Topic 8: Transformation of Data

- Transformations in SAS
  - General overview
  - Log transformation
  - Power transformation
  - The pitfalls of interpreting interactions in transformed data

**Transformations in SAS**

"Data transformation" is a fancy term for changing the values of observations through some mathematical operation. Such transformations are simple in SAS and assume a form that should be very familiar to you by now:

```sas
Data Transform;
  Input Treatment $ Counts;
  Trans = SQRT(Counts);
Cards;
...```

The above code tells SAS to create a new data set named "Transform" that consists of two variables, Treatment and Counts. It then instructs SAS to create a third variable called "Trans," equal to the square root of the variable Counts, for each line of inputted data. SAS executes this **Data Step** once for each row of data, reading the values entered via the Input statement, then calculating the value of Trans for that step. If SAS does not encounter the end of the cards (" ; "), it returns for another execution of the Data Step. So, if there are twenty data lines, each containing the two input values, the Data Step executes twenty times; and the new built data set “Transform” will consist of twenty rows, each containing three variables (Treatment, Counts, and Trans). We’ve seen this before (e.g. Lab 1, Example 3).

While SAS can handle just about any mathematical operation you can throw at it, the syntax for such things is not always intuitive (it is SAS, after all). So here are some other examples that we could have used in the above sample code:

- `Trans = Counts**3;` Raises Counts to the power of 3 (** means exponent in SAS)
- `Trans = Counts**(1/9);` Takes the ninth root of Counts
- `Trans = Log(Counts);` Takes the natural logarithm (ln) of Counts
- `Trans = Log10(Counts);` Takes the base-10 logarithm of Counts
- `Trans = Sin(Counts);` Calculates the sine of Counts
- `Trans = Arsin(Counts);` Calculates the inverse sine (arcsine) of Counts

Etc…
Log Transformation

Example 5.1

In this experiment, the effect of vitamin supplements on weight gain is being investigated in three animal species (mice, chickens, and sheep). The experiment is designed as an RCBD with one replication (i.e. animal) per block*treatment combination. The six treatment levels are MC (mouse control), MV (mouse + vitamin), CC (chicken control), CV (chicken + vitamin), SC (sheep control), and SV (sheep + vitamin). The response variable is the weight of the animal at the end of the experiment.

Data Vit;
   Do Trtmt = 'MC', 'MV', 'CC', 'CV', 'SC', 'SV';
      Do Block = 1 to 4;
         Input Weight @@;
         Output;
      End;
   End;
Cards;
   0.18  0.30  0.28  0.44
   0.32  0.40  0.42  0.46
   2.0   3.0   1.8   2.8
   2.5   3.3   2.5   3.3
   108.0 140.0 135.0 165.0
   127.0 153.0 148.0 176.0;
Proc GLM Data = Vit Order = Data;
   Class Block Trtmt;
   Model Weight = Block Trtmt;
   Output Out = VitPR p = Pred r = Res;
   Contrast 'Vitamin' Trtmt 1 -1 1 -1 1 -1; * Test vitamin effect;
Proc Univariate Normal Data = VitPR; * Test normality of residuals;
   Var Res;
Proc GLM Data = Vit; * Levene's test for Trtmt (one-way ANOVA);
   Class Trtmt;
   Model Weight = Trtmt;
   Means Trtmt / hovtest = Levene;
Proc GLM Data = VitPR; * Tukey nonadditivity test;
   Class Block Trtmt;
   Model Weight = Block Trtmt Pred*Pred;
Proc Plot vpercent = 70 hpercent = 100; * v- and h-% tell SAS the size;
   Plot Res*Pred;
Proc Gplot Data = VitPR; * Makes a res vs. pred plot in another window;
   Plot Res*Pred;
Run;
Quit;
Output

The ANOVA

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block</td>
<td>3</td>
<td>984.0000</td>
<td>328.0000</td>
<td>2.63</td>
<td>0.0881 NS</td>
</tr>
<tr>
<td>Trtmt</td>
<td>5</td>
<td>108713.6800</td>
<td>21742.7360</td>
<td>174.43</td>
<td>&lt;.0001 ***</td>
</tr>
</tbody>
</table>
Contrast                    DF    Contrast SS    Mean Square    F Value    Pr > F
Vitamin                    1     142.1066667  142.1066667    1.14     0.3025  NS

Test for normality of residuals

Test                  --Statistic---    -----p Value------
Shapiro-Wilk          W     0.953596    Pr < W      0.3236  NS

Test for homogeneity of variance among treatments

Levene's Test for Homogeneity of Weight Variance
ANOVA of Squared Deviations from Group Means

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trtmt</td>
<td>5</td>
<td>699888</td>
<td>139978</td>
<td>2.51</td>
<td>0.0686  NS</td>
</tr>
<tr>
<td>Error</td>
<td>18</td>
<td>1005322</td>
<td>55851.2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Levene's Test is NS, but one can clearly see that it is borderline. The res vs. pred plot will illustrate this.

Test for nonadditivity

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block</td>
<td>3</td>
<td>984.0000</td>
<td>328.0000</td>
<td>98.15</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Trtmt</td>
<td>5</td>
<td>108713.6800</td>
<td>21742.7360</td>
<td>6506.42</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Tukey</td>
<td>1</td>
<td>1822.9405</td>
<td>1822.9405</td>
<td>545.51</td>
<td>&lt;.0001 ***</td>
</tr>
</tbody>
</table>

DANGER DANGER WILL ROBINSON!!!
SIGNIFICANT NON-ADDITIVE EFFECT! MUST TRANSFORM DATA!

Status: We violated our assumption of additivity, and Levene's Test for Treatment is almost significant. What to do? First thing's first: Read your tea leaves...
It's smiling at you.

And take a look at the means, standard deviations, and variances:

<table>
<thead>
<tr>
<th>Trtmt</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC</td>
<td>0.3000000</td>
<td>0.1070825</td>
<td>0.0114667</td>
</tr>
<tr>
<td>MV</td>
<td>0.4000000</td>
<td>0.0588784</td>
<td>0.0034667</td>
</tr>
<tr>
<td>CC</td>
<td>2.4000000</td>
<td>0.5887841</td>
<td>0.3466667</td>
</tr>
<tr>
<td>CV</td>
<td>2.9000000</td>
<td>0.4618802</td>
<td>0.2133333</td>
</tr>
<tr>
<td>SC</td>
<td>137.0000000</td>
<td>23.3666429</td>
<td>546.0000000</td>
</tr>
<tr>
<td>SV</td>
<td>151.0000000</td>
<td>20.1163284</td>
<td>404.6666667</td>
</tr>
</tbody>
</table>

Between mice and sheep, the mean increases by a factor of about 400, the standard deviation increases by a factor of about 270, and the variance increases by a factor of about 73,000!

The situation we face is this:

1. Significant Tukey Test for Nonadditivity
2. The standard deviation scales with the mean
3. The Res vs. Pred plot is smiling tauntingly at you

The best transformation under these conditions is a LOG transformation.

Example 5.2

Data Vit;
  Do Trtmt = 'MC', 'MV', 'CC', 'CV', 'SC', 'SV';
  Do Block = 1 to 4;
    Input BadWeight @@;
    Weight = Log10(BadWeight); * The ole ID switcheroo;
    Output;
  End;
End;
Cards;
...

Output

The ANOVA of the transformed data

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block</td>
<td>3</td>
<td>0.12049601</td>
<td>0.04016534</td>
<td>13.04</td>
<td>0.0002 ***</td>
</tr>
<tr>
<td>Trtmt</td>
<td>5</td>
<td>28.63231572</td>
<td>5.72646314</td>
<td>1859.57</td>
<td>&lt;.0001 ***</td>
</tr>
<tr>
<td>Contrast</td>
<td></td>
<td>Contrast SS</td>
<td>Mean Square</td>
<td>F Value</td>
<td>Pr &gt; F</td>
</tr>
<tr>
<td>Vitamin</td>
<td>1</td>
<td>0.05036523</td>
<td>0.05036523</td>
<td>16.36</td>
<td>0.0011 ***</td>
</tr>
</tbody>
</table>
Test for normality of residuals of the transformed data

<table>
<thead>
<tr>
<th>Test</th>
<th>Statistic</th>
<th>p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk</td>
<td>W 0.965975</td>
<td>0.5694 NS</td>
</tr>
</tbody>
</table>

Test for homogeneity of variance among transformed treatments

Levene's Test for Homogeneity of Weight Variance
ANOVA of Squared Deviations from Group Means

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trtmt</td>
<td>5</td>
<td>0.000795</td>
<td>0.000159</td>
<td>1.78</td>
<td>0.1686 NS</td>
</tr>
<tr>
<td>Error</td>
<td>18</td>
<td>0.00161</td>
<td>0.000090</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Test for nonadditivity in the transformed data

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type I SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Block</td>
<td>3</td>
<td>0.12049601</td>
<td>0.04016534</td>
<td>13.68</td>
<td>0.0002</td>
</tr>
<tr>
<td>Trtmt</td>
<td>5</td>
<td>28.63231572</td>
<td>5.72646314</td>
<td>1950.93</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Tukey</td>
<td>1</td>
<td>0.00509824</td>
<td>0.00509824</td>
<td>1.74</td>
<td>0.2087 NS</td>
</tr>
</tbody>
</table>

So all of our tests are good. Notice how much better the residuals look now:

At this point then, you may make conclusions about differences among treatments, etc. But be careful how you state your conclusions because you are making them based on transformed data. It is also customary to use the detransformed means in your final conclusions. "But aren't the detransformed means just the original means reclaimed?" NO:
When the mean of the logarithms is detransformed back to the original scale, what results is a geometric mean (not arithmetic mean) of the original data:

<table>
<thead>
<tr>
<th>Y</th>
<th>20</th>
<th>40</th>
<th>50</th>
<th>60</th>
<th>80</th>
<th>50</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(Y)</td>
<td>2.9957</td>
<td>3.6889</td>
<td>3.9120</td>
<td>4.0943</td>
<td>3.820</td>
<td>3.8146</td>
</tr>
</tbody>
</table>

The geometric mean of the original data \( G = (20*40*50*60*80)^{1/5} = 45.3586 \), exactly what you get if you detransform the log(Y) mean: \( 10^{3.8146} = 45.3586 \).

**Some final remarks about the Log transformation**

Data with negative values cannot be transformed this way. If there are zeros in the data, we are faced with the problem that Log(0) = - \( \infty \). To get around this, it is recommended that 1 be added to every data point before transforming. Logarithms to any base can be used, but log10 is most common. Before transforming, it is also legitimate to multiply all data points by a constant since this has no effect on subsequent analyses. This is a good idea if any data points are less than 1, for in this way you can avoid negative logarithms (Little and Hills).

**Power Transformation**

**Example 3**

This experiment is a generic CRD with six treatments and five replications per treatment.

```sas
Data Power;
  Do Trtmt = 'A', 'B', 'C', 'D', 'E', 'F';
    Do Rep = 1 to 5;
      Input Response @@;
      Output;
    End;
  End;
Cards;
  220 200 311 196 262
  96 213 142 154 151
  62 75 94 92 88
  378 323 228 177 265
  197 100 139 198 131
  77 80 123 118 101
;
Proc GLM Data = Power;
  Class Trtmt;
  Model Response = Trtmt;
  Means Trtmt / hovtest = Levene;
  Means Trtmt / Tukey;
  Output Out = PowerPR p = Pred r = Res;
Proc Univariate Normal Data = PowerPR;
  Var Res;
Proc Plot vpercent = 60;
  Plot Res*Pred = Trtmt;
  * Trtmt' labels each point according to treatment;
Proc Plot vpercent = 60;
```
Plot Res*Pred; * no '=' Trtmt' gives same plot but without treatment information;
Run;
Quit;

Note: There is no Tukey 1-df Test for Nonadditivity because this is a CRD.

Output

The ANOVA

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>5</td>
<td>143272.9667</td>
<td>28654.5933</td>
<td>13.44</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>24</td>
<td>51180.0000</td>
<td>2132.5000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>29</td>
<td>194452.9667</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type III SS</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trtmt</td>
<td>5</td>
<td>143272.9667</td>
<td>28654.5933</td>
<td>13.44</td>
<td>&lt;.0001  ***</td>
</tr>
</tbody>
</table>

Test for normality of residuals

<table>
<thead>
<tr>
<th>Test</th>
<th>--Statistic---</th>
<th>-----p Value------</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shapiro-Wilk</td>
<td>W 0.982662</td>
<td>Pr &lt; W 0.8910 NS</td>
</tr>
</tbody>
</table>

Test for homogeneity of variance among treatments

Levene's Test for Homogeneity of Response Variance
ANOVA of Squared Deviations from Group Means

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trtmt</td>
<td>5</td>
<td>75259223</td>
<td>15051845</td>
<td>2.82</td>
<td>0.0386  *</td>
</tr>
<tr>
<td>Error</td>
<td>24</td>
<td>1.2817E8</td>
<td>5340548</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

DANGER DANGER!!!
Significant Levene's Test! Must transform data!

The tea leaves
The significant Levene's Test is reflected in the Res*Pred plot above. The funnel shape of the data indicates that the magnitude of the residuals is increasing as the mean increases. This is verified by the table of means and standard deviations found below the Levene’s Test:

<table>
<thead>
<tr>
<th>Level of Trtmt</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>237.800000</td>
<td>48.5715966</td>
</tr>
<tr>
<td>B</td>
<td>5</td>
<td>151.200000</td>
<td>41.7097111</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td>82.200000</td>
<td>13.4981480</td>
</tr>
<tr>
<td>D</td>
<td>5</td>
<td>274.200000</td>
<td>78.7762655</td>
</tr>
<tr>
<td>E</td>
<td>5</td>
<td>153.000000</td>
<td>43.1566913</td>
</tr>
<tr>
<td>F</td>
<td>5</td>
<td>99.800000</td>
<td>21.1116082</td>
</tr>
</tbody>
</table>

In this situation, a power transformation will likely restore the data; but what is the appropriate power to use? There is a slick procedure for finding this information, and it involves performing a regression of the logarithms of the variances vs. the logarithms of the means of the original data. The code:

Example 4  
**Calculating the power for a power transformation [Lab5ex4.sas]**

```sas
Data Power2;
   Input Mean Stdev; * Treatment means and stddevs from original data;
   LogMean = Log10(Mean); * Calculate the log of treatment means;
   LogVar = Log10(Stdev*Stdev); * Calculate the log of treatment variances;
Cards;
   237.800000 48.5715966
   151.200000 41.7097111
   82.200000 13.4981480
   274.200000 78.7762655
   153.000000 43.1566913
   99.800000 21.1116082
;
Proc GLM; * Running the regression by Proc GLM, no Class statement;
   Model LogVar = LogMean;
Proc Reg; * Running the regression by Proc Reg (same results);
   Model LogVar = LogMean;
Run; Quit;
Output
```
Locate the slope of the regression. In this case, slope = 2.581433078. Now calculate the appropriate power of the transformation, where Power = 1 – (b/2). In this case,

\[
\text{Power} = 1 - \left(\frac{2.581433078}{2}\right) = -0.29
\]

To use this magic number, return to the original SAS code and make the following highlighted changes:

```sas
Data Power;
   Do Trtmt = 'A', 'B', 'C', 'D', 'E', 'F';
      Do Rep = 1 to 5;
         Input BadResponse @@;
         Response = BadResponse**(-0.29);
         Output;
      End;
   End;
Cards;
```

As before in the log transformation, what we have done is a little ID shuffle so that we do not have to chase our variable through the rest of the code. The results?

Output

Again, we have a significant ANOVA and a NS Shapiro-Wilk test. But our Levene's Test result has changed dramatically:

```
Levene's Test for Homogeneity of Response Variance
ANOVA of Squared Deviations from Group Means

Source    DF  Sum of Squares  Mean Square  F Value   Pr > F
Trtmt      5   1.683E-7    3.365E-8   0.51       0.7655 NS!
Error      24   1.582E-6    6.59E-8
```

And this result is confirmed by the Res*Pred plot for the transformed data, shown below. Notice that the strong funnel shape is now gone and the variances have lost their previous correlation to the means.
The suggested power transformation restored the homogeneity of variances and eliminated the obvious correlation between means and dispersion. Mean comparisons based on the transformed data are valid, but those based on the untransformed (i.e. original) data are not. This is because in the ANOVA of the original data, you used an average variance (MSE) that is not really representative of the different variances present across the different treatments.

To present a table of mean comparisons from this experiment, first perform the mean comparison analysis on the transformed data. The results:

<table>
<thead>
<tr>
<th>Tukey Grouping</th>
<th>Mean</th>
<th>N</th>
<th>Trtmt</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.27965</td>
<td>5</td>
<td>C</td>
</tr>
<tr>
<td>B</td>
<td>0.26500</td>
<td>5</td>
<td>F</td>
</tr>
<tr>
<td>B C</td>
<td>0.23609</td>
<td>5</td>
<td>B</td>
</tr>
<tr>
<td>B C</td>
<td>0.23543</td>
<td>5</td>
<td>E</td>
</tr>
<tr>
<td>D C</td>
<td>0.20580</td>
<td>5</td>
<td>A</td>
</tr>
<tr>
<td>D</td>
<td>0.19887</td>
<td>5</td>
<td>D</td>
</tr>
</tbody>
</table>

While the Tukey Groupings (i.e. significance groups) shown in this table are correct, it is customary to present the means in the original data scale. To do this, you should detransform the means of the transformed data, using the inverse operation of the original transformation:

\[ \text{detransformed mean} = (\text{original mean})^{-1/0.29} \]

<table>
<thead>
<tr>
<th>Tukey Grouping</th>
<th>Mean</th>
<th>N</th>
<th>Trtmt</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>262.2567</td>
<td>5</td>
<td>D</td>
</tr>
<tr>
<td>A B</td>
<td>233.0396</td>
<td>5</td>
<td>A</td>
</tr>
<tr>
<td>C B</td>
<td>146.5527</td>
<td>5</td>
<td>E</td>
</tr>
<tr>
<td>C B</td>
<td>145.1448</td>
<td>5</td>
<td>B</td>
</tr>
<tr>
<td>C D</td>
<td>97.45572</td>
<td>5</td>
<td>F</td>
</tr>
<tr>
<td>D</td>
<td>80.95147</td>
<td>5</td>
<td>C</td>
</tr>
</tbody>
</table>

Notice how it was necessary to flip the sequence of the treatments and shuffle the letters of the significance groupings in order to keep the means listed from largest to smallest.

**THE TAKE-HOME MESSAGE**

USE THE DATA THAT BETTER FIT THE ANOVA ASSUMPTIONS, NOT THE DATA THAT BETTER FIT YOUR ASSUMPTIONS ABOUT NATURE.
The Pitfalls of Interpreting Interactions in Transformed Data

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>A</th>
<th>B</th>
<th>AB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>20</td>
<td>30</td>
<td>35</td>
<td>45</td>
</tr>
<tr>
<td>Y^2</td>
<td>400</td>
<td>900</td>
<td>1225</td>
<td>2025</td>
</tr>
</tbody>
</table>

Our transformation $y^2$