Programming the NLIN Procedure Using the Negative Binomial Yield Model for ASIC Semiconductors

Jon A. Patrick
IBM Microelectronics Division, Essex Junction, VT 05452

Abstract
This presentation will show how easily the negative binomial yield model is programmed using the SAS® NLIN procedure, and describe the characteristics of the output parameters: alpha - the cluster parameter, lambda - the average fault density, and \( y_0 \) - the gross yield. The predicted output will be overlaid with the actual data and plotted with SAS/GRAF®. The benefits to this method of characterization are:

- Maverick product is easily identified.
- Yield projections on new, larger size die can be easily determined.

Introduction
The SAS NLIN (NonLinear regression) procedure is a very useful tool for analyzing wafer final test yield of application specific integrated circuit (ASIC) semiconductors. The IBM Microelectronics Division fabricators in Essex Junction, Vermont, manufacture a wide range of OEM and internal ASICs with different chip or die areas. The overall area of a die is mostly a function of how many I/Os are necessary to perform the ASIC's particular function and is not always the most discrete variable to model. A finer measurement would be the number of logic circuits or transistors per die, but this is hard to obtain for each ASIC. In general, the industry assumes fully populated ASICs, therefore, this analysis models yield based on die area in sq mm.

Negative Binomial Yield Model
The negative binomial model takes the form of

\[
Yield = y_0 \times \left[1 + \frac{\lambda}{\alpha} \right]^{-\alpha}
\]

The model is used throughout IBM to project semiconductor yields and has been written about extensively by Dr. C. H. Stapper. Before looking at the results of the model, the parameters must be briefly described. \( y_0 \) is the y intercept or the maximum yield your fab is producing at the time your data are analyzed. It's effect is to lower the predicted yield, and is referred to as a gross yield parameter. One way to determine \( y_0 \), on a mature process line, is by plotting the mean die yield across 15-20 batches of a particular ASIC as a function of the distance from the edge of the wafer (Figure 1). Each point on the plot is a single die averaged across all wafers in the sample and \( y_0 \) is the ratio of the overall yield across the entire wafer.
sample divided by the higher yield of die greater than 6-7mm from the wafer's edge. This edge loss is usually related to extra handling damage, increased thickness of spun-on materials such as resist or polyimide, or loss of parametric control across 100% of the wafer. Typical values for \( \gamma_0 \) on a mature process line range between .94 -to- 1.0, which are good starting values for \( \gamma_0 \) when beginning the NLIN procedure.

![Die Yield Chart](image)

**Figure 1.** Plot of die yield as a function of distance to edge of wafer (mm)

\[
\gamma_0 = .97  \\
\text{Die Size} = 88.4 \text{ sq mm}
\]

\( \lambda \), the average fault density per die, is expressed as

\[
\lambda = N_a \left( \frac{mm^2}{die} \right) \times \lambda_e \left( \frac{\text{faults}}{mm^2} \right)
\]

where \( N_a \) is die area in sq mm and \( \lambda_e \) is the average number of faults per sq mm. The NLIN procedure is coded using the expression \( N_a \times \lambda_e \), so the units of the parameter estimate, \( \lambda_e \), are in faults per sq mm. If you have available the number of circuits per die this can be substituted for \( N_a \), which changes the units of lambda to faults per circuit. Good starting values for lambda range between .001 -to-.010 faults per sq mm.

The term "alpha" is considered a defect cluster parameter or a measure of how defects tend to concentrate in close proximity of one another. The higher the alpha the more random the defect distribution across a wafer and the lower the wafer yield. Low values of alpha indicate high clustering and higher yield since multiple defects will destroy the same die. Good starting values for alpha range between .2 -to- 2.5, with the typical value for alpha being 2.0. When modeling yields in the high 80's and low 90's, expect a lower value for alpha.

When projecting yields on larger area die be slightly conservative and restrict alpha's range with the BOUNDS statement to a value close to 2.0. This effect will lower the predicted yield on the large area die your fabricator plans to build.

### Getting Started

The input data to the model are yield and area, while the results or parameter estimates are \( \gamma_0 \), \( \lambda \), and alpha, which provide the best fit to the data. The SAS NLIN procedure has five different techniques for initiating the analysis. The DUD method is the easiest to program because no derivatives are necessary. The modified Gauss-Newton method requires derivatives for each of the variables, which are provided in the code following this text. As with any regression analysis, the first step is to look at the data. Critical data are individual batch yields and sample size within each batch. Concentrate on fixing problems that are systematic across all batches of a particular ASIC. Don't include yield results for a poorly processed batch if the mode of the yield distribution for that ASIC is significantly higher. Data should be screened to the latest six months of process history. When ready to average data across an ASIC, reflect the weight a particular batch has on the overall mean by the number of wafers within each batch or, simply, the total number of good die divided by the total number of die tested for each ASIC. By using PROC MEANS; BY ASIC; the result gives equal weight per batch, regardless if one batch has 48 wafers and the other only 8 wafers. Another way to give equal weight per batch is using the FREQ X statement within PROC MEANS, were X equals the number of wafers per batch.
Output

The most important part of the NLIN output is whether or not convergence was met. If NLIN fails to converge, the asymptotic standard error is set to zero and the upper and lower 95% confidence intervals are equal, resulting in unreliable parameter estimates. If convergence is not met, expand the starting values in the PARMS statement; however, avoid using too small a increment in the grid of starting values because the job may exceed the time allotted trying to converge. The BOUNDS statement should also be used to prevent the possibility of alpha becoming negative. The output dataset contains yhat— the predicted yield, yresid— the residual yield, and the parameter estimates, \( \beta_0 \), \( \lambda \), and alpha. Always output the residuals and review them, as one or two observations with high positive residuals may have been coded with the wrong die area. The parameter estimates are placed into macro variables so they can be printed in the title of the graph. This helps keep track of parameter estimates after changes have been made. ASIC's with high negative residuals should be the focus of your analysis. This is the case in Figure 2, where the maverick ASIC clearly stands out from other ASICs of the same die size. Figure 3 shows the same data, only the mean yield across all ASICs of the same die size are plotted. This results in more accurate parameter estimates because the asymptotic 95% confidence intervals are tighter and the error associated with each lower.

Conclusion

The SAS NLIN procedure is a useful tool in semiconductor yield analysis. Results can be plotted and maverick ASICs can be easily identified, which helps define and focus work on low yield product. Dissecting the test yield into individual components, especially, the total static leakage current (Idd) is the first step. Previous analysis found that a particular maverick ASIC had 20-25% of the die failing Idd (and only Idd) while passing all other logic pattern tests. The background level of these fails was 4-9%, depending on die size. This failure analysis information was used to start a major process change at polysilicon which reduced our across-chip line variation (ACLV) by 3x, thereby virtually eliminating all but 1% of Idd-only loss.
References


- The author may be contacted on Internet: JON_PATRICK@VNET.IBM.COM

Appendix

Sample format of incoming data. None of the data are actual.

<table>
<thead>
<tr>
<th>OBS</th>
<th>ASIC</th>
<th>YIELD</th>
<th>AREA</th>
<th>Batches</th>
<th>Wafers</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0003849725</td>
<td>0.99</td>
<td>58.4</td>
<td>4</td>
<td>32</td>
</tr>
<tr>
<td>4</td>
<td>0007396256</td>
<td>0.87</td>
<td>66.2</td>
<td>2</td>
<td>23</td>
</tr>
<tr>
<td>11</td>
<td>0007639529</td>
<td>0.98</td>
<td>60.3</td>
<td>2</td>
<td>35</td>
</tr>
<tr>
<td>12</td>
<td>0001650955</td>
<td>0.71</td>
<td>98.4</td>
<td>8</td>
<td>323</td>
</tr>
<tr>
<td>13</td>
<td>0001875955</td>
<td>0.72</td>
<td>98.4</td>
<td>4</td>
<td>97</td>
</tr>
<tr>
<td>20</td>
<td>0006825661</td>
<td>0.72</td>
<td>98.4</td>
<td>4</td>
<td>172</td>
</tr>
<tr>
<td>49</td>
<td>0000237996</td>
<td>0.89</td>
<td>143.3</td>
<td>5</td>
<td>324</td>
</tr>
<tr>
<td>13</td>
<td>0006602844</td>
<td>0.94</td>
<td>143.3</td>
<td>8</td>
<td>736</td>
</tr>
<tr>
<td>55</td>
<td>0001380957</td>
<td>0.88</td>
<td>181.3</td>
<td>9</td>
<td>123</td>
</tr>
</tbody>
</table>

The SAS code to apply the complete model is:

```sas
/* coding the NLIN procedure using the GAUSS method */
PROC NLIN DATA=SUMMARY METHOD=GAUSS;
   PARMS ALPHA= 20 TO 2.5 BY .5;
   YD= .99 TO 1.0 BY .5;
   L = .5 TO .99 BY .02;
   BOUNDS ALPHA= .5, ALPHA= 3.5;
   MODEL YIELD= 1/((1-C70 + (AREA*/ALPHA))**ALPHA);   
   DEFL= 1;
   DFLX= -(AREA*/ALPHA)**(1/(1-(ALPHA)));
   DFLX2= (1/ALPHA)**(3/2)*(1/ALPHA));
   OUTPUT OUT= PREDICTED PREDICTED PREDICTED=ALPHA YD L;   */
```

---

1175