Handout 3
36-601 Fall, 2014

Finding the Best Transformation

From UNDERSTANDING DATA
ERICKSON + NOSANCHUK

In the last chapter one very important transformation, the log, was described, not only for its own sake but also to help illustrate how transformation works in general and what some of the reasons for transformation are. This chapter is necessary because batches* can have many different shapes. So we need to know several kinds of transformations, each a fit for a different shape; and we need to know how to choose among them. We explain the use of several transforms; we offer two procedures that may help you to find an appropriate transformation quickly; and we give some guidelines for deciding when a transformation is appropriate.

Other Useful Transformations

Relax: we are not about to dazzle you with a mass of weird and unfamiliar functions. Almost all the shapes you'll run into can be handled quite adequately with logs or one of a very few familiar alternatives. First, consider batches that straggle up, similar to those in the last chapter.

Sometimes when we try logs on such data, we find that we have overshot the mark: the new batch now straggles down because we've overcorrected. So we turn to a more moderate correction than the log transform, the square roots of the data. This is easy to handle, as many statistics books and books of statistical tables have square root tables more than adequate for exploration, and many inexpensive electronic calculators will give square roots. Suppose instead that logs undercorrect the data, that is, the batch still straggles up after logging. Then we want something stronger and might turn to the negative reciprocal (or $-1/x$), a somewhat stronger transform. You use negative reciprocals because that keeps the order of the numbers the same. For example, 2 is bigger than 1 but when you take reciprocals 1/2 is smaller than 1/1; using $-.5$ and -1 is fine because $-.5$ is bigger (less negative) than -1.

Let's try these three transformations on the same batch to see how they work. Table 6.1 gives data on the number of housing starts in the

*BATCH = A DATA SET

Table 6.1
Private Nonfarm Housing Starts, U.S.A., 1966-1968

	Number of Units Started		
Month	1966	1967	1968
January	78,500	57,700	79,800
February	74,800	60,200	82,800
March	115,900	89,200	123,900
April	138,600	112,000	159,100
May	126,700	129,700	139,000
June	118,200	123,400	136,000
July	97,600	124,000	137,300
August	99,600	123,600	134,500
September	86,900	119,500	132,400
October	74,400	133,100	138,100
November	71,400	116,800	125,100
December	58,900	79,100	95,500
Entire Year	1,141,500	1,268,400	1,483,600

Source: Bureau of the Census.
NOTE: Components may not add to totals due to rounding.

United States, 1966–68. The data have been stemmed-and-leaved and plotted in Table 6.2 so we can get a better look at them. Some things are clear at once; in Table 6.1 we can easily see that the time of year makes a big difference to housing starts, which is simple enough to explain: it's much harder to pour foundations in January than in July in most of the United States. In Table 6.2 we see that the level of housing starts went up quickly, and that none of the months in any of the years seemed different enough from the others to be outliers. Looking a bit more closely we see that the shapes of the batches are interesting. The 1966 batch straggles upward just a bit and 1967 and 1968 straggle down rather substantially. The fact that the shapes are different from year to year is intriguing and suggestive, and we will discuss that more later.

Take 1966. It straggles up; and in fact most of the social science data you will ever see will straggle up. This is in part because we often deal with variables which have clear floors but where ceilings are vague or absent: income, age, population, gross national product and so on. So the transformations that make upwardly straggling data symmetric are particularly useful. Square roots, logs, and negative reciprocals are given for the 1966 batch in Table 6.3. With this batch of numbers, square roots are not a strong enough transformation; the batch still straggles up. The logged data look good, essentially no straggle still remains. Taking negative reciprocals is fine for the middle mass of the data, but makes the data beyond the quartiles very asymmetric. Overall, the logged data are probably the best compromise, though as usual, there is some judgement involved. More generally, you can

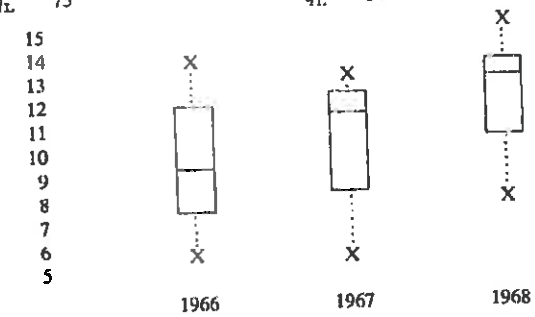
Table 6.2

Private Nonfarm Housing Starts, U.S.A.

1966		1967		1968	
				15	9
				14	
				13	967528
13	9	13	30	12	45
12	7	12	3440	11	
11	68	11	27	10	
10	0	10		9	5
9	8	9		8	30
8	7	8	9	7	
7	9541	7	9	6	
6		6	0		
5	9	5	8		

stems: ten thousands
leaves: rounded to nearest thousand

q_U	117	q_U	124	q_U	138
Md	93	Md	119	Md	134
q_L	75	q_L	84	q_L	109



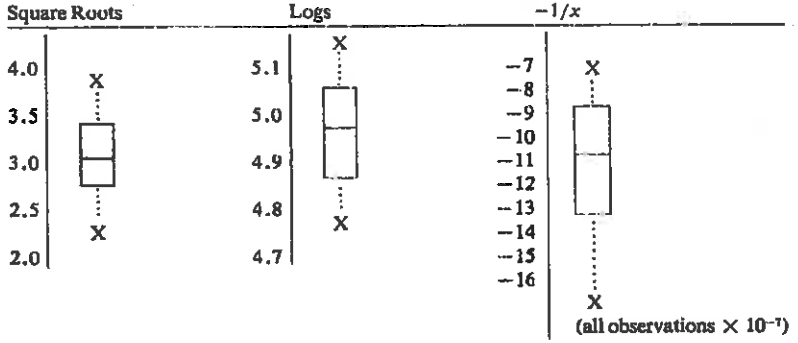
see that these transformations have different effects and so will be appropriate for different data sets.

What do you do if a batch straggles down? You want to spread out the higher values and pull the lower ones together to make the batch symmetric. This is easy to do by trying powers of the batch numbers: x^2, x^3, x^4, or whatever it takes. This is pretty simple so we have not bothered to work out examples for you. The antilog transformation corrects for downward straggle too (being the opposite of logs) but is awkward if the data are large numbers; the antilogs can really get enormous. Consequently this transform will rarely be used as a way of handling shape, though it is used in getting results from log transforms back to the original data units. For most data which straggle down, finding squares or cubes is likely to be adequate. You should feel free to try powers or roots or negative reciprocals or any other function of the data that will make them amenable, so become familiar with enough options to be flexible.

Table 6.3

1966 Housing Starts, Transformed

Square Roots		Logs		$-1/x$	
3	70, 60	5.1	40	−7	29
3	40, 40, 20, 10	5.0	670	−8	56
2	94, 82, 76, 70, 64	4.9	940	−9	
2	46	4.8	875	−10	02
		4.7	7	−11	5
				−12	7
				−13	35
				−14	1
				−15	
				−16	9
stem: hundreds		*stems*: units and tenths		*stem*: ten millionths	
q_U	340	q_U	5.07	q_U	−86
Md	302	Md	4.97	Md	−109
q_L	276	q_L	4.88	q_L	−134



Choosing an Appropriate Transformation

By now you have seen, and can easily work with, several kinds of transformation. Different transformations have different effects on data; some spread the data out where values are high and condense them where values are low, some do the opposite. In addition some change spread a lot more than others. How do you decide which to use? Let us first consider choosing a transformation for one batch, and then for several related batches.

When working with a single batch you want to get it as close as possible to standard form: single-peaked, symmetric, and falling off smoothly on both sides. The middle part of the data is particularly important and, other things being equal, we choose that transform which symmetrizes the middle. Thus, we chose a log transform for the Canadian population data (Table 5.5). But if we have two transforms that are pretty good for the middle we can choose between them by looking at how they handle the observations beyond

the quartiles. Consequently, we rejected the negative reciprocal in favour of logs for our housing start data (Table 6.3).

Now we want to show how you can quickly choose symmetrizing transformations for whatever batches come your way. We want a better choice procedure than hit-or-miss, trial and error, because that takes time and work. The first step is to get an orderly idea of what transformations do what kinds of things. In Table 6.3 we saw that square roots, logs, and negative reciprocals all correct for upward straggle to varying extents; the square root transform is weaker in its effect than logging, which is in turn weaker than negative reciprocals. We can also see with a little thought that powers of x correct for downward straggle and do so to varying extents. Consider a little example: the numbers 5,6,7. If we transform these by squaring we get 25, 36, and 49. The smaller pair are one unit apart in the original form and are 11 apart in the squared form. The larger pair are again one unit apart in the original form, but are further apart than the smaller pair after squaring: 13 vs. 11. And these are numbers that are not very different in the first place. Now suppose we try x^3, getting 125, 216, and 343. Now the smaller pair are 91 units apart and the larger pair are 127 apart. Again the larger numbers have been spread out relatively more, with a higher power of x exaggerating the effect. In general, the higher the power of x the more the larger numbers are stretched compared to the smaller ones (or the more the smaller ones are squeezed together compared to the larger ones).

We can sum up this useful information by what Tukey calls the *ladder of transformations*:

$-\frac{1}{x^2}$	$-\frac{1}{x}$	$\log x$	$\sqrt{x}$	x	x^2	x^3	antilog x
STRONGER		MILD		NO SHAPE CHANGE	MILD		STRONGER

CORRECT UPWARD STRAGGLE ← → CORRECT DOWNWARD STRAGGLE

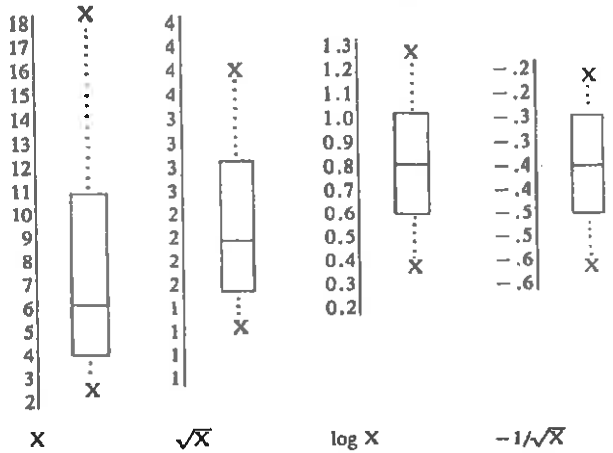
Moving to the right spreads out larger x values and clumps the smaller ones; moving to the left spreads out smaller values and clumps the larger ones. Suppose you start with original data that straggle down. You might try x^3 and, alas, this straggles up. Try x^2 then, of course. One or two stabs are usually enough to identify a good transformation. But even this sounds like work! With bad luck we could end up doing several transformations on a batch, taking up time that we would rather spend on thinking about what the data mean. There is an easier way to handle this problem. Think back to the numerical summary. We can get the extremes, quartiles, and the median of a batch quickly by now and we will have gotten them anyway when we first looked at the raw data, so that is no problem. Transforming five numbers (and perhaps a few others for added reliability if there is time)

Table 6.4

Finding a Good Transform Using Numerical Summaries

Canadian Population = X

	X	$\sqrt{X}$	Log X	$-1/\sqrt{X}$
X_U	18.24	4.27	1.26	−.23
q_U	10.95	3.31	1.04	−.30
Md	6.29	2.51	.80	−.40
q_L	4.01	2.00	.60	−.50
X_L	2.44	1.56	.39	−.64



is not hard or time-consuming. And the summary numbers are good guides to the batches they summarize: if the transformed quartiles and extremes look symmetrical around the median then we probably have a transformation that will work on the whole batch. To get a quick feeling for how symmetric the summary numbers are after transformation we again can use a fast and familiar tool: box-and-dot plots.* Consider Table 6.4, in which we look at trial transformations on the Canadian population data (because we already know what the "right" answer is). We can see very easily that the raw population figures straggle up; the square root figures are a bit more symmetrical but still straggle up; and the negative reciprocals of the square roots straggle down slightly. Clearly the square root transform undercorrects and the negative reciprocal overcorrects, while the log transform is "just right." As we have already seen, the message from the summary numbers is accurate in this case since a full-scale log transform of all the population figures was very effective.

Using summary numbers and the ladder of transformations is easy and fast and works most of the time, which is just as well, since it wouldn't be worth much otherwise. Once in a while you will run into intermediate cases.

* ...-Box Plot

For example, you might find that x^2 undercorrects your batch and x^3 overcorrects it. What then? If you are extremely precise you may want to find the transformation that symmetrizes the data exactly — but that will be something in between x^2 and x^3, say $x^{2.314}$. Now you can do this if you insist (you do it with the help of logs) but it is a lot of work and only rarely worth the trouble, even in confirmatory work. For most exploratory work we suggest not taking the trouble. Decide which simple transform looks better and go ahead with it.

If the decision between the two transformations is really close, you might pick the one that is easier to do (that's quite reasonable) or the one that makes more sense. For example, we saw that the U.S.A. population figures were slightly overcorrected by a log transformation, and one could argue for taking the square root of the batch or possibly even leaving it alone. But the log transform makes a lot of sense for populations so we used it anyway, and then we got some discussion out of the slightly imperfect nature of the transformation — the transformed data straggled down a little and that led us to think about declining immigration rates and other things interfering with the theoretically normal pattern of constant increase. Like a lot of data analysis, the choice of a good transformation is often a matter of judgement and taste. Besides, we can do more than one analysis.

Choosing a Transformation for Several Related Batches

Balancing Batches

For one batch we try to even up within-batch spread, so that the upper and lower halves of the batch straggle about the same amount. For several related batches we try to even up spread within each batch and between batches as well. The ideal multi-batch transformation will result in all the batches having medians well centred between both the quartiles; the extremes equally far from the quartiles; and each of the batches having similar spreads even if their levels differ. How often do you get a set of batches which can be transformed so neatly? Well, let's get back to the real world.

It can be tricky to decide what the best overall transformation for a set of batches is. The best overall may be the one that makes all the batches balanced, or at least makes the most of them balanced. This is just our single-batch criterion expanded to several related batches. Consider the five related batches in Table 6.5. Each batch is defined by a "stage" of economic growth, with the stages defined by levels of Gross National Product per capita for about 1957. The batch entries are the number of students enrolled in higher education per 100 000 population, for about 1960. (The exact values of these figures would be different now, no doubt higher on the whole, but the

Table 6.5

Enrollment in Higher Education at Different Levels of Economic Growth

Country	Higher Ed. per 100,000	Country	Higher Ed. per 100,000	Country	Higher Ed. per 100,000
Stage I GNPC 45–64		Stage III GNPC 108–239		Stage IV GNPC 262–794	
Nepal	56	Iran	90	Mexico	258
Afghanistan	12	Paraguay	188	Colombia	296
Laos	4	Ceylon	56	Yugoslavia	524
Ethiopia	5	Indonesia	62	Hong Kong	176
Burma	63	Rhodesia and Nyasaland	3	Brazil	132
Libya	49	Egypt	399	Spain	258
Sudan	34	Morocco	40	Japan	750
Tanganyika	9	Surinam	109	Jamaica	42
Uganda	14	South Korea	397	Panama	371
		Iraq	173	Greece	320
		Nicaragua	110	Malaya	475
		Taiwan	329	Costa Rica	326
		Saudi Arabia	6	Romania	226
Stage II GNPC 70–105		Ghana	29	Lebanon	345
		Syria	223	Bulgaria	456
Pakistan	165	Tunisia	64	Malta	142
China (Mainland)	69	Albania	145	Chile	257
India	220	Algeria	70	South Africa	189
South Vietnam	83	Peru	253	Singapore	437
Nigeria	4	Ecuador	193	Trinidad and Tobago	61
Kenya	5	Guatemala	135	Cyprus	78
Madagascar	21	Honduras	78	Poland	351
Congo (Leopoldville)	4	Barbados	24	Uruguay	541
Thailand	251	El Salvador	89	Argentina	827
Bolivia	166	Philippines	976	Hungary	258
Cambodia	18	Turkey	255	Italy	362
Haiti	29	Portugal	272	Ireland	362
		Mauritius	14	Puerto Rico	1,192
		British Guiana	27	Iceland	445
		Dominican Republic	149	U.S.S.R.	539
				Venezuela	355
Stage V GNPC 836–2577				Austria	546
				Czechoslovakia	398
Netherlands	923	New Zealand	839	Israel	668
West Germany	528	Australia	856	Finland	529
France	667	Sweden	401		
Denmark	570	Luxembourg	36		
Norway	258	Switzerland	398		
United Kingdom	460	Canada	645		
Belgium	536	United States	1,983		

GNPC = Gross National Product per Capita, U.S. dollar equivalent, circa 1957.

Higher Ed. per 100,000 = number of students enrolled in higher education per 100,000 of total population; primary and secondary schools, adult education and technical training excluded.

Source: Bruce M. Russett et al., *World Handbook of Political and Social Indicators*; Yale University Press, New Haven 1964. Table B.2, pp. 294–298.