FES 524: Natural Resources Data Analysis

Reading 4.2: Transformations

Contents

1 Transformations

Transformations are still commonly used in statistical analyses. It is important to understand why we sometimes use transformations as well as modern alternatives to transformations.

The use of transformations is generally to change the scale of the response variable. Because the spacing changes, transformations also lead to changes in relationships among variables. Make sure you can recognize what amounts to a transformation vs a location or scale shift after seeing the examples below.

1.1 Changes to relative spacing

Changing the scale of the response variable with a transformation means we are changing the relative spacing among observations.

Figure 1 is a graphic showing two plots. In the top plot you can see observations from a variable Y labeled as A-O based on the rank of the particular value (i.e., A represents the lowest observed value and O represents the highest observed value). The bottom plot shows the same set of values after taking the natural logarithm of Y.

You can see that the relative spacing among the observations changed after taking the natural logarithm of Y compared to the relative spacing among the original Y values. Taking the logarithm of a variable is a

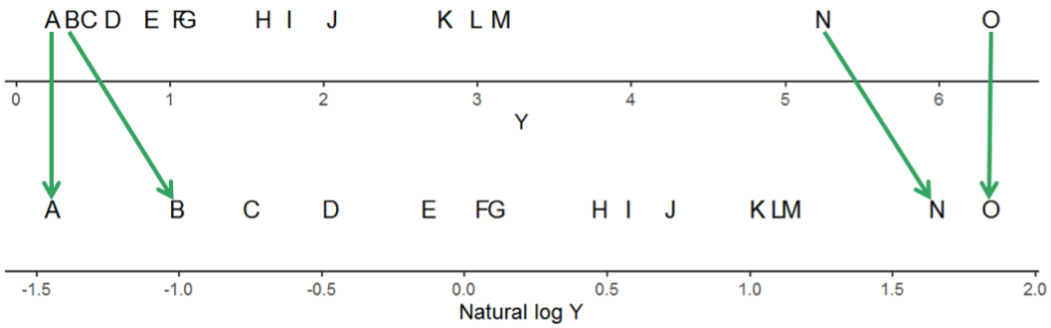

Figure 1: Two number lines, one showing the original space with specific points labeled A - O, the second showing the log-transformed space and where points A - O land on the transformed line.

transformation. Notice that we are not talking about what happens to specific values of Y when we pass them through a certain function, but actually what happens to the whole number line and how a transformation stretches or squishes the original space. For this example, all the space that did lie in the interval (0, 1) got stretched out to $(-\infty, 0)$, while all the space that did lie above 1 got sucked back in, closer to the origin. This dependence between which region of the original Y axis we are talking about and how much that region gets squished or stretched is what defines a *non-linear transformation.*

We can see something similar when using the square root of Y. The relative spacing of the observations changed compared to the original spacing of Y. Taking the square root of a variable is another transformation.

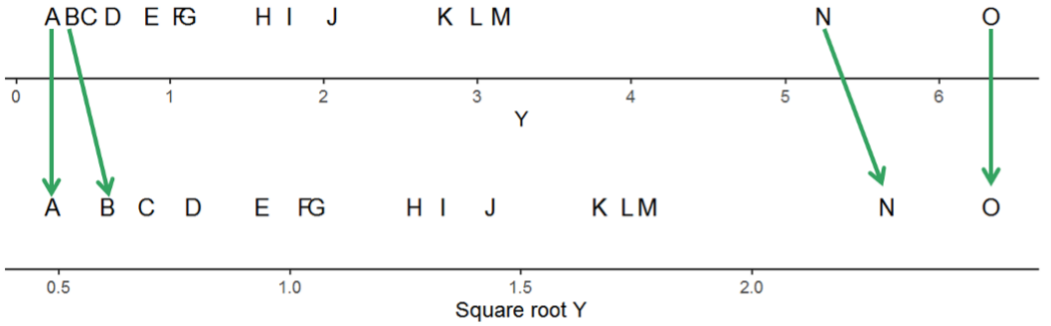

Figure 2: Two number lines, one showing the original space with specific points labeled A - O, the second showing the square root-transformed space and where points A - O land on the transformed line.

Figures 1 and 2 show nonlinear transformations. On the other hand, linear transformations, such as additive shifts, do not change the relative spacing. In Figure 3 you can see that the values on the axis have changed due to a location shift but the spacing among observations is unchanged. Adding a value to Y is a linear transformation, so it doesn't change relative spacing.

Multiplying by a certain constant, or *scaling* the data, is also a linear transformation. While the values after multiplying Y by 3 have changed, the *relative* spacing of observations is still the same (Figure 4).

1.2 Changing relationships

We will generally refer to the transformations that change relative spacing as *transformations* in this class, specifying the special case of *linear transformations* when necessary. A (non-linear) transformation changes

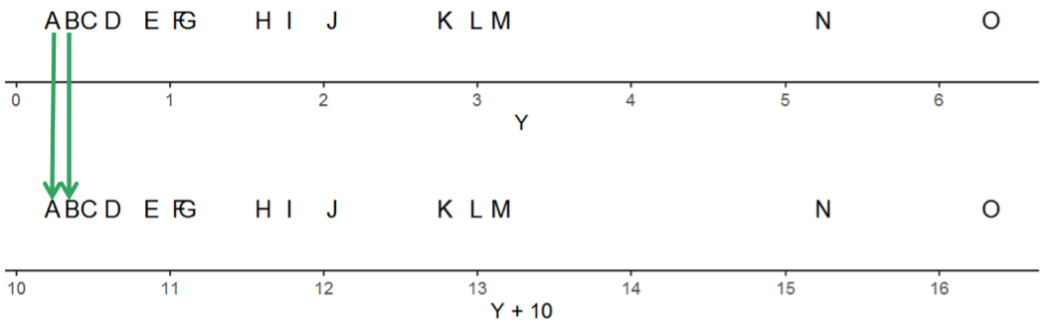

Figure 3: Two number lines, one showing the original space with specific points labeled A - O, the second showing a linear transformation that shifts the number line by 10.

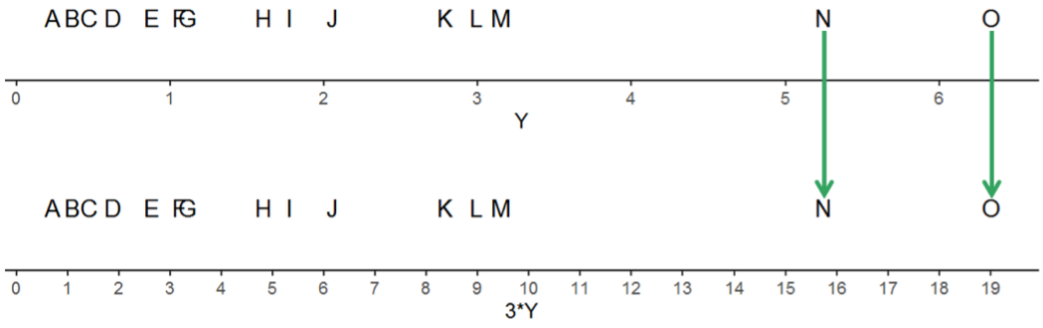

Figure 4: Two number lines, one showing the original space with specific points labeled A - O, the second showing a linear transformation that scales the number line by a factor of 3.

relationships between variables, while a linear transformation does not. This is demonstrated in Figure 5, where four different variables based on Y are plotted against the same X variable. Compare the subplot of the raw values of Y in the upper left-hand corner of the plot with the other three subplots. The strip label indicates how the original Y values were changed for each subplot.

Only when Y is transformed does the relationship of Y vs X change (shown in the lower right subplot).

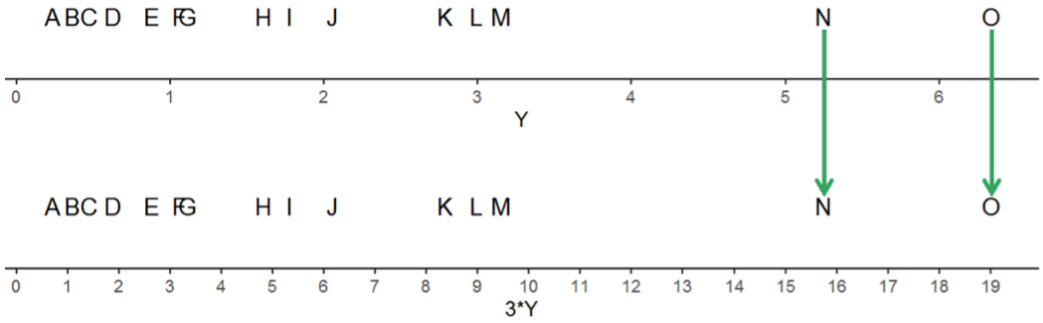

Figure 5: Panels of Y plotted against X. The first three show linear transformations of Y, while the final, in the lower-left, shows a non-linear transform of Y.

1.3 Why transform?

Most often, investigators do transformations in an attempt to meet assumptions of the linear model that aren't met when using the raw data. Most common transformations you may have heard of are what are called "variance stabilizing transformations". While many practitioners focus on normality of the errors when thinking about transformations, you know that issues of nonconstant variance is a bigger problem for linear models than departures from normality. Addressing issues of nonconstant variance is generally the purpose of transformations.

1.4 Variance-stabilizing transformations

There are many different transformations available that can help stabilize variances.

Here are just a few examples:

- Reciprocal: Y^{-1}
- Square: Y^2

Square root: $\sqrt{Y}$

Log base 10: $\log_{10}(Y)$

How do you choose a transformation?

The first thing to ask is what scale you are interested in making inference on. If you want to talk about things on the original scale of the data, will a back-transformation allow you to do that? Or, alternatively, maybe there is a transformation that allows for a natural interpretation. For example, working with the cube root of volume can be useful in some fields where it makes sense to talk about the results as lengths. Similarly, log base 2 can be useful in genetics and genomics where there are natural duplicating processes

based on making copies of existing entities (e.g., DNA fragments). The log base 2 can be translated into the number of duplication events.

Below are a couple of examples of the kind of inference that can be made after different transformations of the response variable:
- Differences in mean square root biomass

- Estimated slope between mean growth cubed and rainfall

Are such results interesting? This is a question you always need to be asking yourself when you are considering using a transformation.

I would argue that, no, the example results above are not very interesting because they are not easily interpreted. I see the focus on transformations as only a way to meet model assumptions as pretty old-fashioned. I still see a lot of discussion online about Box-Cox and other transformations, but for practical problems like we are doing in this class where the goal is estimation, these recommendations may be obsolete. We want estimates on an interpretable scale, not to force things into a certain shape so we can meet model assumptions.

That being said, do note that transformations can be more useful when the goal is prediction instead of estimation. Even in that case, though, we often have other options to consider.

Unlike other transformations, the log transformation can often be quite useful, particularly in regression contexts. We will talk more about the log transformation, the log-normal distribution, and how to make inference after a log transformation in the next section.

Another potentially useful transformation when working with continuous proportion data is the logit transformation (i.e., log odds). We will come back to this transformation when we get to logistic regression later in the class, but if you are struggling with meeting assumptions with continuous proportion data, take a look at the Warton and Hui 2011 paper on transformations for proportions.

1.5 Alternatives to transformation

In modern statistics we have many alternatives to doing transformations. Below are three options, but there are more than this.

- For relaxing assumption of constant variance, we can "extend" the linear model. In R this can be done using the `gls()` function in R from package `nlme`. Linear mixed models can also be extended in the `lme()` function. We will see a brief code example of this in lab this week and later this quarter.
- Generalized linear models (GLMs) using distributions with an appropriate mean-variance relationship are a common analysis tool for some kinds of variables, such as counts and counted proportions. You will be introduced to GLMs later this quarter.
- Robust regression can be used if the main problem you are having is extreme outliers. There are also options to deal with nonconstant variance and correlations among errors, although these modeling approaches have not been as widely adopted in natural resources fields as they have been in, e.g., economics.

2 The log-normal distribution

The log-normal distribution is the distribution of a continuous variable whose logarithm is normally distributed. I will be talking about the natural logarithm, but this is true for any logarithm base. The log-normal distribution is the distribution you are using if you log transform a response variable, so it is

important to understand this distribution in order to figure out when it is going to be useful to you in difference scenarios.

Features of the log-normal distribution: 1. The support of the log-normal distribution is strictly positive. This means the range of values that a variable following the log-normal distribution can have are positive and cannot contain 0. 2. A variable following the log-normal distribution is continuous. It is not, for example, integer-valued like counts. 3. A variable (or model residuals) that comes from the log-normal distribution will show right skew, even if the skew is minor. 4. The variance of a log-normally distributed variable increases with the mean. This is something you would see in, for example, a residuals vs fitted plot.

Seeing some examples of data drawn from a log-normal distribution can help show these features.

Figure 6 showing a log-normal distribution. The raw data is on the left and the log-transformed data is on the right, showing that it is approximately normally distributed after transformation. The range of the variable is listed in the title so you can see the support. In this case, which is common with log-normally distributed variables, the range is quite large. This particular log-normal distribution is very right-skewed.

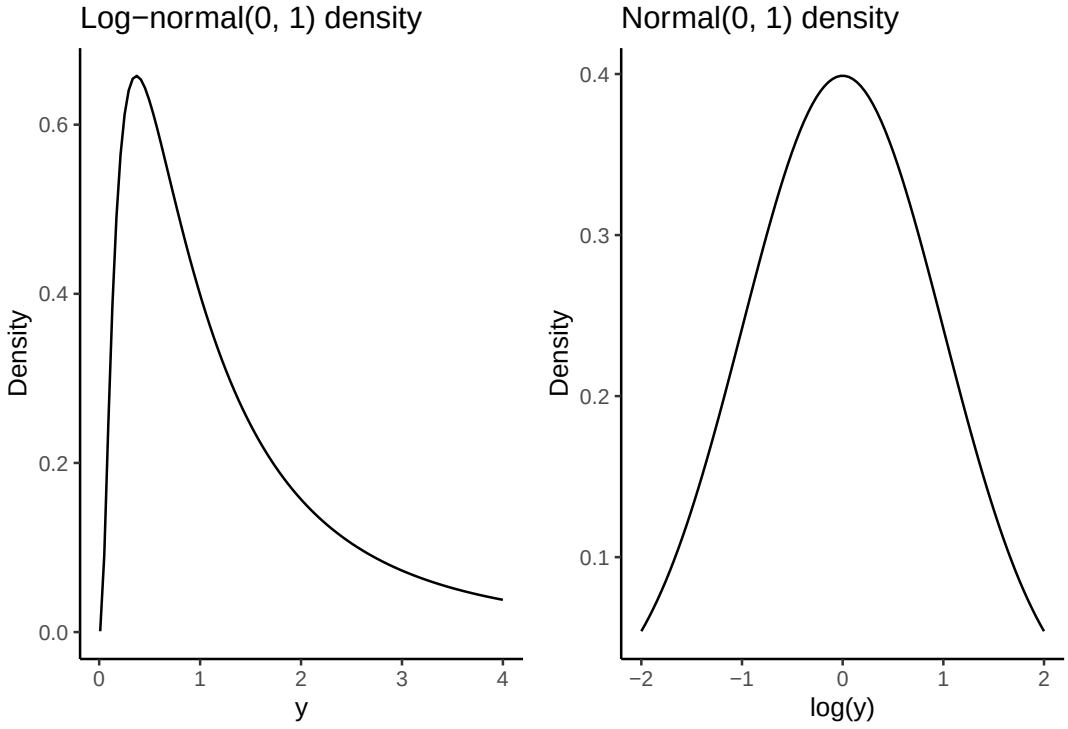

Figure 6: Log-normal(0, 1) density compared to a normal(0,1) density. If a random variable Y is distributed log-normal(μ, σ^2), then $\log(Y)$ is distributed as $\mathcal{N}(\mu, \sigma^2)$.

2.1 What about (true) zeros?

Now that you know more about the log-normal distribution you can see that this distribution does not contain 0 values. If you have 0 values in a variable that you want to log-transform you will need to carefully think through your options. I unfortunately see people thoughtlessly add something to the 0 values and go ahead with a log transformation without spending time thinking about what the 0 values in the data mean and whether or not they can justify adding something to those values. Adding 1 is particularly attractive to people in this situation, which can fail rather spectacularly for certain types of data.

You may not be surprised to learn that the actual steps you should take are more complicated than simply adding an arbitrary value and transforming the variable. Finding a reasonable analysis approach when you

have 0 values along with positive, continuous data involves careful thought and scientific expertise. It is possible that you will end up justifying that you can add some value and shifting your entire distribution away from 0. However, you first have to make sure other options are not available for your particular situation and that adding something to the zeros makes scientific sense.

If you ever run into a situation where you have 0 values along with right-skewed data and the variance increasing with the mean, you need to do more reading on possible analysis approaches. One place to start is with this blog post, https://aosmith.rbind.io/2018/09/19/the-log-0-problem/, and then going to some of the other discussions linked to from there.

2.2 Multiplicative models

When working with a log-transformed response variable, the model is an additive (linear) model on the log scale. Here is what the statistical model looks like for a two group log-linear model:

$$\log(y_i) = \beta_0 + \beta_1 I\{g_2\}_i + \epsilon_i$$

where $I\{g_2\}_i$ is an *indicator* variable indicating whether observation i belongs to group 2. If so, $I\{g_2\}_i = 1$.

One of the reasons the log transformation is so attractive, though, is that we can still make inference on the original scale and not the log scale. We back-transform from a natural logarithm transformation by exponentiating both sides of the model.

This results in a multiplicative model and estimates need to be expressed multiplicatively (i.e., times or percentage changes) instead of additively. On the original scale, the model can be written as

$$y_i = \exp\{\beta_0\} \times \exp\{\beta_1 I\{g_2\}_i\} \times \exp\{\epsilon_i\}.$$

2.2.1 Means on the log scale become medians on the original scale

After back-transformation, all inference from your model of a log-transformed response variable will be about *medians* instead of means.

The reason for this has to do with symmetric distributions. If you are interested in more details, you can see more discussion on this topic in the Statistical Sleuth chapter 3.

For real distributions, the estimated median based on the model and the observed median from the original data are not identical. This is because we are making inference to the population median but, of course, we only have a sample. However, we can still make inference to medians. Alternatively, you can talk about *geometric* mean if that is of interest in your field.

2.3 Multiplicative differences

When back-transforming additive differences among groups after a log transformation, remember that results will be expressed as *estimated ratios of medians*. If working with slopes (applicable when you have continuous explanatory variables) instead of group differences (categorical explanatory), results are about an estimated multiplicative change in the median of Y for some change in a continuous variable.

2.4 Percentages

When working with a percentage as a response variable, there can be a lot of ambiguity between whether the change expressed is additive (a percentage point change) or multiplicative (a percentage change).

One way to clarify is to use the term "percentage point change" to indicate additive changes for simple differences; i.e., 40% - 20% = 20 percentage point change.

If talking about a multiplicative change, use the term percentage change. In a multiplicative change, the estimated change is relative to a baseline.

2.5 Percentage change vs ratios

We can either use percentage changes or ratios to express multiplicative differences. A lot of us tend to default to using percentage changes. However, percentage changes are not symmetric and so we can confuse the reader if we are not careful.

For example, the inverse of a 100% increase from 2 to 4 is a 50% decrease from 4 to 2. An increase from 3 to 4 is a 33% increase but the inverse, a decrease from 4 to 3, is a 25% decrease. If you change the order of a comparison you will always need to recalculate your percentage changes.

This is not true for ratios. They are always exactly symmetric. Going from 2 to 4 is a doubling (2) and going from 4 to 2 is a halving (1/2). Going from 3 to 4 is an increase by a factor of 4/3 while going from 4 to 3 is a decrease by a factor of 3/4. To change the order of the comparison you only need to invert that ratio.

This feature of ratios is why Frank Harrell argues that we should be using ratios and not percentage changes when reporting multiplicative results. You can read his argument here: http://www.fharrell.com/post/percent/.