Finding an Empirical Model
Using Linear Least-squares Regression

Dr. Brian Hornbuckle

November 8, 2020

A *model* can be used to make predictions. A model is called "empirical" if it is solely based on data. For example, if we had measurements of crop yield, along with the total amount of precipitation that fell during the growing season, we could make an empirical model to predict crop yield, Y, given growing-season precipitation, P.

$$P \Longrightarrow \text{model} \Longrightarrow Y \tag{1}$$

Here P is the independent variable (the information we would use to make a prediction) and Y, what we want to know, is the dependent variable (because Y depends on P). Compare this to a "process-based" model that would try to replicate specific processes like root water-uptake, photosynthesis, etc. that link yield to precipitation.

If the empirical model we create is a *linear* model, then mathematically (1) is written

$$Y = a \times P + b \tag{2}$$

where a is the slope and b is the y-intercept of the best-fit line relating measured values of crop yield to growing-season precipitation.

The predictions of any model will have some uncertainty. Empirical models are only as good as the data used to make the model, and can only be expected to make good predictions in situations that closely resemble the situation in which the data were collected. Process-based models can be more sophisticated, but that doesn't always mean better.

Here is a procedure to develop an empirical model (and the uncertainty associated with its predictions) in a spreadsheet like Excel.

1. Get the data needed to construct the model into Excel by simply opening the data file in Excel, or typing in the data.

2. Insert a scatter-plot chart. Add axes labels.

3. Add a linear least-squares regression best-fit line. For example, in Excel select the "Plot Area" of the chart and then click on "Add Chart Element." Select "Trendline" and then "Linear." Right-click on the best-fit line and select "Format Trendline..." Then click the buttons for "Display equation on chart" and "Display R-squared value on chart."

4. Find the statistics associated with this best-fit line by using the `LINEST` function. In Excel select a cell near your chart. Use the "Insert Function" command from the "Formulas" tab or simply type in `=LINEST(B3:B38,D3:D38,TRUE,TRUE)` where in this example the dependent variable (y-values) is in cells `B3:B38` and the independent variable (x-values) is in cells `D3:D38`. Type "Return" or "Enter."

5. Now you will see either one of two things.

   (a) A 5 × 2 grid of statistical values.

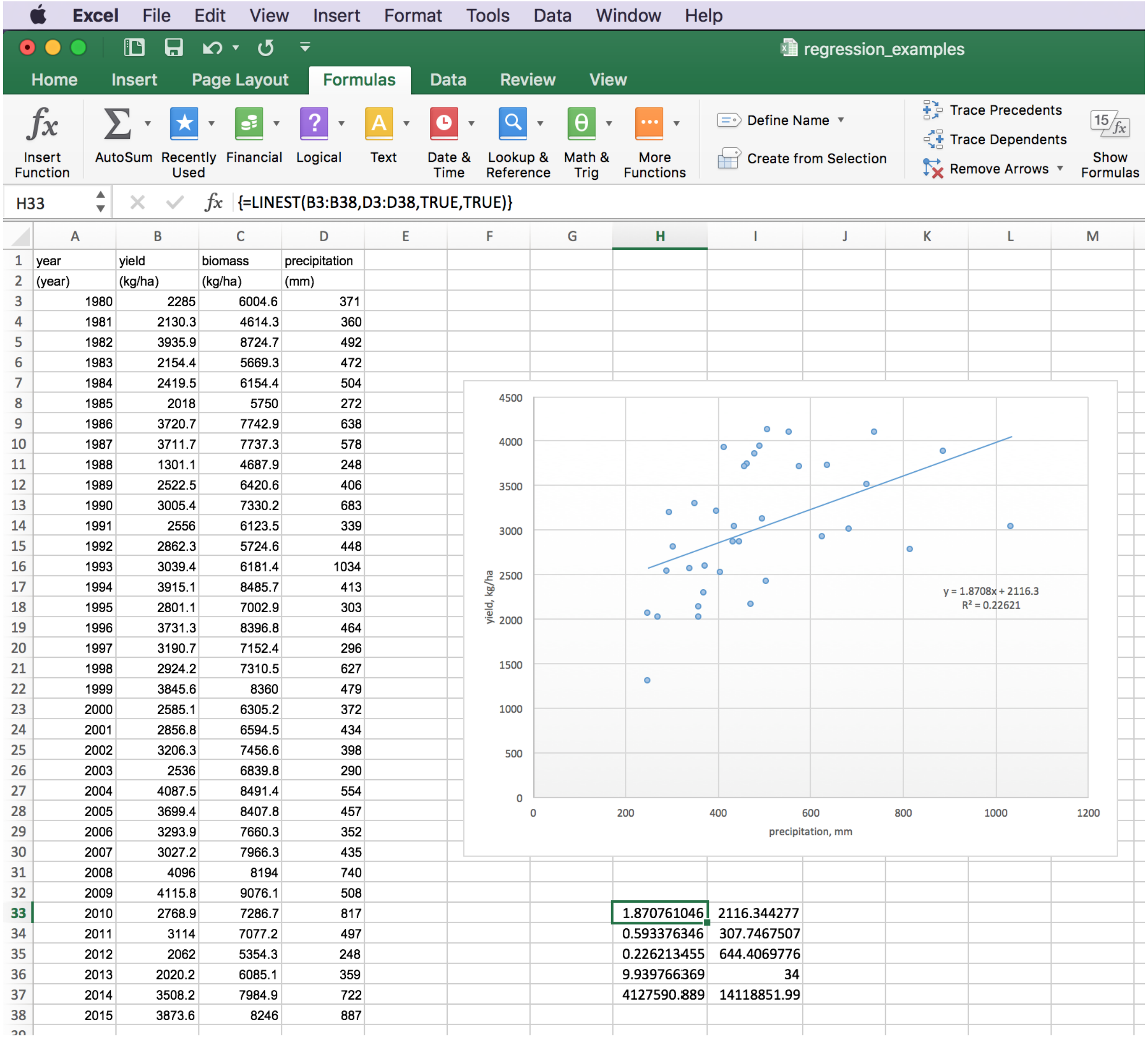
   (b) One value in the cell in which you placed the `LINEST` formula. This value is the slope of your best-fit line. But there is more information about the best-fit line that we want to use. To get this information, click and hold on the cell containing the `LINEST` formula, and highlight a 5-row by 2-column area. Then click in the formula bar (where you see `=LINEST(B3:B38,D3:D38,TRUE,TRUE)`), move the cursor to the end of the command (to the right of the last parentheses), and then type CTRL-SHIFT-ENTER (or CTRL-SHIFT-RETURN).

   These cells have the following values.

slope of best-fit line	y-intercept of best-fit line
uncertainty in slope	uncertainty in y-intercept
square of the correlation coefficient	uncertainty in output
F statistic	degrees of freedom
sum of squares of regression	sum of squares of residuals

6. Note that the equation of the best-fit line displayed in your chart matches the slope and y-intercept displayed in the cells below your chart. If we relate this back to (2), then a is the slope of the best-fit line and b is the y-intercept of the best-fit line. The "square of the correlation coefficient" is exactly that: if R is the correlation coefficient for the relationship between the independent and dependent variables, then the square of the correlation coefficient is R^2. The R^2 value is another measure of the "goodness" of fit of the best-fit line. It can be thought of as the percentage of the variation in the dependent variable that is explained by the independent variable.

7. The "uncertainty in output" is the uncertainty in the output of the model. This is the same uncertainty we have calculated earlier, the standard error. Use this uncertainty and the following rules when using the equation of the best-fit line in (2) as an empirical model. Calculate the output of the model and round the result so that it has the same number of significant digits as the input. Multiply the standard error by 2 (to get 95% confidence) and round up to the least significant digit of the output. For example, see Figure 1. We can predict crop yield given growing-season precipitation. If $P = 6\bar{0}0$ mm (which has two significant digits), then the mean value of expected crop yield would be $Y = 6\bar{0}0 \times 1.871 + 2116 = 3200$ kg ha^{-1}. The uncertainty in the model output is 644 kg ha^{-1}. Multiply 644 by 2 and round up, the result is 1300. This gives a prediction of $Y = 3200 \pm 1300$ kg ha^{-1} for $P = 6\bar{0}0$ mm.



H33 {=LINEST(B3:B38,D3:D38,TRUE,TRUE)}

	A	B	C	D
1	year	yield	biomass	precipitation
2	(year)	(kg/ha)	(kg/ha)	(mm)
3	1980	2285	6004.6	371
4	1981	2130.3	4614.3	360
5	1982	3935.9	8724.7	492
6	1983	2154.4	5669.3	472
7	1984	2419.5	6154.4	504
8	1985	2018	5750	272
9	1986	3720.7	7742.9	638
10	1987	3711.7	7737.3	578
11	1988	1301.1	4687.9	248
12	1989	2522.5	6420.6	406
13	1990	3005.4	7330.2	683
14	1991	2556	6123.5	339
15	1992	2862.3	5724.6	448
16	1993	3039.4	6181.4	1034
17	1994	3915.1	8485.7	413
18	1995	2801.1	7002.9	303
19	1996	3731.3	8396.8	464
20	1997	3190.7	7152.4	296
21	1998	2924.2	7310.5	627
22	1999	3845.6	8360	479
23	2000	2585.1	6305.2	372
24	2001	2856.8	6594.5	434
25	2002	3206.3	7456.6	398
26	2003	2536	6839.8	290
27	2004	4087.5	8491.4	554
28	2005	3699.4	8407.8	457
29	2006	3293.9	7660.3	352
30	2007	3027.2	7966.3	435
31	2008	4096	8194	740
32	2009	4115.8	9076.1	508
33	2010	2768.9	7286.7	817
34	2011	3114	7077.2	497
35	2012	2062	5354.3	248
36	2013	2020.2	6085.1	359
37	2014	3508.2	7984.9	722
38	2015	3873.6	8246	887

	H	I
33	1.870761046	2116.344277
34	0.593376346	307.7467507
35	0.226213455	644.4069776
36	9.939766369	34
37	4127590.889	14118851.99

Figure 1: Chart of soybean yield in kg ha^{-1} as a function of precipitation in mm, along with the other information concerning the best-fit line as found by the function `LINEST`.