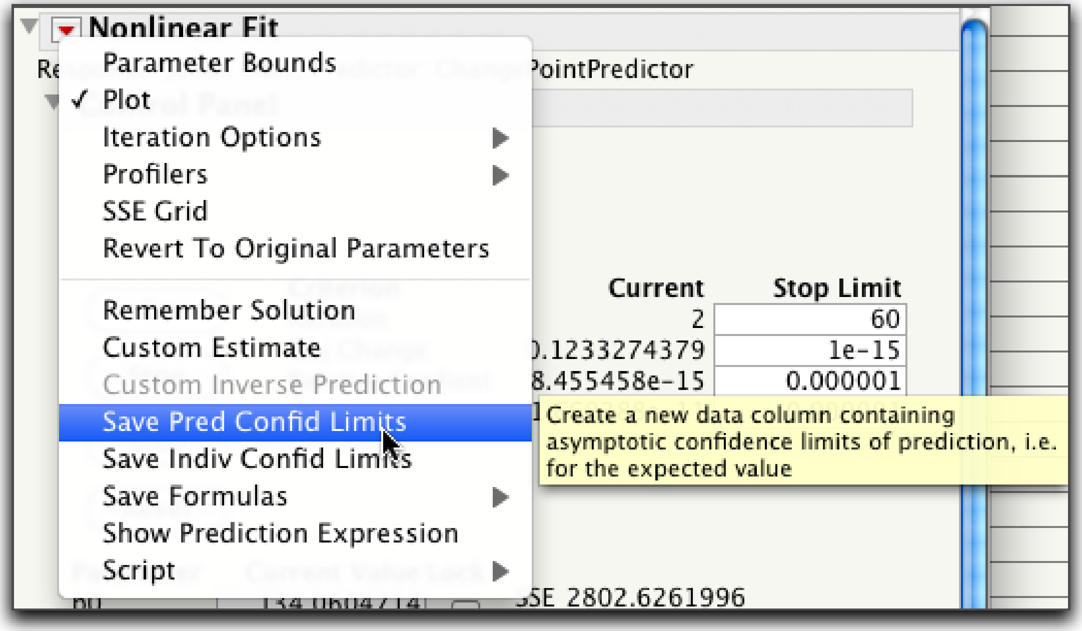
The standard errors are based on large-sample theory. In order to compute a 95% confidence interval for the break point, you could use the standard $estimate \pm 2(se)$, but in small samples, the resulting confidence intervals may not perform well. Toms and Lesperance (2003) recommend that a likelihood ratio confidence interval be computed. *JMP* attempts to compute profile-likelihood confidence intervals when you press the *Confidence Interval* button which gives:



Parameter	Estimate	ApproxStdErr	Lower CL	Upper CL
b0	134.0604714	103.096019	.	.
b1	-0.00370333	0.05308602	.	.
b2	-0.170525936	0.08672183	-0.3634129	.
C	1967.1558103	13.5792228	1917.06467	.

In this case, the profile intervals fail to give upper and lower bounds because the slope after the change point is just on the boundary of statistical significance at ($\alpha = 0.05$). If you change the confidence coefficient form 95% to 90%, the procedure is able to find confidence bounds on the C parameter. Consequently, there may or may not be a change point. Notice that the lower boundary of the confidence interval for C is quite far below the point estimate!

Confidence intervals for the mean response and prediction intervals for a future response are obtained in the usual way and are interpreted in the same way as in ordinary regression. In *JMP*, these are obtained by clicking on the red triangle:



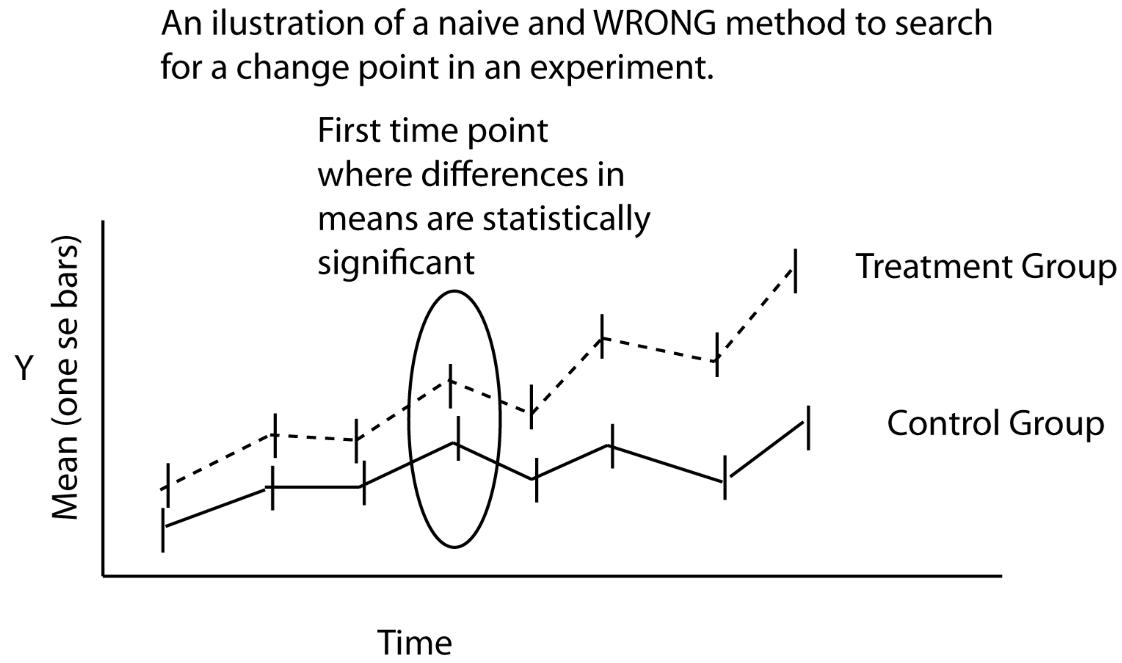
Create a new data column containing asymptotic confidence limits of prediction, i.e. for the expected value

The *Analyze->Modelling ->NonLinear* platform also allows you to "play" with the estimates to investigate the sensitivity of the fit to the parameters. The *Profiler* option under the red triangle is also useful in these cases.

18.3 What is the first time that a treatment mean differ from a control mean

A fairly common request in our Statistical Consulting Service is for help in finding the time at which some treatment gives a difference in response from a control. For example, a group of animals may be fed a control diet and are measured over time, while another group of animals are fed an experimental diet and are measured over time. At which point do the responses between the two groups start to differ?

Let us assume, for simplicity, that separate animals are measured at each time point so that the problem of longitudinal data are ignored. For example, suppose that animals must be sacrificed at each time point to measure the response. A naive analysis starts by plotting the means of the two groups over time and searching for the first time point at which the two means are statistically different:



This is NOT A VALID ANALYSIS! The problem is that the estimate of the change point for this analysis will depend on the sample size and the alpha level (the cutoff to declare statistical significance). If the sample size is small in each group, then the standard error bars are larger, and the estimated change point tends to be larger than if the sample size is large and the standard errors are smaller. If the alpha level is chosen to be $\alpha = 0.10$ rather than $\alpha = 0.05$, then it is easier to detect an effect and so the estimated change point would once again shift.

The actual change point does NOT depend on sample size! All that should happen is that the estimated precision of the change point problem should be worse for smaller sample sizes than for larger sample sizes.

The proper way to search for a change point is to find the DIFFERENCE or log(RATIO) of the means at each time point and then apply the change point analysis to the difference or log(ratio). A model where the difference in means is forced to be zero prior to the unknown change point may be a suitable alternate model.

18.3.1 How long does a bait last for attracting ants?

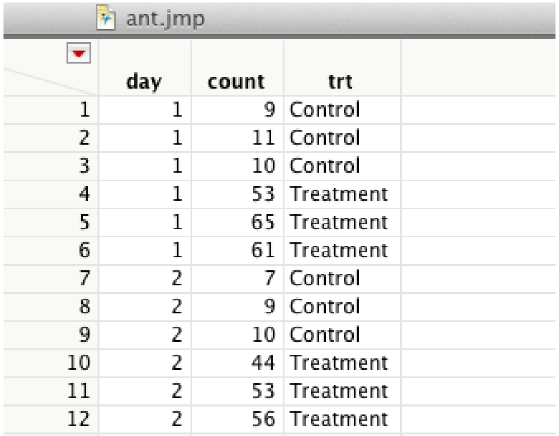
This example is based on a project by Nate Derstine of Biological Sciences at Simon Fraser University. The data are simulated, but illustrative of the process.

Considered one of the worst invasive pest ants, the electric ant, or little fire ant (*Wasmannia auropunctata*) has negatively impacted both biodiversity and agriculture. Its distribution is nearly pantropical, and greenhouse infestations have been reported as far north as Canada and the United Kingdom. Current *W. auropunctata* detection methods commonly employ a food item like peanut butter. An alternative detection method may use pheromone attractants. For W. auropunctata, a one-way trap containing an alarm pheromone has been successfully used to detect little fire ant populations in macadamia nut orchards. What is the longevity of this type of pheromone lure as used in a unique one-way ant trap?

At the beginning of the experiment, 180 control traps and 180 treatment traps were prepared. On each day, three traps of each type were randomized to locations in the orchard where the ant species were known to be present. 24 hours later, the traps were retrieved and the number of ants captured counted, and the trap is discarded.

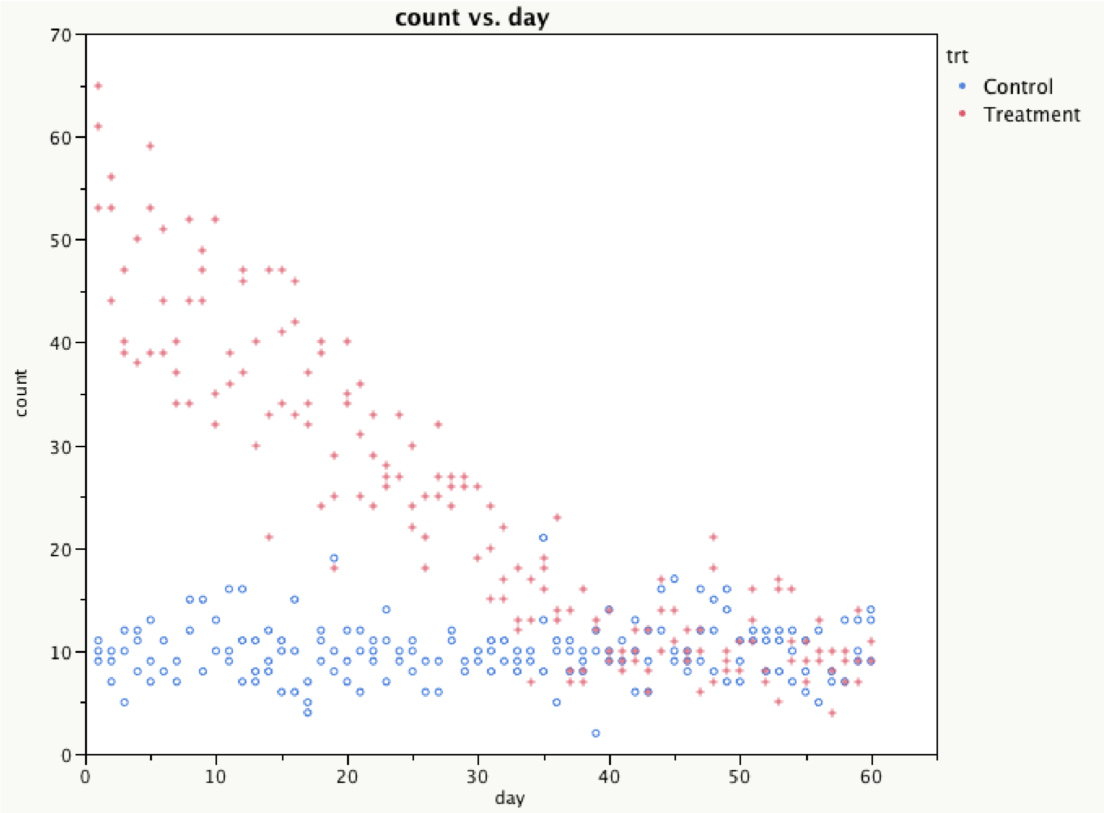
Because separate traps were prepared for each day and randomized each day, each observation can be treated as independent of other observations. This avoids the complications of repeated measures if the same lure is used for multiple days or the same locations used for the experiment – the analysis ideas are similar, but some care is needed to deal with potential correlation in the responses taken from the same trap at the same location over time.

The data is available in the *ants.csv* file in the Sample Program Library at `http://www.stat.sfu.ca/~cschwarz/Stat-650/Notes/MyPrograms`. The data are also available in a *JMP* datafile. Part of the raw data are shown below:


ant.jmp

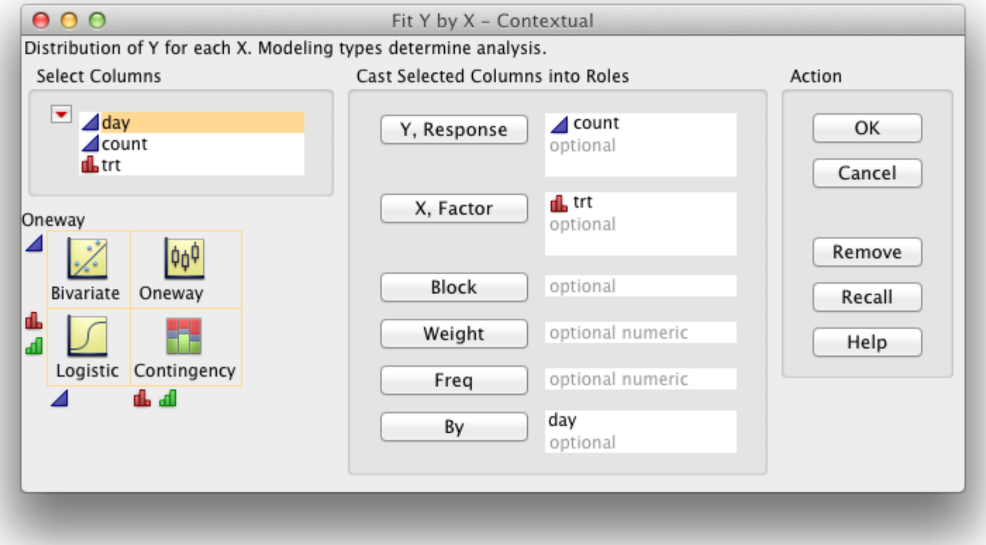
	day	count	trt
1	1	9	Control
2	1	11	Control
3	1	10	Control
4	1	53	Treatment
5	1	65	Treatment
6	1	61	Treatment
7	2	7	Control
8	2	9	Control
9	2	10	Control
10	2	44	Treatment
11	2	53	Treatment
12	2	56	Treatment

Start by plotting the data in the usual way using the *Analyze->Fit Y-by-X* platform after assigning markers to the row based on the *trt* variable:



It would appear from the plot, that the pheromone loses its effectiveness somewhere between day 30 and day 50, but the actual time point is difficult to see because of the noise in the data.

A naive analysis would do a t-test on each day and see when a statistically significant difference was detected. This takes a bit of work in*JMP*, but is relatively straightforward. First, use the *Analyze->Fit Y-by-X* platform to do a t-test for each day, comparing the mean count for the two treatments. Notice the use of day as a *By* variable.



Fit Y by X - Contextual

Distribution of Y for each X. Modeling types determine analysis.

Y, Response: count

X, Factor: trt

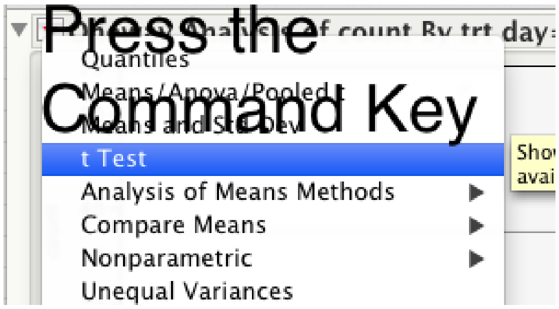
Block: optional

Weight: optional numeric

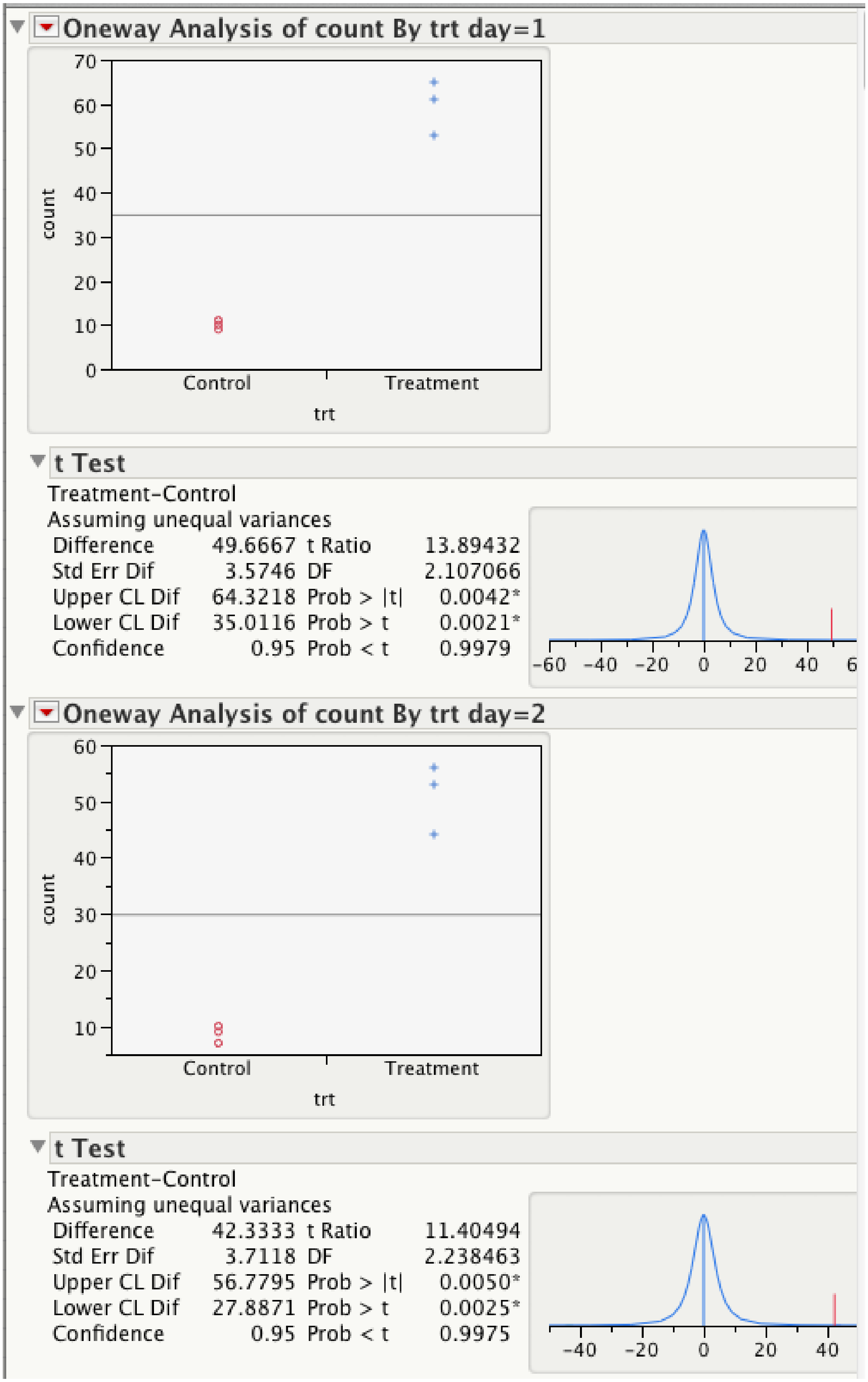
Freq: optional numeric

By: day

This gives a separate plot by *Day*. To fit the t-test for each day, use the **Command** key, click on the red-triangle, and select t-test:



You now have a separate report for EACH day:



Oneway Analysis of count By trt day=1

t Test

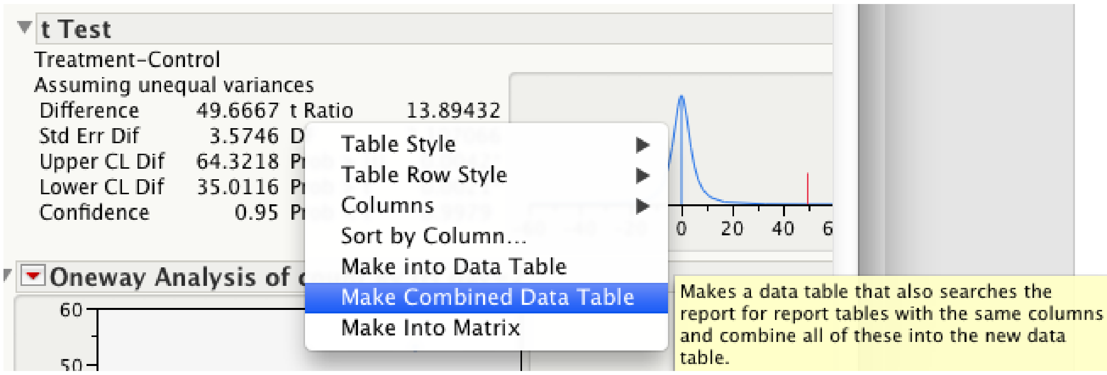
```
Treatment-Control
Assuming unequal variances
 Difference      49.6667  t Ratio       13.89432
 Std Err Dif      3.5746  DF            2.107066
 Upper CL Dif    64.3218  Prob > |t|     0.0042*
 Lower CL Dif    35.0116  Prob > t       0.0021*
 Confidence         0.95  Prob < t       0.9979
```

Oneway Analysis of count By trt day=2

t Test

```
Treatment-Control
Assuming unequal variances
 Difference      42.3333  t Ratio       11.40494
 Std Err Dif      3.7118  DF            2.238463
 Upper CL Dif    56.7795  Prob > |t|     0.0050*
 Lower CL Dif    27.8871  Prob > t       0.0025*
 Confidence         0.95  Prob < t       0.9975
```

We want to extract the p-value for each day. Right click in ANY of the t-test reports and choose *Make Combined Data Table*:

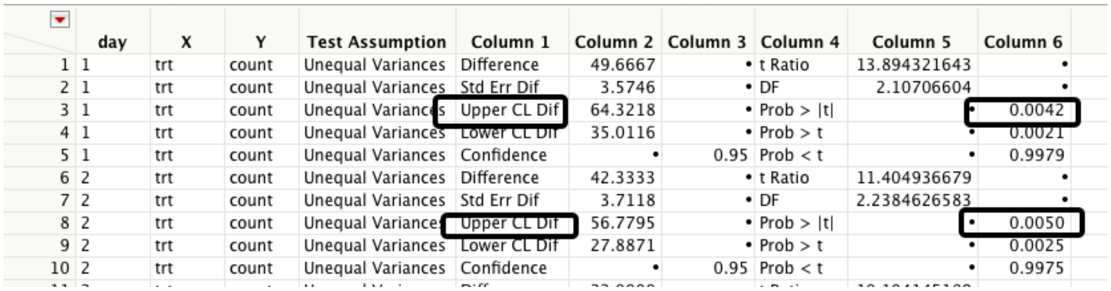


t Test
Treatment-Control
Assuming unequal variances

Difference	49.6667	t Ratio	13.89432
Std Err Dif	3.5746	D	
Upper CL Dif	64.3218	P	
Lower CL Dif	35.0116	P	
Confidence	0.95	P	

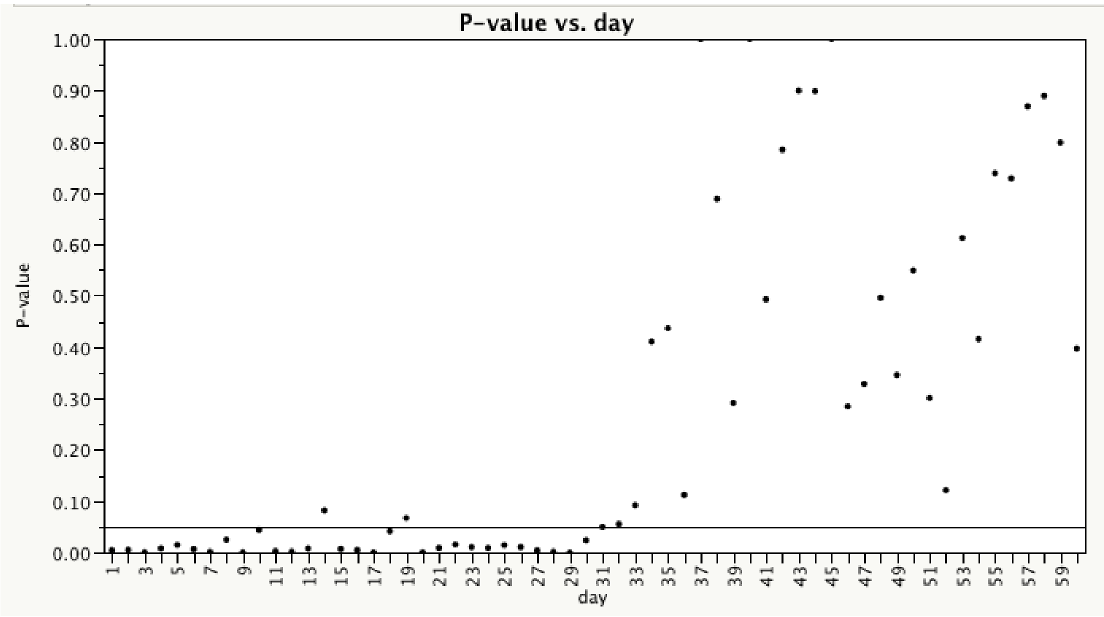
Makes a data table that also searches the report for report tables with the same columns and combine all of these into the new data table.

This creates a MESSY data table:



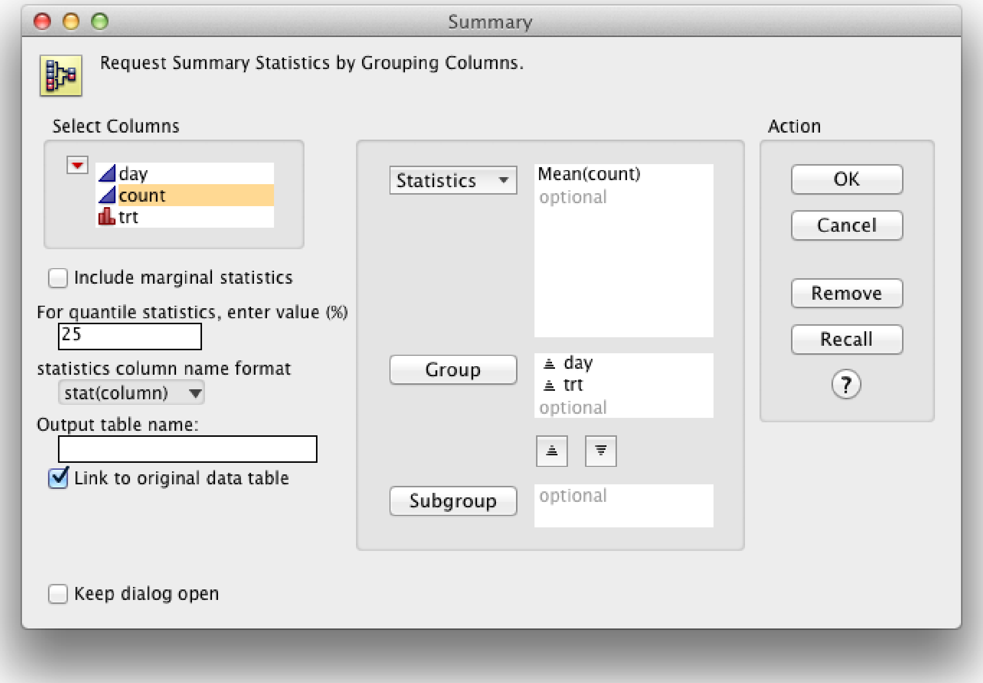
	day	X	Y	Test Assumption	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
1	1	trt	count	Unequal Variances	Difference	49.6667	•	t Ratio	13.894321643	•
2	1	trt	count	Unequal Variances	Std Err Dif	3.5746	•	DF	2.10706604	•
3	1	trt	count	Unequal Variances	Upper CL Dif	64.3218	•	Prob > |t|	•	0.0042
4	1	trt	count	Unequal Variances	Lower CL Dif	35.0116	•	Prob > t	•	0.0021
5	1	trt	count	Unequal Variances	Confidence	•	0.95	Prob < t	•	0.9979
6	2	trt	count	Unequal Variances	Difference	42.3333	•	t Ratio	11.404936679	•
7	2	trt	count	Unequal Variances	Std Err Dif	3.7118	•	DF	2.2384626583	•
8	2	trt	count	Unequal Variances	Upper CL Dif	56.7795	•	Prob > |t|	•	0.0050
9	2	trt	count	Unequal Variances	Lower CL Dif	27.8871	•	Prob > t	•	0.0025
10	2	trt	count	Unequal Variances	Confidence	•	0.95	Prob < t	•	0.9975

Finally, we select only the rows where *Column 1* contains the string "Upper CL Dif" and then plot *Column 6* vs. day, adjust the axes to get the plot:



Based on this naive analysis, it would appear that the pheromone is effective only out to about day 30, but the previous graph shows that the change point should likely occur around day 40ish. Also note that this naive analysis failed to detected a difference in the mean count captured just after day 10 and just before day 20, but the previous graph shows this is likely a false negative. This again illustrates the perils of the naive analysis.

The change-point method of analysis starts by finding the average for each day for treatment and control.



Summary

Request Summary Statistics by Grouping Columns.

Statistics: Mean(count)

Group: day, trt

For quantile statistics, enter value (%): 25

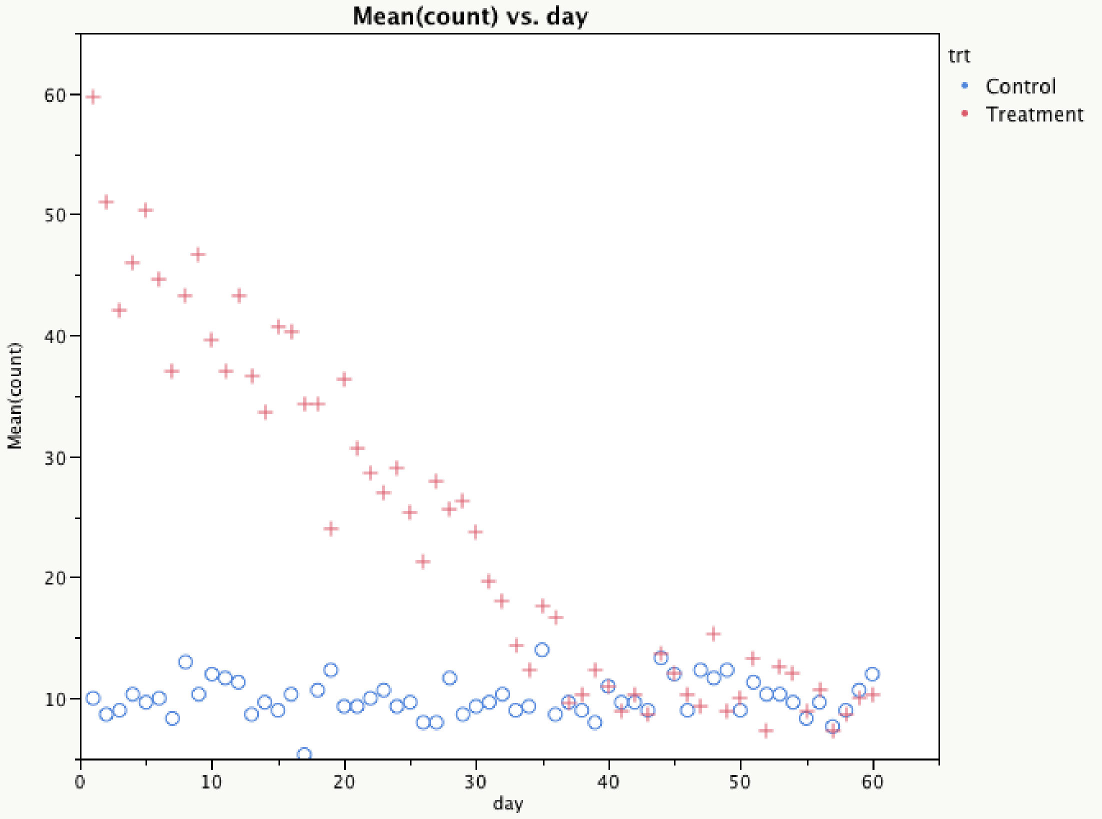
statistics column name format: stat(column)

☑ Link to original data table

□ Include marginal statistics

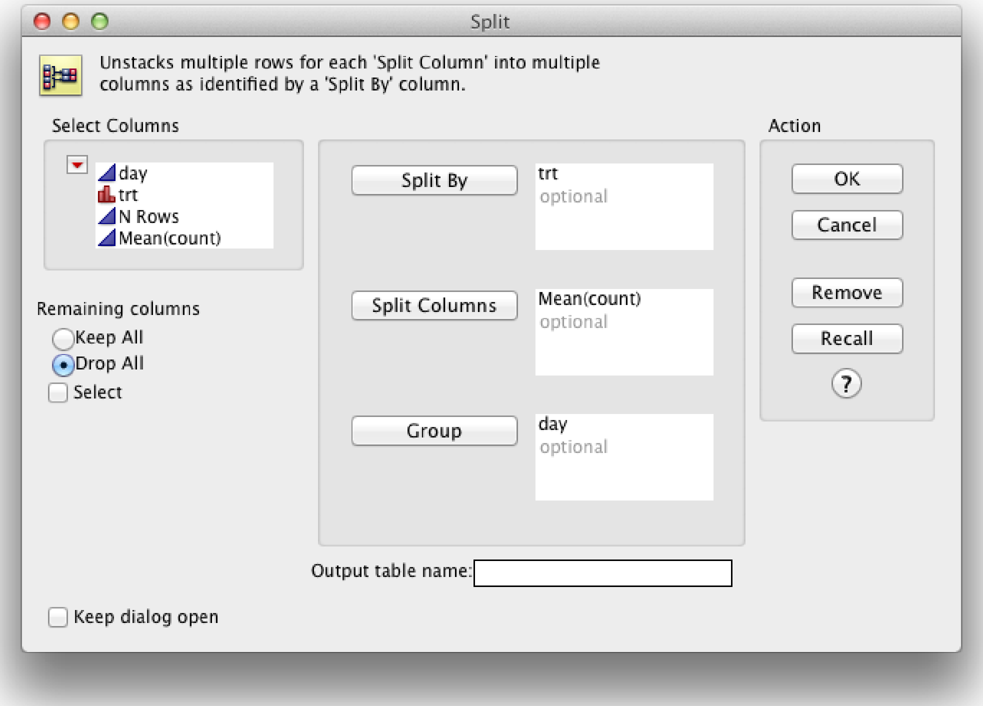
□ Keep dialog open

and then plotting the results to give:



This plot again shows that the pheromone lasts to about day 40ish.

Finally, we find the log(ratio) of the treatment mean to the control mean and plot this ratio by day. We start in *JMP* by splitting the data and then creating a new formula variable:



Unstacks multiple rows for each 'Split Column' into multiple columns as identified by a 'Split By' column.

Select Columns

- day
- trt
- N Rows
- Mean(count)

Split By: trt

Split Columns: Mean(count)

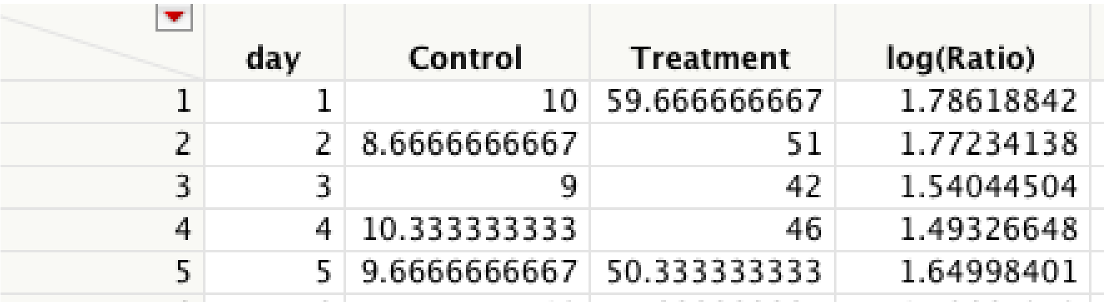
Group: day

Remaining columns

○ Keep All
● Drop All
□ Select

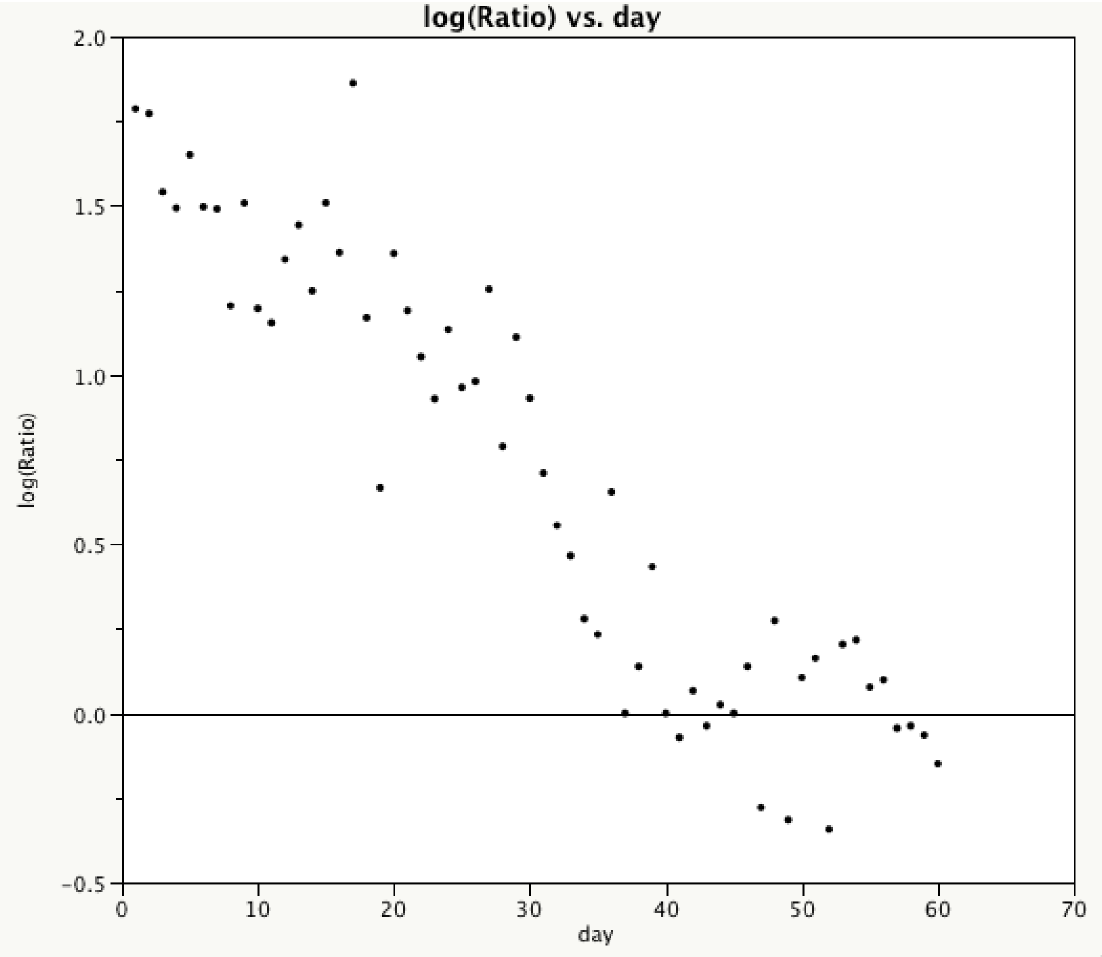
Output table name:

□ Keep dialog open



	day	Control	Treatment	log(Ratio)
1	1	10	59.666666667	1.78618842
2	2	8.6666666667	51	1.77234138
3	3	9	42	1.54044504
4	4	10.333333333	46	1.49326648
5	5	9.6666666667	50.333333333	1.64998401

and then plotting the $\log(ratio)$:



Note that a $\log(ratio) = 0$ implies that the mean counts are the same. The change point appears to be around 40 days again.

The change point model is

$$log(ratio)_i = \beta_0 + \beta_1(Day_i) + \epsilon_i$$

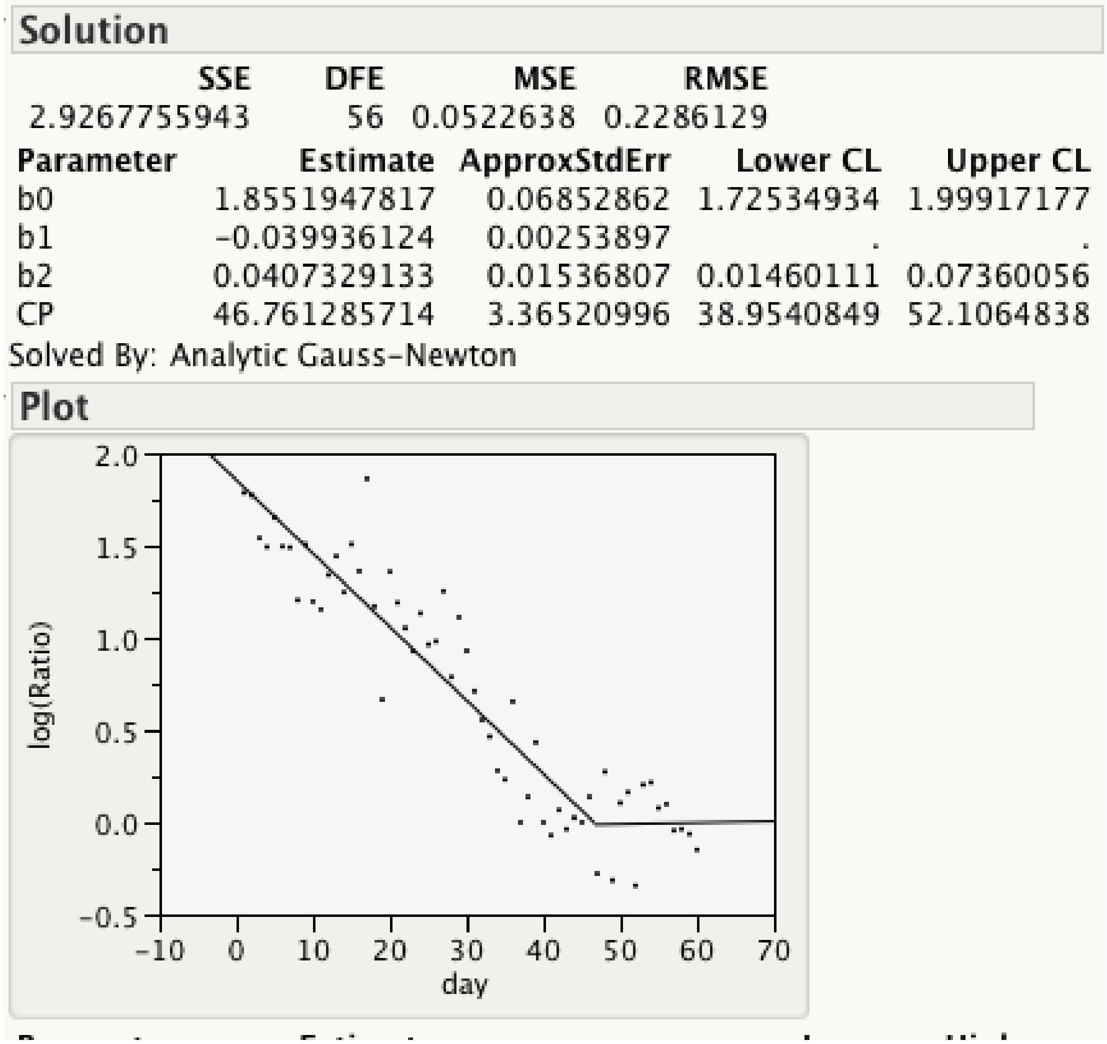
if Day_i is less than the change point (CP), and

$$log(ratio)_i = \beta_0 + \beta_1(Day_i) + \beta_2(Day_i - CP)^+ + \epsilon_i$$

where $(x)^+$ takes the value 0 if $x < 0$ and x if $x > 0$ as outlined earlier in Section 18.2.

We can fit a change-point model to the $\log(ratio)$ using the *NonLinear* platform like we did in Section 18.2.1. Refer to the previous example on how to set up the *NonLinear* model.

The final estimates are:



Solution

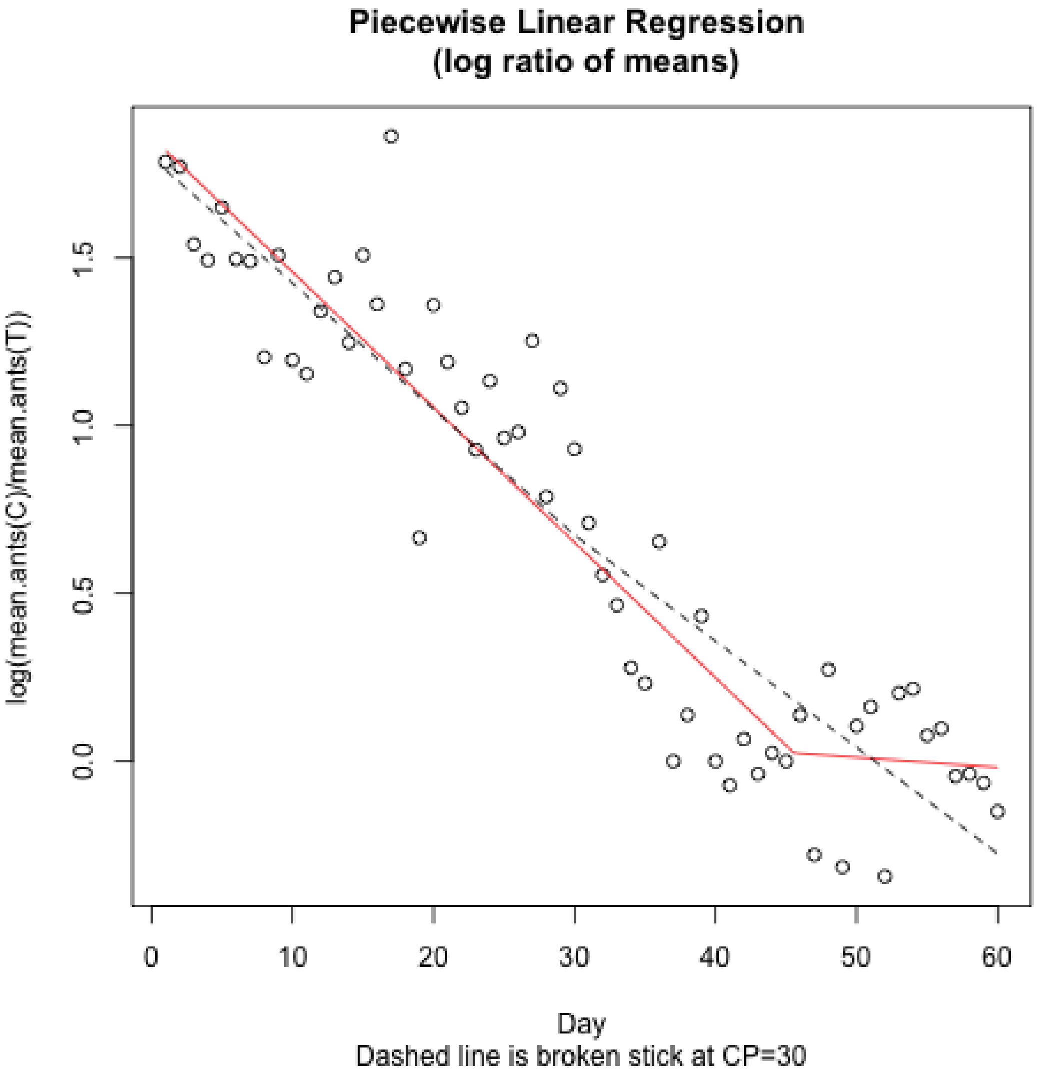
```
          SSE      DFE        MSE        RMSE
 2.9267755943       56  0.0522638   0.2286129
Parameter       Estimate  ApproxStdErr    Lower CL    Upper CL
b0          1.8551947817    0.06852862  1.72534934  1.99917177
b1          -0.039936124    0.00253897           .           .
b2          0.0407329133    0.01536807  0.01460111  0.07360056
CP          46.761285714    3.36520996  38.9540849  52.1064838
Solved By: Analytic Gauss-Newton
```

Plot

The results are not entirely satisfactory[2] - the estimated change point is around 46 days which seems sensible, but the 95% confidence interval for the change point is very wide! Why was the change point estimated so poorly?

The problem is that are no constraints on the slope after the change point. So a model where the change point is, say around 30, gives almost the same fit as the model where the change point is around 45:

[2] Note that *SAS*, *R* and *JMP* give slightly different answers because *R* uses maximum likelihood estimation, while *SAS* and *JMP* uses non-linear least squares. The programs also compute the confidence intervals differently – *R* uses a bootstrap approach while *JMP* and *SAS* use a delta-method approximation. The results are all asymptotically equivalent.

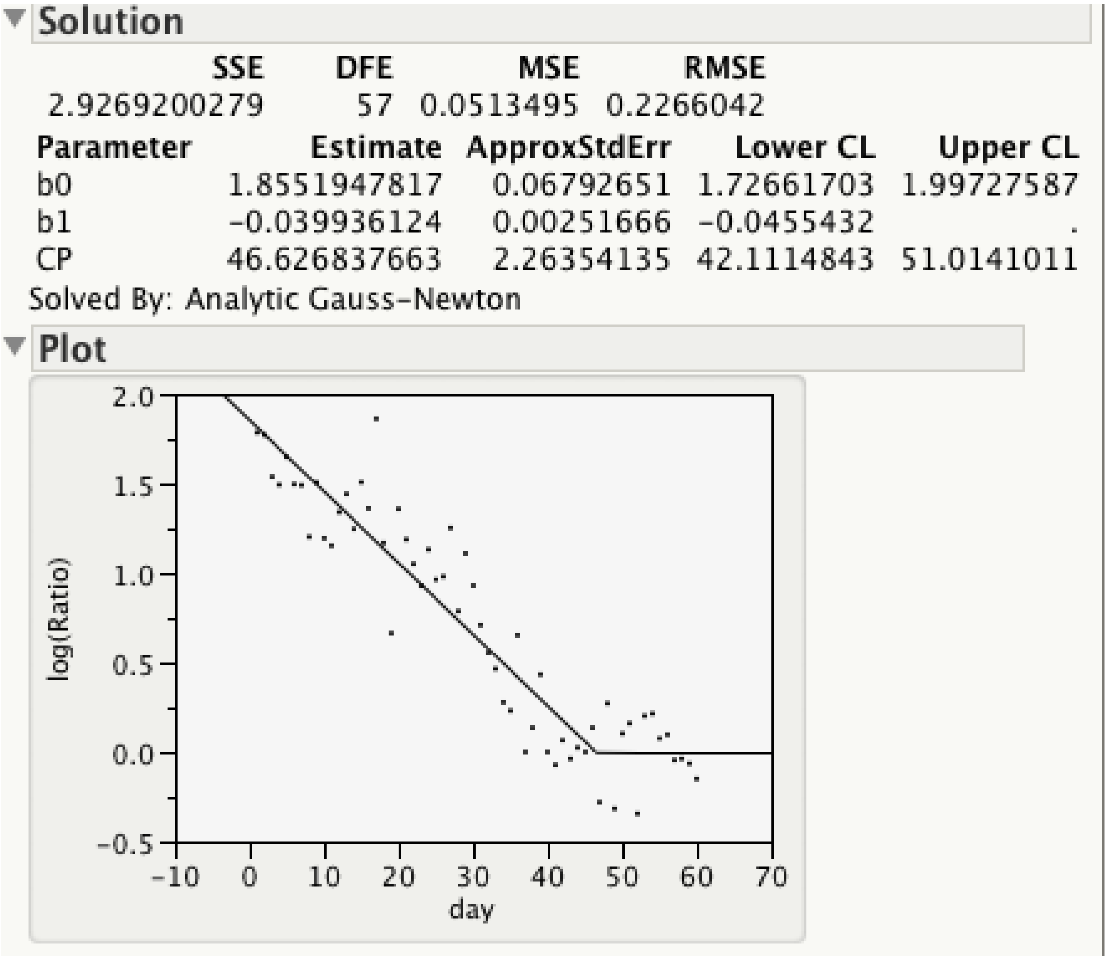


In these cases, what is needed is a broken-stick model where the slope AFTER the change point is forced to be zero. This model is surprisingly easy to fit:

$$Y_i = \beta_0 + \beta_1 min(CP, Day_i)$$

where CP is the change point. Of course, the model must be fit with different values for the CP to find the best fit.

The non-linear least-squares procedure can be easily modified to force the slope to be 0 after the change point. Consult the *JMP* script for details. The final results is:



```
Solution
              SSE       DFE        MSE       RMSE
   2.9269200279        57  0.0513495  0.2266042
Parameter          Estimate  ApproxStdErr     Lower CL     Upper CL
b0             1.8551947817    0.06792651   1.72661703   1.99727587
b1             -0.039936124    0.00251666   -0.0455432            .
CP             46.626837663    2.26354135   42.1114843   51.0141011
Solved By: Analytic Gauss-Newton
Plot
```

The revised fit has an estimated change point around 47 days with a 95% confidence interval for the change point between 43 and 51 days – much tighter than the previous broken-stick change-point model.

The same basic methods are employed in the cases where the same trap and/or location is repeatedly used over time. As a general rule, you would want multiple traps of each type and try and randomize the locations over time as much as possible. It is not necessary to measure the traps every day. For example, at the start of the experiment, you could take a daily measurement every 5 days, and then switch to more intensive monitoring (i.e. daily) after about 30 days.

The broken-stick model is a simplification of reality (there likely isn't such a sharp change at the change point), but will serve as a close approximation to the underlying process.