Minimum Percent Error-Zero Percent Bias Regression for Wine Economists

Neal D. Hulkower, Ph.D.
McMinnville, OR
nhulkower@yahoo.com
Outline

- Motivation
- Terminology
- Limitations of Ordinary Least Squares (OLS)
- Limitations of Log-Log OLS (LLOLS)
- Minimum Percent Error-Zero Percent Bias (MPE-ZPB)
- Calculation
- Measures of Significance
- Example
- Summary
- References
Motivation

• Much of the work of wine economists is in the phenomenological stage of development without a causal basis
• New relationships are derived from empirical data using OLS and LLOLS regression methods
• These centuries old methods have severe limitations and problems
• MPE-ZPB is an alternative that can overcome many of these issues
Terminology

• **Data set**: $n$ points each of which relates $k$ independent variables, $x_{ji}$, $i = 1$ to $k$, to a dependent variable, $y_j$, $j = 1$ to $n$

• **Fit parameter** is a coefficient, exponent, or other number that is not an independent variable that defines a function which is determined by a regression method
  - The fit parameters of $y = f(x_1, x_2, x_3, x_4) = a + bx_1^c x_2^d x_3^e g^{x_4}$ are $a, b, c, d, e, g$
Limitations of Ordinary Least Squares (OLS)

• Can only fit a data set to a linear equation, \( y = f(x) = ax + b \)
• Only accommodates an additive error model, \( y = ax + b + \varepsilon \) where \( \varepsilon \) is the difference between the actual and estimated value of \( y \)
Limitations of Log-Log OLS (LLOLS) (1 of 2)

- Can only fit a data set to a power law, one of the form \( y = f(x) = ax^b \)
  - Must go through origin
  - Converts to OLS by taking log of both sides

- Since OLS uses additive error model, the error, \( E = \log y - (\log a + b \log x) \) has log of whatever units \( y \) has
Limitations of Log-Log OLS (LLOLS) (2 of 2)

• “Error of Estimat[es]…is not minimized
• “LLOLS [estimates] are biased (usually low, but sometimes high)
• “When bias is ‘corrected’ to zero, the [estimate’s] Standard Error and $R^2$ must be recalculated
• “Although the LLOLS error model is multiplicative, the reported standard error has no meaning” (Book, 2012)
Minimum Percent Error-Zero Percent Bias (MPE-ZPB) (1 of 3)

- Can accommodate additive or multiplicative error and any functional form

- Goal: find a function $y = f(x)\varepsilon$ with a multiplicative error, $\varepsilon = \frac{y}{f(x)}$, that fits a data set $(x_i, y_i), i = 1$ to $n$, so that the
  - Fit parameters of $f(x)$ minimize
  \[ \sum (\varepsilon_i - 1)^2 = \sum \left[ \frac{y_i - f(x_i)}{f(x_i)} \right]^2, \text{ where } \varepsilon_i = \frac{y_i}{f(x_i)} \]
  - Bias $= \frac{1}{n} \sum \left[ \frac{f(x_i) - y_i}{f(x_i)} \right]$ is constrained to be zero
Minimum Percent Error-Zero Percent Bias (MPE-ZPB) (2 of 3)

• Quality Metrics
  – Percentage Standard Error =
  \[
  \sqrt{\frac{1}{n-k} \sum_{i=1}^{n} \left( \frac{y_i - f(x_i)}{f(x_i)} \right)^2} \times 100\%
  \]
  where \( k \) is the number of fit parameters
  • Want this minimized
  – Percentage Bias =
  \[
  \frac{1}{n} \sum \left( \frac{f(x_i) - y_i}{f(x_i)} \right) \times 100\%
  \]
  • Want this constrained to 0
Minimum Percent Error-Zero Percent Bias (MPE-ZPB) (3 of 3)

• Quality Metrics (Continued)
  – Pearson’s Correlation Squared, the squared correlation between the estimated values, \( f(x_i) \), and the actual values, \( y_i \)

\[
\text{Pearson's } r^2 = \left[ \frac{n \sum_{i=1}^{n} y_i f(x_i) - \sum_{i=1}^{n} y_i \sum_{i=1}^{n} f(x_i)}{\sqrt{n \sum_{i=1}^{n} y_i^2 - \left( \sum_{i=1}^{n} y_i \right)^2}} \sqrt{\frac{n \sum_{i=1}^{n} f(x_i)^2 - \left( \sum_{i=1}^{n} f(x_i) \right)^2}{n \sum_{i=1}^{n} f(x_i)^2}} \right]^2
\]

• Want this as close to 1 as possible
(References: Book and Lao, 1998; Book, 2012)
Calculation

• MPE-ZPB requires the solution of a constrained nonlinear optimization problem
• While Excel “Solver” can do this, one is never guaranteed to find the global minimum without manual intervention
• One must also be cautious to avoid division by zero
Measures of Significance (1 of 3)  
Reference: Anderson, 2009

• Idea: need a test for the significance of the fit parameters since the classic ones don’t apply

• By “nullifying” a fit parameter, we can assess the significance of each

• For multiplicative error, two measures of significance developed, $\text{SIG}_{\text{Mean}}$ and $\text{SIG}_{\text{SPE}}$ (SPE is standard percent error)
  – $\text{SIG}_{\text{SE}}$ defined for additive error instead of $\text{SIG}_{\text{SPE}}$ (SE is standard error)
Measures of Significance (2 of 3)
Reference: Anderson, 2009

“Pseudo-mean” of $f(X)$ = \( f_{\bar{Y}}(X) = f(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_N) \)

\[
SIG_{\text{Mean}} = \frac{f_{\bar{Y}}(X)_{\text{Reduced}} - f_{\bar{Y}}(X)_{\text{Full}}}{f_{\bar{Y}}(X)_{\text{Full}}}
\]

\[
SIG_{\text{SPE}} = \frac{SPE_{\text{Reduced}} - SPE_{\text{Full}}}{SPE_{\text{Full}}}
\]

- The “Pseudo-mean” of $f(X)$ is the same as the mean of $f(X)$ only if $f(X)$ is linear
- “Full” means includes all the independent variables
- “Reduced” means one of the fit parameters is “nullified”
Measures of Significance (3 of 3)
Reference: Anderson, 2009

- Can combine: \( \text{SIG}_{\text{Total}} = \text{SIG}_{\text{SPE}} + |\text{SIG}_{\text{Mean}}| \)
- Anderson suggests as default values:
  - \( \text{SIG}_{\text{SPE}}, \text{SIG}_{\text{Mean}} < 5\% \) indicate fit parameter is insignificant
  - \( \text{SIG}_{\text{Total}} < 10\% \) indicates fit parameter is insignificant
Example (1 of 4)
Reference: Anderson, 2009

Use MPE-ZPB to fit:

\[ Y = f(X) = a + bX_1^c X_2^d X_3^e X_4^f \]

to data with multiplicative error

number of data points, \( n = 28 \)
number of fit parameters = 6
degrees of freedom = 22

<table>
<thead>
<tr>
<th>( Y )</th>
<th>( X_1 )</th>
<th>( X_2 )</th>
<th>( X_3 )</th>
<th>( X_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>124</td>
<td>209</td>
<td>3</td>
<td>5.2</td>
<td>0</td>
</tr>
<tr>
<td>126</td>
<td>106</td>
<td>4</td>
<td>16.7</td>
<td>0</td>
</tr>
<tr>
<td>105</td>
<td>35</td>
<td>10</td>
<td>9.2</td>
<td>0</td>
</tr>
<tr>
<td>337</td>
<td>255</td>
<td>6</td>
<td>9.6</td>
<td>0</td>
</tr>
<tr>
<td>281</td>
<td>206</td>
<td>7</td>
<td>2.9</td>
<td>0</td>
</tr>
<tr>
<td>131</td>
<td>109</td>
<td>12</td>
<td>4.9</td>
<td>0</td>
</tr>
<tr>
<td>155</td>
<td>120</td>
<td>2</td>
<td>12.7</td>
<td>0</td>
</tr>
<tr>
<td>46</td>
<td>27</td>
<td>5</td>
<td>3.4</td>
<td>0</td>
</tr>
<tr>
<td>224</td>
<td>286</td>
<td>1</td>
<td>9.8</td>
<td>0</td>
</tr>
<tr>
<td>233</td>
<td>257</td>
<td>4</td>
<td>36.8</td>
<td>0</td>
</tr>
<tr>
<td>357</td>
<td>134</td>
<td>4</td>
<td>21.1</td>
<td>1</td>
</tr>
<tr>
<td>240</td>
<td>149</td>
<td>5</td>
<td>3.3</td>
<td>1</td>
</tr>
<tr>
<td>97</td>
<td>10</td>
<td>44</td>
<td>5.0</td>
<td>1</td>
</tr>
<tr>
<td>134</td>
<td>32</td>
<td>1</td>
<td>25.9</td>
<td>1</td>
</tr>
<tr>
<td>149</td>
<td>46</td>
<td>18</td>
<td>5.5</td>
<td>1</td>
</tr>
<tr>
<td>422</td>
<td>253</td>
<td>4</td>
<td>18.3</td>
<td>1</td>
</tr>
<tr>
<td>262</td>
<td>81</td>
<td>20</td>
<td>13.7</td>
<td>1</td>
</tr>
<tr>
<td>126</td>
<td>37</td>
<td>18</td>
<td>4.7</td>
<td>1</td>
</tr>
<tr>
<td>152</td>
<td>28</td>
<td>18</td>
<td>6.8</td>
<td>1</td>
</tr>
<tr>
<td>331</td>
<td>183</td>
<td>2</td>
<td>5.6</td>
<td>1</td>
</tr>
<tr>
<td>204</td>
<td>24</td>
<td>37</td>
<td>2.7</td>
<td>1</td>
</tr>
<tr>
<td>59</td>
<td>6</td>
<td>20</td>
<td>6.5</td>
<td>1</td>
</tr>
<tr>
<td>134</td>
<td>46</td>
<td>4</td>
<td>1.1</td>
<td>1</td>
</tr>
<tr>
<td>232</td>
<td>108</td>
<td>3</td>
<td>20.8</td>
<td>1</td>
</tr>
<tr>
<td>48</td>
<td>5</td>
<td>28</td>
<td>3.2</td>
<td>1</td>
</tr>
<tr>
<td>281</td>
<td>80</td>
<td>28</td>
<td>11.7</td>
<td>1</td>
</tr>
<tr>
<td>272</td>
<td>149</td>
<td>9</td>
<td>1.2</td>
<td>1</td>
</tr>
<tr>
<td>142</td>
<td>39</td>
<td>8</td>
<td>8.8</td>
<td>1</td>
</tr>
</tbody>
</table>

Average: 108 12 9.9 0.6
Example (2 of 4)
Reference: Anderson, 2009

• Using Excel Solver with initial guesses a = c = d= e =0 and b = f = 1,

\[ Y = -79.65 + 31.35(X_1)^{0.3664}(X_2)^{0.1094}(X_3)^{0.0576}(1.44)^{X_4} \]

SPE = 21.4%
Pearson's r² = 84.6%
Bias = 0%

• Using average values of the \( X_i \), the pseudo-mean is

\[ f_{\bar{Y}}(X)_{\text{Full}} = -79.65 + 31.35(108)^{0.3664}(12)^{0.1094}(9.9)^{0.0576}(1.44)^{0.6} \]

\[ = 245.22 \]
Example (3 of 4)
Reference: Anderson, 2009

- To test the significance of $c$, the fit parameter associated with $X_1$, “nullify” it by setting it to 0 and re-optimize to get:

$$f_Y(X)_{\text{Reduced}} = 151.96 + 9.03(\bar{X}_2)^{-0.3739}(\bar{X}_3)^{0.5594}(3.60)^{\bar{X}_4}$$

- Substituting the average values:

$$f_Y(X)_{\text{Reduced}} = 151.96 + 9.03(12)^{-0.3739}(9.9)^{0.5594}(3.60)^{0.6} = 179.69$$

- So then

$$SIG_{\text{Mean}} = \frac{179.69 - 245.22}{245.22} = -0.267 = -26.7\%$$

which implies that $c$ is significant.
Example (4 of 4)
Reference: Anderson, 2009

• The $\text{SPE}_{\text{Reduced}} = 52.7\%$ so that

$$
SIG_{SPE} = \frac{52.7\% - 21.4\%}{21.4\%} = 1.463 = 146.3\%
$$

which reinforces the fact that $c$ is significant
## Summary

<table>
<thead>
<tr>
<th></th>
<th>OLS</th>
<th>LLOLS</th>
<th>MPE-ZPB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Functional form</strong></td>
<td>( y = ax + b ), also multivariate</td>
<td>( y = ax^b ), also multivariate</td>
<td>Any and multivariate</td>
</tr>
<tr>
<td><strong>Error Model</strong></td>
<td>Additive only</td>
<td>Additive in the log of the error</td>
<td>Additive or multiplicative</td>
</tr>
<tr>
<td><strong>Computation</strong></td>
<td>Solve simultaneous linear equations</td>
<td>Take log of both sides and proceed as OLS</td>
<td>Constrained nonlinear optimization</td>
</tr>
<tr>
<td><strong>Measures of Quality</strong></td>
<td>Sample SE, Sample Bias, ( R^2 )</td>
<td>Same as OLS but in log space</td>
<td>SPE or SE, Bias and Pearson’s ( r^2 )</td>
</tr>
<tr>
<td><strong>Measures of Significance</strong></td>
<td>( p ), student ( t )</td>
<td>Same as OLS but in log space</td>
<td>( \text{SIG}<em>{SE} ) or ( \text{SIG}</em>{SPE} ), ( \text{SIG}<em>{\text{Mean}} ), ( \text{SIG}</em>{\text{Total}} )</td>
</tr>
</tbody>
</table>
References

