

Regression Line: How to Construct the Best Fit

HAWK WATCH STATISTICS

Trudy Battaly and Drew Panko

Updated 8/6/2003

This document supplements our article,
Analyzing Hawk Count Data: Graphs that show Year to Year Trends,
published in the [HMANA Hawk Migration Studies](#), September, 2003.

Table of Contents

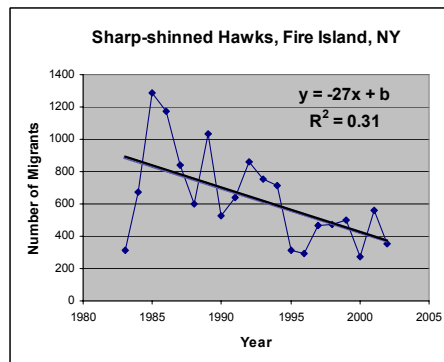
(Use bookmark tab on left.)

A. Start with a Graph	p.2
B. Trendline Options	p.2
C. The Regression Line – What Is It?	p.3
D. Making the Equation and R^2 more Readable	p.3
E. Changing the Source Data	p.4
F. New Regression Equation	p.5
G. Assignment	p.5

For data entry, manipulation, and sorting, please see [Introductory Excel](#).

For creating graphs, please see [Graphing Your Data](#)

We welcome your comments, criticism, or questions. [Contact the Authors](#)

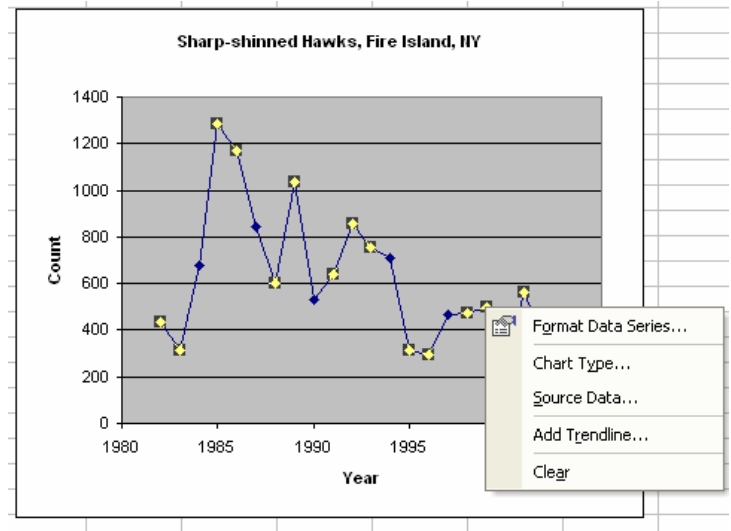


(This document may be browsed on the web or printed for use at your computer.)

A. Start with a Graph

1. Open Excel and the file with your data and graph. Or, download the practice file which we will use in this lesson: [practice.xls](#). (Right click and save.)
2. Find the graph on the spreadsheet. If you do not have an existing graph, you can create one using the previous lesson, [Graphing Your Data](#). Or, in the practice file, click on the Tab at the bottom left that is labeled SS Graph to find the Sharp-shinned Hawk graph.

3. **Right click on any of the data points.** This highlights the points and brings up a menu. From the menu, select “Add Trendline.”



B. Trendline Options

1. There are 6 different types of regression. For our purposes, select **Linear**. This will result in the best fitting straight line.
2. Click on the **Options** Tab. Select the bottom two options: **Display equation on chart**, and **Display R-squared value on chart**.
3. Click OK, using the default values for the other options. The regression line, equation, and R^2 appear on the graph.

C. The Regression Line – What Is It?

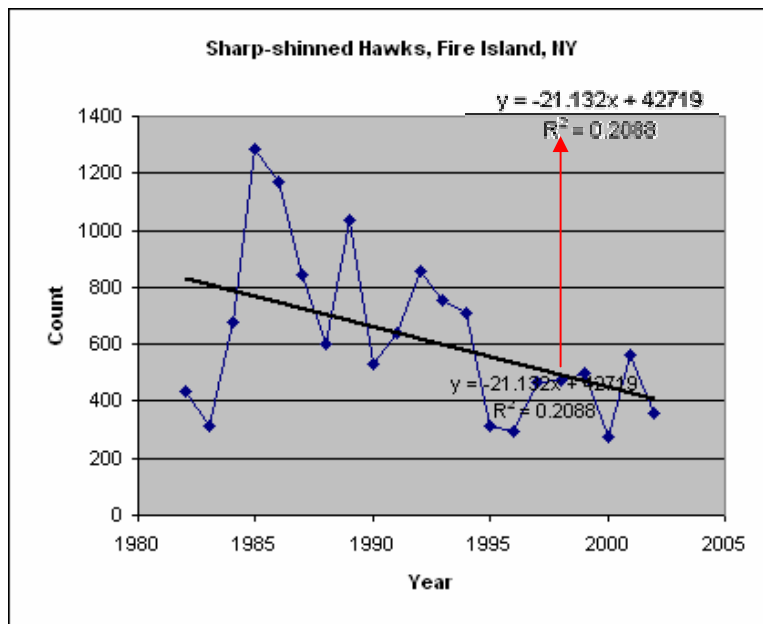
The regression line has two aliases: the “best fit line,” and the “least squares line.”

Given all the lines that could be used to represent the trend in the data, the regression line is the closest to all the points, or the “best fit.” The statistical procedure that guarantees this considers the vertical distance of each of the points from the line. Some of the points are above the line, having a positive distance, some are below, with a negative distance, and the sum of the distances is zero. To obtain a non-zero result, each distance is squared before the sum is computed. The regression line is the one line (of many possible lines) with the lowest sum of squares, or the “least squares line.”

If we used a non-linear option in step B1 we would refer to the regression **equation** rather than the regression line. Assumptions for the correct use of the regression line will be addressed in a separate article.

D. Making the Equation and R^2 more Readable

1. **Move the equation to a visible location.** The equation and R^2 usually appear in the middle of the graph and are difficult to read. Click on them and drag them up to a space where you can see them better.



2. **Change the font.** All options for font are available for your equation, including size, color, bold, italic, etc. **Enlarge or change the font** in either of 2 ways:

- Click on the equation to highlight it, and **use the excel formatting menu** to make the changes.
- Or, right click on the equation, select **Format Data Labels**, and choose the Font tab.

E. Changing the Source Data

1. Notice that the equation and R^2 are **different from those generated in our paper.**

	slope	R^2
Uncorrected totals	-27	.31
All years - 9/15-10/20	-32	.50
'92 - '02 9/10-11/7	-35	.37

2. **Why** the difference? In the first lesson, where we created the graph, we focused on how, not what or why. We did not consider which years should be included in the graph. (See article.) Now that we have used regression to look for trends, we need to be more careful about which years are comparable.

3. Return to the table in the practice file.

YEAR	OS	BE	NH	SS	CH	NG	RS	BW	RT	RL	GE	AK	ML	PG	U	TH	HRS	DAYS	Range	Reg	Coverage
1982	49	0	65	435	1	0	0	1	0	0	0	1204	226	30	15	2026	86	14			
1983	52	0	78	312	8	0	0	1	2	0	0	989	497	39	43	2021	213	33	9/15 - 10/20		
1984	108	0	151	676	8	0	0	1	0	0	0	999	491	53	31	2518	212	31	9/16 - 10/16		
1985	167	1	239	1284	6	0	0	1	0	0	0	1613	1027	40	30	4408	260	37	9/11 - 10/20		
1986	135	1	252	1171	23	0	0	1	0	0	0	1851	1093	110	35	4672	286	42	9/9 - 10/20		
1987	258	1	175	841	16	0	0	0	0	0	0	3245	1181	169	10	5896	287	42	9/10 - 10/20		
1988	301	2	207	602	3	0	0	0	0	1	0	2536	1342	99	35	5128	334	46	9/10 - 10/30		
1989	267	0	237	1036	6	0	0	1	0	1	0	2214	1348	199	16	5325	362	46	9/9 - 10/25		
1990	235	0	191	530	9	1	0	0	0	0	0	1538	1118	249	32	3903	427	55	9/7 - 10/31		
1991	278	0	131	639	10	0	0	1	2	1	0	2051	1092	168	33	4406	445	59	9/6 - 10/30		
1992	257	5	153	857	28	0	0	0	1	2	1	2311	1203	114	36	4968	483	60	9/9 - 11/8		
1993	321	3	328	753	34	4	2	2	1	0	0	3271	1249	112	24	6104	429	57	9/8 - 11/7		
1994	227	4	391	712	36	0	0	0	0	1	0	3523	1638	78	44	6654	480	65	9/10 - 11/13		
1995	350	0	200	311	13	0	0	1	2	0	0	1121	1084	162	29	3273	472	65	9/9 - 11/13		
1996	241	3	98	291	18	1	0	0	1	0	0	982	1123	171	30	2946	411	59	9/11 - 11/12		
1997	237	2	133	464	17	0	0	0	2	0	0	1229	851	148	13	3096	424	66	9/11 - 11/6		
1998	409	6	328	474	11	0	0	0	1	0	0	1598	1600	214	28	4669	470	65	9/7 - 11/9		
1999	390	5	228	501	24	1	0	0	1	3	0	1503	1581	253	35	4504	340	51	9/11 - 11/7		
2000	325	4	208	273	8	5	1	0	1	1	0	1222	888	148	32	3116	445	65	9/5 - 11/8		
2001	383	4	249	561	32	7	1	0	1	1	0	888	1382	185	27	3722	471	68	9/10 - 11/12		
2002	287	1	169	355	22	1	0	0	3	0	0	1105	1415	169	15	3543	399	55	9/5 - 11/8		

Notice there were **only 14 days of coverage for 1982**, less than half that in all other years. Therefore, 1982 is not comparable and should not be included in the analysis.

4. **Copy the Graph.**

- Left click on the graph, then select copy from the standard bar.
- Click in a cell below the graph and Paste.

5. **Change the Source Data to remove 1982**

- Right click on the new graph. Select Source Data, then Series.
- Click on the right square in the X-values box. Highlight 1983 to 2002. Click the right square again to return to the Series window.
- Click on the right square in the Y-values box. Highlight 312 to 355 in the SS column. Click the right square again to return to the Series window. Click OK.

F. New Regression Equation

1. Notice that the equation and R^2 automatically changed when you changed the data source. Neat stuff! You don't have to repeat the whole process for a different data set!
2. Check the regression summary table. Your new results should agree with the "Uncorrected totals."
3. If you have read our article, *Analyzing Hawk Count Data: Graphs that show Year to Year Trends*, you know that Excel provides more digits in regression than is meaningful. **To edit the equation:**
 - a) Click on the equation once, and then click again.
 - b) Change the equation and R^2 to: $y = -27x + b$ and $R^2 = 0.31$

Notes: The number 55294 is meaningless for our purposes, so we use a general y-intercept, b. The -27 suggests that, on average, in each year there were 27 fewer Sharpies than in the previous year. The R^2 suggests that about 31% of the variation in the data can be attributed to the passage of time (years).

4. **Reserve any edits to the equations until the last step.** The edit prevents the automatic change to the regression results when you change the source data.

G. Assignment:

1. a) Either: Volunteer to cover the hawk watch at the Fire Island. 😊

OR

b) Repeat steps E and F using data for the years 1992 to 2002, and compare your results with the regression table.

2. Starting from scratch, create the graph and the regression line for the years 1992 to 2002. Compare your results to those from assignment 1b above and to the regression table in E.

And: Let us know who you are and how you are doing: [Contact the Authors](#)