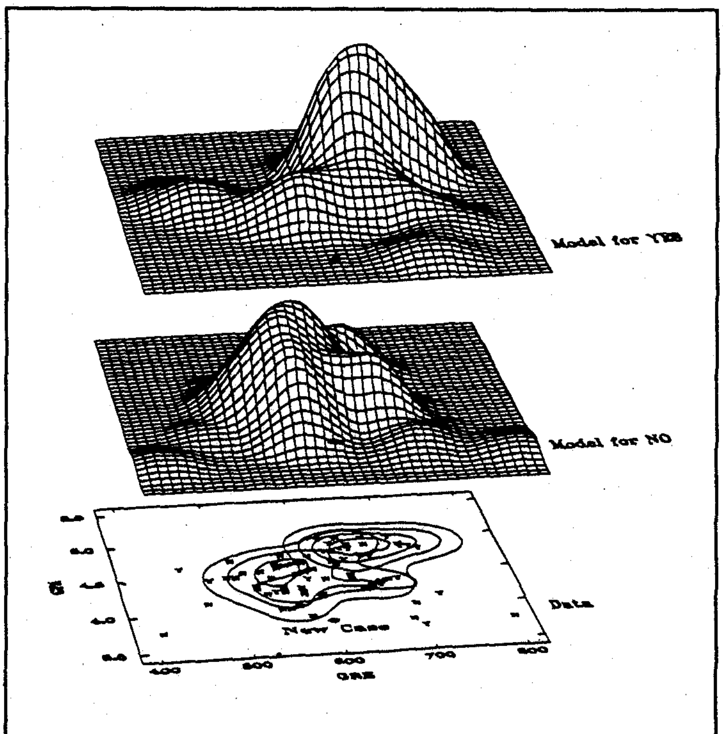
DESKTOP DATA ANALYSIS WITH SYSTAT



LELAND WILKINSON
GRANT BLANK
CHRISTIAN GRUBER

22

Transformations

Transformations (sometimes called reëxpressions) can simplify interpretation of statistical results and improve subsequent analysis. The simpler the model, the more readily we can interpret it. For example, a nonlinear relationship between an independent and the dependent variable can often be transformed into a straight-line relationship. Or, skewed residuals from a model may be made to look like a normal distribution through a simple transformation. Or, nonconstant variance in residuals can be made more uniform. Transformations themselves are not the goal of data analysis. They are means which can help you understand the information in your data more easily or meet the assumptions necessary to do statistical inference.

This chapter will deal mostly with data having positive values. There are two reasons for this. First, most problems requiring transformations occur with positive data like counts and proportions. Second, transforming negative and zero values is a non-trivial task which gets beyond the scope of this book. The one exception here is for correlations, which can be negative. We will discuss a simple transformation for correlations.

We'll begin with the problem of standardizing a variable to a selected location and scale. Then we'll discuss transformations to make a variable look normally distributed, or at least lumpy in the middle and symmetric. Then we'll cover the more general issue of transforming data so a linear model can be used, if that is possible. Finally, we'll describe a specific method for normalizing variances.

22.1 Standardizing a variable

Figure 22.3 shows DIT density plots of the male life expectancy variable from the world data in Chapter 24. The bottom density shows the raw data, which range from 40 to 75. The middle plot shows the same density after standardizing male life expectancy to have mean = 0 and standard deviation = 1. This was done by the SYSTAT command:

```
STANDARDIZE MALE / SD
```

The top plot shows the same density after standardizing so that the range of the data is 1. This was done by the SYSTAT command:

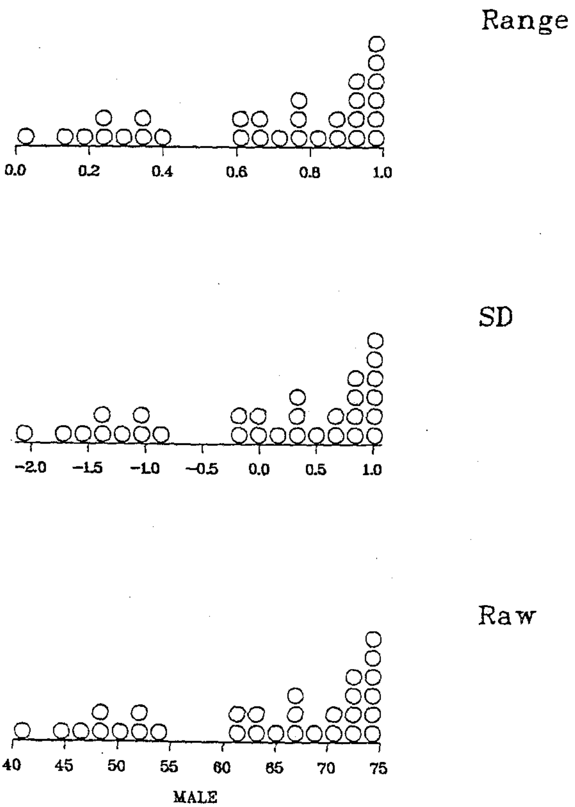
```
STANDARDIZE MALE / RANGE
```

All three plots are the same, except for the scale at the bottom of each. Some think that "standardizing" changes the shape of the data, perhaps because the transformed values, or **z-scores** (mean = 0, SD = 1) are associated with the normal distribution. Another reason for this mistaken belief may be that the shape of histograms is sensitive to the scale, so that people who look at typical histograms of their data before and after standardizing them may think something changed about the distribution because the histograms look different. Figure 24.35 in Chapter 24 shows this phenomenon. In any case, keep in mind that standardizing changes nothing but the scale on which the data are measured.

Why do people standardize data, then? There are several reasons. Social scientists often standardize data in order to compare measurements across groups. For example, an educator who grades "on the curve" wants to equate average test difficulty and range of difficulty for two different classes or tests. For example, a teacher may give the same test two different years and decide that the difference in average performance between the two classes is a nuisance. Under pressure to give roughly the same number of A's, B's, C's, and D's each year, the teacher standardizes the grades and then cuts the distribution at the same places each year. This method penalizes an outstanding class (and helps a poor class) but the procedure has a consistency about it that makes it attractive. Our experience teaching suggests that most teachers and professors tend to grade "on the curve" across years, although some may do it rather informally without resort to a computer. This is true, in our experience, even when other factors, such as grade inflation, contribute to a long-term trend in grades.

Another form of grading "on the curve" is to equate tests. That is, if a teacher gives three tests during a semester, he or she may decide to standardize each test and then average the standard scores. This gives each test equal weight in determining the final grade. Like standardizing across groups, this procedure involves inequities. Nevertheless, it prevents a single extremely difficult test passed by only one or two students, for example, from determining almost the entire final grade. Modern test theory (underlying, for example, the ACT's, College Boards, and Graduate Record tests) involves much more sophisticated methods of test equating, but this simple form of standardizing is still used widely.

Figure 22.1
Standardizing a Variable



There is another type of standardizing (called "normalizing") that is illustrated in Figure 22.2. It really belongs in the next section of this chapter, but we'll discuss it here as a preview and to illustrate what some may think standardizing actually does (which it doesn't). This is a form of the **rankit transformation.**

First, we sort MALE. Next we calculate the rankits. These are taken from the inverse normal distribution function (ZIF) of the fractiles of the data: (CASE-.5)/N. Other fractiles could be computed, such as CASE/(N+1), but the smallest value may not be zero and the largest may not be 1. Then we rescale the numbers with the mean (63.333) and standard deviation (10.864).

The lower part of Figure 22.2 shows the histogram for our normalized data with a kernel overlaid. The kernel looks like a bell curve. If you did a probability plot (PPLOT) of these transformed values, it would show a straight line. That is not surprising, because the transformation used to produce the vertical scale of a PPLOT is the same as in Figure 22.2.

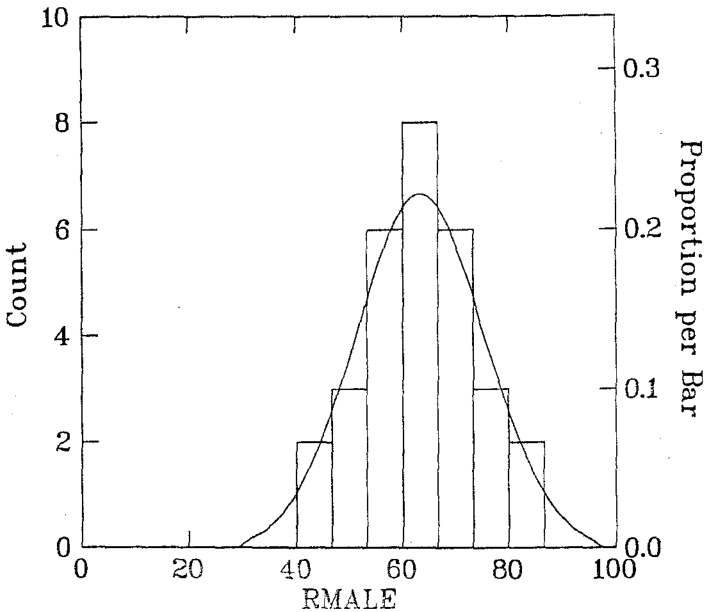
Stigler (1986) discusses the role that this transformation played in Francis Galton's theories on heredity. More than a century later, Galton's "statistical scale" continues to influence psychometrics and contemporary discussions of IQ (Herrnstein & Murray, 1994; Fienberg, Devlin, Resnick & Roeder, 1995).

22.2 Transforming a variable to look normal

If a variable is normally distributed, then we can use classical distributions like z, t, and F to make various inferences about the population mean and other parameters. Even if we are not making inferences, simple description using statistics like the mean (sample average) make more sense when the data are symmetric and lumpy in the middle. These reasons are why we seek transformations for data that look skewed to the left or right.

Figure 22.2
Normalizing a Variable

```
USE WORLD
SORT MALE
LET RMALE=63.333+10.864*ZIF((CASE-.5)/30)
BEGIN
DEN RMALE/HIST,XMIN=0,XMAX=100,BARS=15,YMAX=10
DEN RMALE/KERN,XMIN=0,XMAX=100,BARS=15,YMAX=10,AX=0,SC=0
END
```



The ladder of powers

Tukey (1977) introduced an ordering of transformations from higher to lower powers which he called the **ladder of power transformations**. Figure 22.3 summarizes this ladder. The histogram labeled "xpow=1" (third down from the top) is sampled from a normal distribution and rescaled to have all positive values. The histograms at the bottom of the figure can be transformed to normality by one of the methods in the ladder, assuming the lowest values of the data are near zero. Let's look at them in more detail.

At the top of the ladder in the figure is a **cubic transformation**. In SYSTAT:

```
LET YNEW = Y^3
```

Next is the **quadratic transformation**:

```
LET YNEW = Y^2
```

These will transform histograms like the top two into ones like the third, namely a normal. The next transformation is "no transformation" because a value to the first power is itself.

The next lower on the ladder is the **square root transformation**:

```
LET YNEW = SQR(Y), or
LET YNEW = Y^.5
```

It will transform a positively skewed histogram, like the fourth one shown, into a normal. The square root of a variable is a common transformation for counts: the number of items produced in a manufacturing study, the number of bar-presses in animal research, or the number of tumor sites in a medical study. The square root does the opposite of the square. It shrinks the larger values much more than smaller values and reduces the right-hand tail of the frequency distribution.

The next one down is the **logarithmic transformation**:

```
LET YNEW =  LOG(Y)
```

If you prefer log to the base 10, use

```
LET YNEW = L10(Y), or
LET YNEW = LOG(Y)/LOG(10)
```

It's counterintuitive, but common statistical inferential procedures don't care whether you log to the base 10, base 2, or any other base. That's because logs turn multiplicative effects into additive effects, and different bases affect only the additive constant. Adding a constant value to your data (as in the linear transformation) does not change results for hypothesis tests. Also, if you're wondering how a logarithm can be equivalent to a power of zero (well, the neighborhood of zero), Tukey and Mosteller (1977) explain this in detail. Otherwise, trust us, it works.

The logarithmic transformation is frequently applied to growth data because growth is often proportional to the size of the population. That is, the percentage change is constant. Common data for a log transform are population growth, economic data, or death rates as a function of age. As a reshaping transform, logs are more powerful than square roots, shrinking the right-hand tail of the frequency distribution even more than the square root.

Logging is common in economics, where taking logs of both the independent and dependent variables has important theoretical implications for analysis of economic production functions. It implies a constant relationship between the percentage rate of change, called constant marginal elasticity, in the independent and dependent variables, for small changes.

Next come the **reciprocal transformations**:

```
LET YNEW = -1/Y^.5, and
LET YNEW = -1/Y
```

We've omitted higher powers in the denominator because you can see that the simple inverse (-1/Y) already transforms an extremely skewed distribution. The negative sign preserves the original order of the values, but it is not necessary as long as you keep in mind the direction of the relationship with other variables.

To find the appropriate transformation, plot a variable then compare your histogram to the histograms in Figure 22.3. In the middle column, you will find the appropriate transformation that makes the variable look Normal. Version 6 includes a popup graph tool that allows you to examine these transformations in real time. Just tap the button and the data replot themselves. This helps you hunt for a transformation visually and then code the value of the power for further analysis. It's not a video game, but it's worth a few giggles.

The arcsine transformation for proportions and percents

Proportions range between zero and one, producing a truncated distribution, nonconstant variance, and skewness when the mean is not near .5. Batches of proportions look like the leftmost histograms in the lower panel of Figure 22.3. When the population proportion is .5, a histogram of a batch of sample proportions looks like the normal histogram (pow=1). When the population proportion is near zero, sample histograms look positively skewed (e.g. pow=-1) and when it is near one, sample histograms look negatively skewed (e.g. pow=3).

The arcsine transformation stabilizes the variance and spreads the tails of the distribution. The arcsine transformation doesn't have a parameter to worry about (like power), so you can simply apply it with the command:

```
LET YNEW = 2*ASN(SQR(P))
```

where P is a proportion. Percentages can be converted to proportions by dividing by 100. Another name for this transformation is the angular or the inverse sine transformation.

If you have access to the original data from which the proportions were calculated, an improved arcsine transformation due to Freeman and Tukey (1950) is:

```
LET YNEW = .5*(ASN(SQR(X/(N+1)))+ASN(SQR((X+1)/(N+1))))
```

where X/N = P, a proportion.

Fisher's *z* transformation for correlation coefficients

Correlations range between -1 and 1, producing a truncated distribution, nonconstant variance, and skewness when the mean is not near .0. Batches of correlations look like the leftmost histograms in the lower panel of Figure 22.3. When the population correlation is 0, sample histograms look like the normal histogram (pow = 1). When the population correlation is near -1, sample histograms look positively skewed (pow = -1) and when it is near +1, sample histograms are negatively skewed (pow = 3).

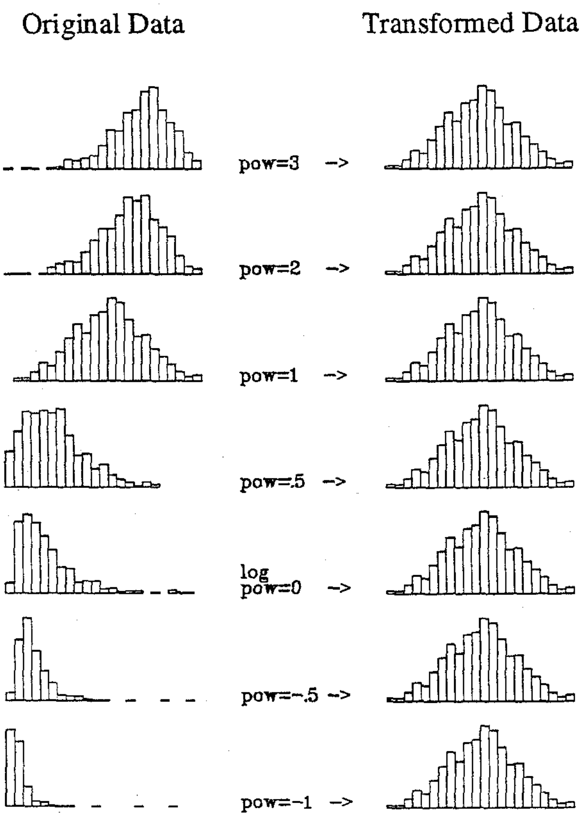
Fisher's z transformation normalizes them in a transformation without extra parameters:

```
LET YNEW = ATH(R)
```

The function used is called the arc hyperbolic tangent. You should use this transformation regularly when doing statistical analyses on a batch of correlations. If the correlations are not skewed, the transformation will leave the shape unchanged. Otherwise, it will straighten out the skewness. This transformation will work adequately for other correlations derived from the Pearson, such as Spearman. It is even satisfactory for others which aren't, provided they range between -1 and +1.

Figure 22.3
Ladder of Power Transformations

Power P	SYSTAT BASIC	Name	Notes
P	Y^P	power	DOWN: shorten upper tail.
⋮	⋮	⋮	⋮
3	Y^3	Cube	Not commonly used.
2	Y^2	Square	The highest commonly used power.
—> 1	Y^1	Original data	No transformation.
1/2	Y^(1/2)	Square root	Commonly used for counts.
"0"	LOG(Y)	Logarithm	Commonly used for financial data.
-1/2	-1/Y^(1/2)	Reciprocal root	The minus sign preserves order.
-1	-1/Y	Reciprocal	Lowest commonly used power.
-2	-1/Y^2	Reciprocal square	
⋮	⋮	⋮	⋮
– P	-1/Y^P	Reciprocal power	UP: shorten lower tail.



22.3 Transformations to linearize

Certain nonlinear functions can be made into linear ones by a simple transformation. For example, the function:

$$y = ax^{b}\varepsilon$$

can be linearized with a log function on both sides of the equation:

$$\log y = \log a + b\log x + \log\varepsilon$$

Now the parameters are different, but the function is linear and we can analyze it with a linear model routine. If data are generated with linearizable functions like this, then we can simplify things with a transformation.

Another example is the function:

$$y = b\log x + \varepsilon$$

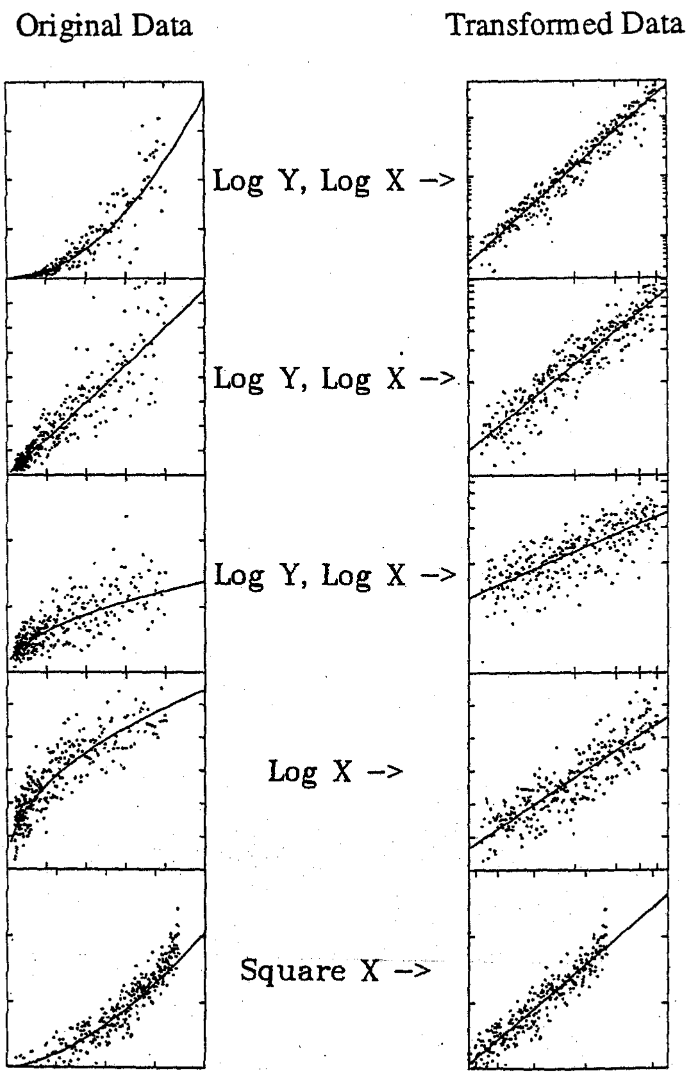
By transforming x with a log, we can fit a curvilinear plot with a straight line.

The trick is to know whether this is plausible. In many fields, you will not have prior knowledge of the nonlinear function so you will have to use a scatterplot of the data or the residuals to help find a suitable transformation. The left column of Figure 22.4 gives some examples of how nonlinear relationships appear. We added a rough smoother to the plots (i.e. DWLS) to make the curvature clearer.

The first thing to notice in the figure about this class of functions is whether the error is multiplicative or additive. In the top three plots on the left, the errors are multiplicative; the points fan out to the right, so the residuals are **heteroscedastic**. In the bottom two plots on the left, the errors are additive; the points are homogeneous about the regression line, so the residuals are **homoscedastic**. The top three plots are representative of the first type of equation above and the last two are typical of the second.

Now, the text next to each plot indicates the transformation which produces the plot on the right. The multiplicative models respond to a log-log transformation and the additive models respond to a nonlinear transformation on x only. The particular transformation chosen depends on the shape of the curved line in the left plot. When you're exploring this kind of scatterplot to figure out a transformation, it's a good idea to use DWLS or LOWESS to fit a rough smoother so that you can concentrate on these two key points: 1) are the residuals heteroscedastic? and 2) what is the type of curvature (log, square, square root, inverse)? Note that each plot on the right looks like a good candidate for a linear model. We superimposed the regression line to make that clearer.

Figure 22.4
Transformations on Scatterplots



What do we do if we have more than one predictor? Our suggestion is to do pairwise scatterplots of y with each predictor (e.g. with SPLOMs). If every one is heteroscedastic, consider logging all the variables. Otherwise, try logs on some and other transformations on the others. In the end, the most important final diagnostic tool is the residuals plot. If you did everything right, the residuals of fit to the the transformed data will look normally distributed and homogeneous across the predicted values.

22.4 Transformations for constant variance

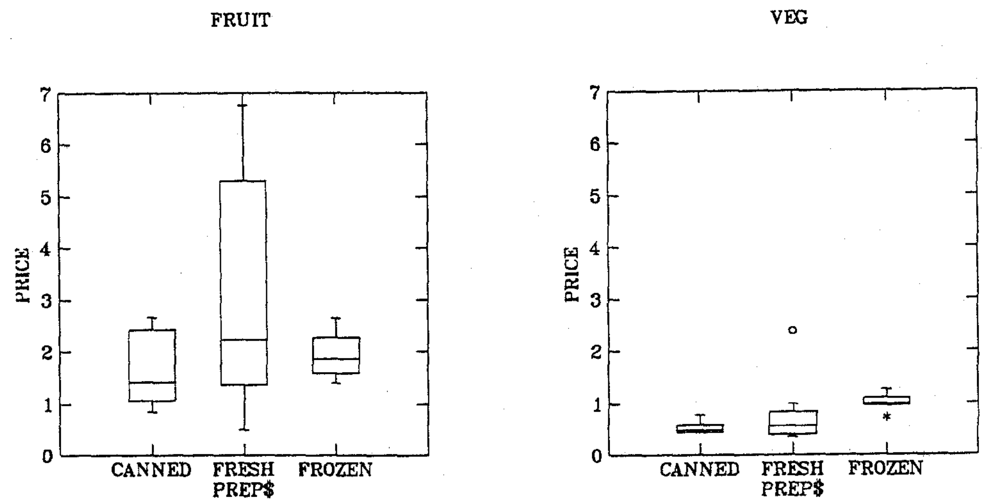
For hypothesis testing in ANOVA, we require that the data have constant variance as in regression. Heteroscedasticity can bias F-tests and produce artifactual (scale dependent) higher-order interactions. This makes the model more complex and harder to interpret. Correction of heteroscedasticity may permit a simpler interpretation at the level of main effects or lower-order interactions.

Is heteroscedasticity present?

Heteroscedasticity in ANOVA is most easily seen in box plots. In the box plots, we would like to see all boxes and all whiskers of about the same width. If some are substantially larger than others, then the data may be heteroscedastic. Figure 22.5 shows a grouped box plot of the food data from Chapter 14. We have plotted the two-way design of preparation (canned, fresh, frozen) by type of food (vegetable, fruit). The boxes vary considerably across conditions. We should consider a transformation.

Figure 22.5
Food Dataset
Box Plot of PRICE by PREP by TYPE

```
USE PRODUCE1
BOX PRICE*PREP$ / GROUP=TYPE$
```



Which transformation?

A tool called a **spread-versus-level plot** can guide us to a transformation to remove heteroscedasticity. To use this tool, we must assume that the variance (or, more generally, the spread) changes with the level. A typical relationship is that the spread gets wider as the values of the variable become larger. If we remove the relationship between the level and the spread, then we have removed the heteroscedasticity. To see the relationship we can plot a measure of the level against a measure of spread.

Suppose, for example, we assume that the standard deviation of a variable is proportional to its mean. Then if we plot the logarithm of the mean against the logarithm of the standard deviation we should obtain a line. (For the mathematical derivation of why this is so, see the notes at the end of this chapter.)

The use of a spread-versus-level plot has four steps. First, find the level and spread of each of the groups. Second, plot the logarithm of the level against the logarithm of the spread. Third, if the plot looks linear, find the slope of the best fitting line. Finally, if b

is the slope of the spread-versus-level plot, then P = 1 - *b* is an estimate of the power transformation which stabilizes the spread. As with other power transformations, if P = 0, use logarithms. (The family of power transformations is presented in Figure 22.3.)

Figure 22.6
Spread vs. Level Plot of Food Dataset

```
STATS
   USE PRODUCE1
   SAVE PROD / AG
   BY PREP$,TYPE$
   STATISTICS PRICE / MEAN,SD
   USE PROD
   LET LMEAN = LOG(ME1PRICE)
   LET LSD=LOG(SD1PRICE)
   PLOT LSD*LMEAN / SMOOTH=LINEAR,SHORT,XMIN=-3,XMAX=3,YMIN=-3,YMAX=3
```

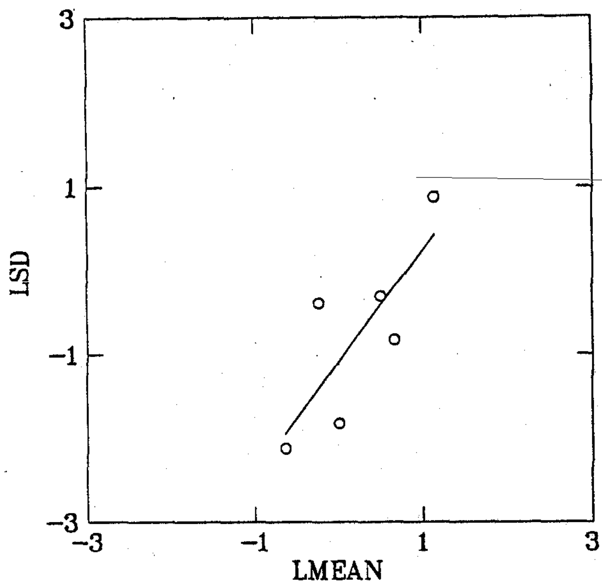


Figure 22.6 shows how to do this plot in SYSTAT. There are a few things to review. The AG option to SAVE in STATS saves the statistics in an aggregate format so that the mean and standard deviation can be on the same record. We do the calculations by both factors in the design (BY PREP$,TYPE$). The reason for the funny names for the means and standard deviations (ME1PRICE, SD1PRICE) is that the AG format allows many variables on the same record. The SYSTAT manual contains other examples of this convention. We then log the statistics and do the plot. The scales are fixed in order

to make the physical slope the same as the numerical. We've omitted the STATS output to save space.

Since the spread-versus-level plot in Figure 22.6 shows that the log of the mean and the log of the standard deviation are linearly related, we can use the slope of the line to find an estimate of the appropriate power transformation, P. Figure 22.7 gives the regression used to estimate the slope. Using the formula $P = 1 - b$ and the estimate $b = 1.337$ from Figure 22.7, we find that $P = -.337$. As with other power transformations, if $P = 0$, use logarithms.

Figure 22.7
Computing Slope for Spread vs. Level Plot of Food Dataset

```
GLM
   MODEL LSD = CONSTANT + LMEAN
   ESTIMATE
```

The output is:

```
Dep Var: LSD N: 6 Multiple R:  0.796 Squared multiple R:  0.633

Adjusted squared multiple R:  0.542    Standard error of estimate:        0.739

Effect            Coefficient     Std Error     Std Coef Tolerance      t    P(2 Tail)

CONSTANT              -1.088         0.326        0.0          .      -3.339    0.029
LMEAN                  1.337         0.508        0.796      1.000     2.629    0.058

                              Analysis of Variance

Source             Sum-of-Squares   DF  Mean-Square      F-Ratio        P

Regression                 3.774     1        3.774        6.913       0.058
Residual                   2.183     4        0.546
```

Regression transformations can also be found using spread-versus-level plots. Since regression variables are continuous, you will have to sort your data on a variable that is related to the heteroscedasticity. Heteroscedastic variances appear in many circumstances. One common circumstance occurs when they are related to some measure of the "size" of the cases; good candidates are variables that measure quantities like weight, height, income, expenses, production, population, speed, or other variables where the ratio of the smallest to the largest case is 1-to-100 or more. Then divide the sorted data into equal sized groups, as a rule of thumb at least 6 and preferably 8-10 groups. The remaining steps are identical to those we followed in the ANOVA example above. It is not always possible to find a transformation that stabilizes the variance of a regression model. A better alternative approach, weighted least squares, is presented in Section 6.4.

Keep several cautions in mind as you use spread-versus-level plots. First, they do not give you a transformation to correct all forms of heteroscedasticity. They work only when you can find some measure of spread (e.g. variance, standard deviation, interquartile range) which is proportional to some measure of level (e.g. mean, median). This is a common form of heteroscedasticity, but it is not the only form. See Judge et al. (1985, Chapter 11) for further discussion. Second, if the data in the spread-versus-level plot do not form a straight line, you may not be using the correct measure of spread. Try a different one. Third, means and measures based on them (e.g. variance, standard deviation) are sensitive to outliers. Correct outliers before you estimate the transformation. Finally, see Emerson (1991), Emerson and Strenio (1983) and Emerson and Stoto (1983) for other examples of the use of spread-versus-level plots in exploratory data analysis.

In addition to the spread-versus-level plot, several rules of thumb can help you choose a transformation. First, if the sample means seem approximately proportional to the variances, then a square root transformation often makes the variance constant. Second, if the sample means seem to be proportional to the standard deviation or to the range of each sample, then the log transform may be an effective way to stabilize the variance. Always check the results of the transformation. You will often need to try several transformations before deciding which is best.

In analysis of variance, you occasionally find that your data are from normally distributed populations but the variance is not constant. This is a serious problem because any transformation to make the variance homoscedastic also makes the data nonnormal. This is part of the Behrens-Fisher problem, and solutions to it are beyond the scope of this book. See Scheffe (1970) and Lee and Gurland (1975) for more information.

22.5 Conclusion

First, look at plots. Scatterplots, box plots, SPLOMS, and residuals plots reveal nonlinearity, while normal probability plots show non-normality. In both cases, the histogram and the ladder of power transformations (Figure 22.3) can be used to select a transformation. Scatterplots, especially residuals plots, also reveal heteroscedasticity. Spread-versus-level plots can help find a transformation to correct it.

Transformations have the greatest impact when the data deviate from a straight line in smooth envelopes. In addition, the minimum value should be near zero, and the ratio of the minimum value to the maximum value should be at least 2- or 3-to-1 before a transformation makes much difference. Sometimes you may want to subtract a constant from all values to increase this ratio. Under the best circumstances, interpretability may be improved and you may gain up to about 10% in explained variance by using the transformed data over the untransformed variables. On the other hand, irregular or multimodal data often cannot be fixed by a transformation. Furthermore, outliers and other pathological problems with your data can make the plots misleading or difficult to interpret. Clear up these problems before you try any transformations.

It is important to keep transformations in perspective. They are not a panacea for all the ills of your data. Draper and Smith caution:

> *...keep in mind that there is no guarantee that use of transformations will necessarily be better than analyzing the data on the original scale; much depends on the data. The effectiveness of a transformation is best assessed by trying it on the data and then checking the fit of the model and the resulting pattern of residuals (1981, p. 239).*

Notes

Which transformation?

Here is a more mathematical discussion of why, when we assume the mean is proportional to the standard deviation, we actually plot the log of the mean against the log of the standard deviation. Mathematically our assumption of proportionality can be expressed as:

$$\text{mean} = c\ (\text{standard deviation})^b$$

Taking logarithms of both sides we obtain:

$$\log(\text{mean}) = \log(c) + b \log(\text{standard deviation})$$

If we let $Y = \log(\text{mean})$, $k = \log(c)$ and $X = \log(\text{standard deviation})$ we have:

$$Y = k + bX$$

This is the equation for a straight line. Thus, the logarithm of the mean is linearly related to the logarithm of the standard deviation. If you would like more mathematical details, see Emerson and Strenio (1983, especially Section 3D).

Further reading

Excellent discussions of the practice of transformation are in Mosteller and Tukey (1977), Tukey (1977), and Velleman and Hoaglin (1981). These books use the word "reëxpression" in place of transformation. A variety of more complex transformations are discussed in Atkinson (1985).

References

Atkinson, A. C. (1985). *Plots, Transformations, and Regression.* New York: Oxford University Press.

Box, G. E. P. & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B, 26*, 211-252.

Draper, N. & Smith, H. (1981). *Applied Regression Analysis* (2nd ed.). New York: John Wiley & Sons.

Emerson, J.D. (1991). Introduction to transformation. In D. C. Hoaglin, F. Mosteller & J. W. Tukey (Eds.), *Fundamentals of Exploratory Analysis of Variance* (pp. 365-400). New York: John Wiley & Sons.

Emerson, J. D. & Stoto, M. A. (1983). Transforming data. In D. C. Hoaglin, F. Mosteller & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 97-128). New York: John Wiley & Sons.

Emerson, J. D. & Strenio, J. (1983). Boxplots and batch comparison. In D. C. Hoaglin, F. Mosteller & J. W. Tukey (Eds.), *Understanding robust and exploratory data analysis* (pp. 58-96). New York: John Wiley & Sons.

Fienberg, S.E., Devlin, B., Resnick, D.P., and Roeder, K. (1995). *Heritability, IQ, and Life Outcomes.* New York: Springer Verlag.

Freeman, M. F. & Tukey, J. W. (1954). Transformations related to the angular and the square root. *Annals of Mathematical Statistics, 21*, 607-611.

Herrnstein, R.J. and Murray, C. (1994). *The Bell Curve: Intelligence and Class Structure in American Life.* New York: The Free Press.

Judge, G. G., Griffiths, W. E., Hill, R. C. & Lee, T. (1985). *The Theory and Practice of Econometrics* (2nd ed.). New York: John Wiley & Sons.

Lee, A. F. S. & Gurland, J. (1975). Size and power of tests for equality of means of two normal populations with unequal variances. *Journal of the American Statistical Association, 70*, 933-941.

Mosteller, F., & Tukey, J. W. (1977). *Data Analysis and Regression*. Reading, MA: Addison-Wesley.

Scheffe, H. (1970). Practical solutions for the Behrens-Fisher problem. *Journal of the American Statistical Association, 65*, 1501-1508.

Stigler, S.M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900.* Cambridge: Harvard University Press.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Velleman, P. F. & Hoaglin, D. C. (1981). *Applications, Basics, and Computing of Exploratory Data Analysis*. Boston: Duxbury.