TECOLOTE RESEARCH, INC.

Bridging Engineering and Economics
Since 1973

The Impact of Using Log-Error CERs Outside the Data Range and PING Factor

Dr. Shu-Ping Hu

15 June 2005

■ Los Angeles ■ Washington, D.C. ■ Boston ■ Chantilly ■ Huntsville ■ Dayton ■ Santa Barbara
■ Albuquerque ■ Colorado Springs ■ Columbus ■ Ft. Meade ■ Ft. Monmouth ■ Montgomery ■ Ogden ■ Patuxent River ■ Pensacola ■ San Diego
■ Charleston ■ Cleveland ■ Denver ■ New Orleans ■ Oklahoma City ■ Silver Spring ■ Warner Robins AFB ■ Vandenberg AFB

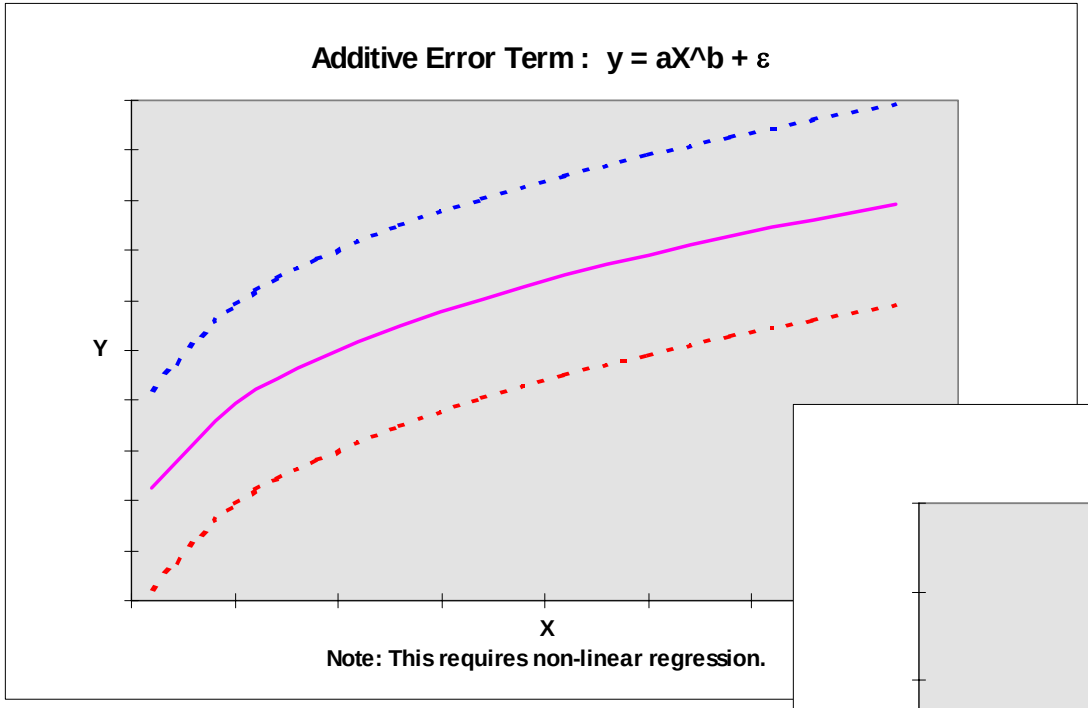
Objectives

- Define Minimum Unbiased Percent Error (MUPE) and Log-Linear CERs
- List pros and cons of MUPE vs. log-error CERs
- Review proper use of log-error CERs
- Explain why a correction factor is required for log-error CERs
- Describe and compare two popular log-error correction factors: Goldberg and PING

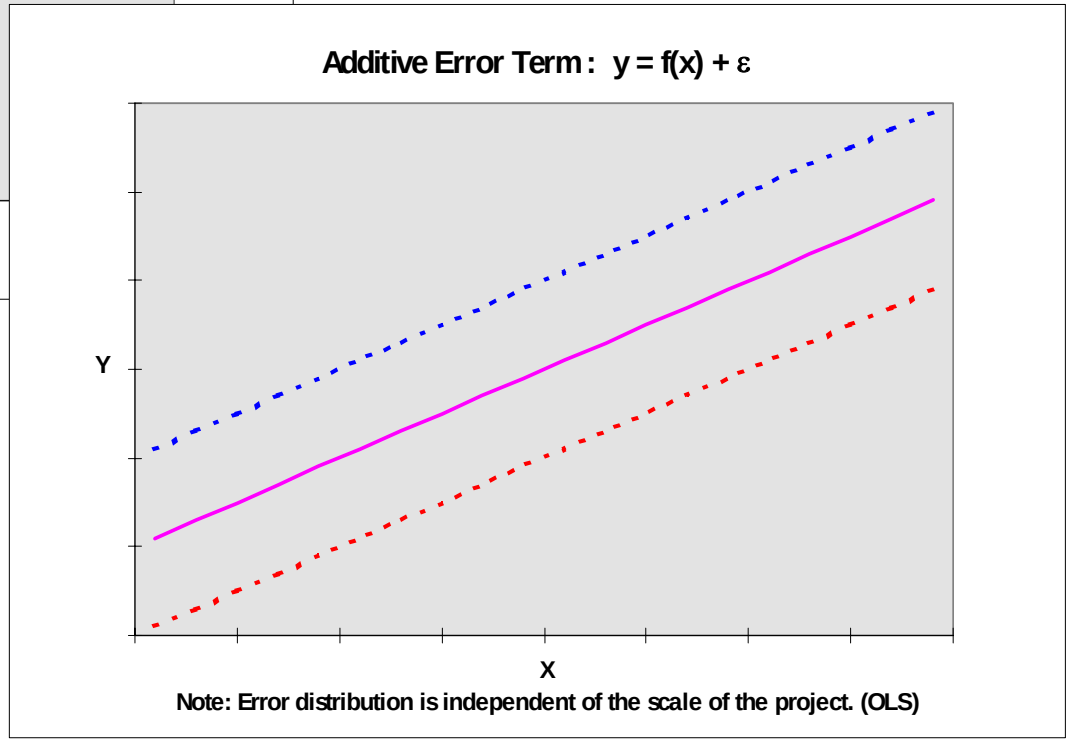
Outline

- **Objectives**
- **Introduction**
  - Error Term Assumption (Additive vs. Multiplicative)
  - Multiplicative Error Model (Log-Error vs. MUPE)
- **Properties of Log-Error CERs**
- **Common Concerns about Log-Error CERs**
- **Pros and Cons of MUPE and Log-Error CERs**
- **Derivations of Correction Factors (Goldberger/PING)**
- **Comparing Three Ways to Use Log-Linear Equations:**
  No Correction, Goldberger Factor Correction, PING Factor Correction
- **Conclusions**

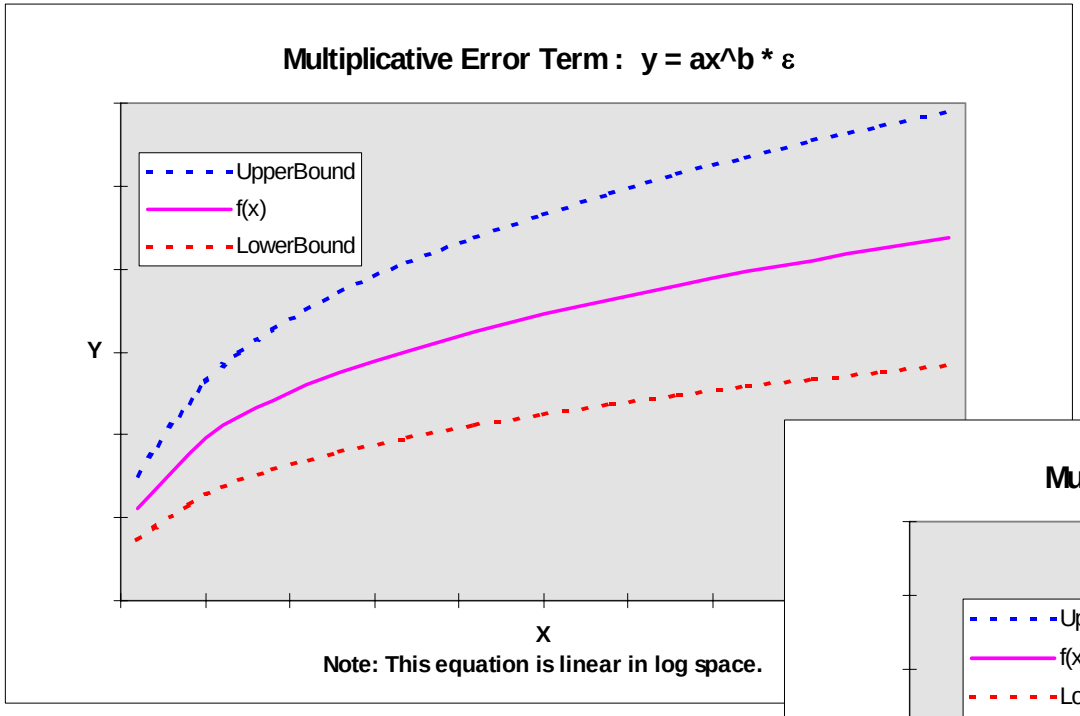
Additive Error Term

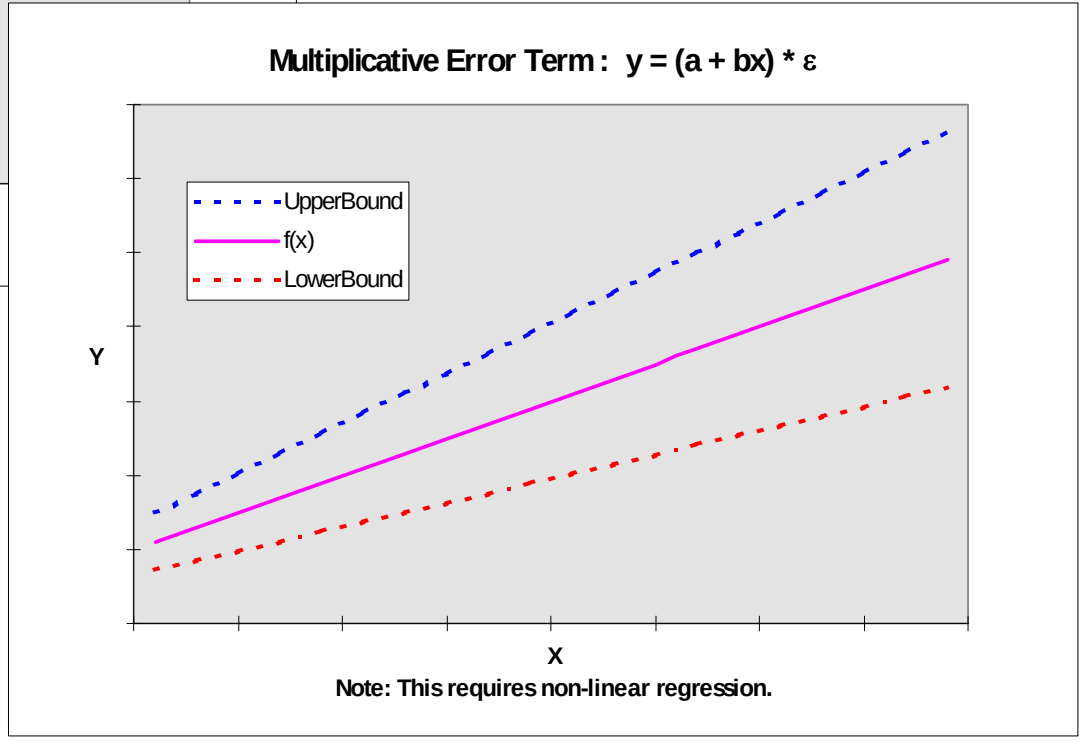


Cost variation is independent of the scale of the project



Multiplicative Error Term





Cost variation is proportional to the scale of the project

Error Term Assumptions – Background

	ADDITIVE ERROR	MULTIPLICATIVE ERROR (Log Error)
Distribution Assumption	$N(0, \sigma^2)$, independent	$LN(0, \sigma^2)$, independent
Typical Model Form	Linear form – $y = a + b\,x + \varepsilon$	Power form – $y = a\,x^{b}\,\varepsilon$
Legitimate Reasons	Absolute Errors	Proportional Errors
What should be cost errors?	Cost variation is independent of the scale of the project	Cost variation is proportional to the scale of the project
Statistical measures	Traditional statistical measures can be used	Traditional statistical measures can be used in the **log** space
Shortcomings	Not a good method if data not homogenous or data range over one order of magnitude	Need correction factor to adjust for the mean in unit space

- **Model form should not drive error term assumption**
- **Error term should not drive model form**

Multiplicative Error Model – MUPE vs. Log-Error

Definition of cost variation for $Y = f(x)*\varepsilon$

- **Log-Error: $\varepsilon \sim LN(0, \sigma^2)$ ⇨ Least squares in log space**
  - Error = Log (y_i) - Log $f(x_i)$
  - Minimize Σ_i (Log (y_i) - Log $f(x_i)$)^2

- **MUPE: $E(\varepsilon) = 1$, $V(\varepsilon) = \sigma^2$ ⇨ Least squares in unit space**
  - Error = $(Y-f(x)) / f(x)$
  - Minimize $\Sigma_i \{(y_i - f(x_i))/ f_{k-1}(x_i)\}^2$

    where k is the iteration number

Note:

$E((Y-f(x)) / f(x)) = 0$

$V((Y-f(x)) / f(x)) = \sigma^2$

Properties of Log-Error CERs

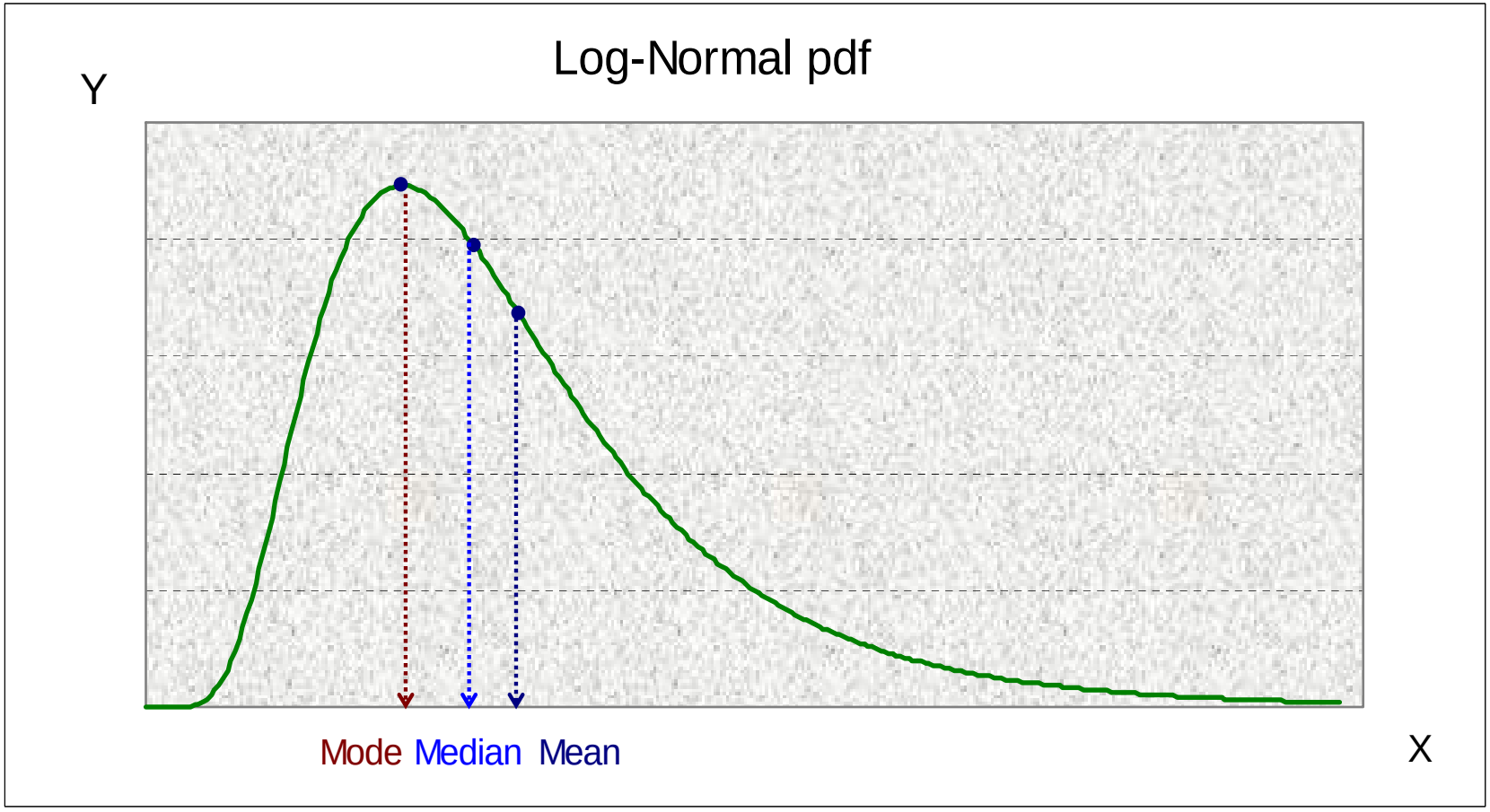
At a given x value x_o, if $Y = f(x)\varepsilon$, where $\varepsilon \sim LN(0, \sigma^2)$, then

- **$E(Y/x = x_o) = \mu_A = f(x_o)e^{\sigma^2/2}$**
- **Median $(Y/x = x_o) = M_A = f(x_o)$**
- **Mode $(Y/x = x_o) = f(x_o)e^{-\sigma^2}$**
- **Stdev $(Y/x = x_o) = f(x_o)(e^{\sigma^2} - 1)^{0.5}$**

$$\mu_A / M_A = \exp(\sigma^2 / 2)$$

The log-linear equation will be biased low if we do not apply correction factors

Log-Normal Distribution



Comparing Mean, Median, and Mode for Log-normal Distribution

Common Concerns about Log-Error CERs (1/2)

- **Errors not expressed in meaningful units ("log dollars")**
- **Minimizing $\sum(\log \varepsilon_i)^2$ is not the same as minimizing $\sum \varepsilon_i^2$**
- **Std Error of Estimate (SEE) in nonlinear case $\sqrt{\frac{1}{n-2}\sum(\log\varepsilon_i)^2}$ cannot be compared with SEE in linear case $\sqrt{\frac{1}{n-2}\sum \varepsilon_i^2}$ to see which functional form is the better estimator**
- **We must choose the power form CER to use the log error assumption**
  - If you choose nonlinear functional form, you must assume multiplicative-error model; if you choose linear functional form, you must assume additive error model
  - The major casualty is that you do not have access to the non-linear functional form $y = a + b\ x^c$
- **Correction factors needed because the resultant equation will be biased low in unit space**

Common Concerns about Log-Error CERs (2/2)

- **If $Y = f(x)*\varepsilon$ and $\varepsilon > 0$ ⇨ $(Y - f(x)) / f(x) = \varepsilon - 1 \cong \log(\varepsilon)$**
  - $\text{Log}(\varepsilon) \approx \varepsilon - 1 = (y - f(x))/f(x)$ by Taylor Series expansion
- **SEE in fit space should not be compared across different models**
  - For example, comparing the SEEs between an additive model and a multiplicative model is meaningless and we should not select a model based upon the comparison between two fit measures
- **Log-error assumption can be applied to any functional forms**
  - The choices between functional form and error term should be made independently of each other
- **MUPE method is suggested for modeling multiplicative error directly *in unit space* to avoid the use of correction factors**

Pros and Cons of MUPE CERs

- The MUPE CER has zero proportional error for all points in the database (no sample bias)
- The MUPE method requires no transformation and no correction factor adjustment
- Goodness-of-fit measures (or asymptotic goodness-of-fit measures) can be applied to judge the quality of the model under the normality assumption
- The MUPE CER produces consistent estimates of the parameters and the mean of the equation
- The estimated parameters using the MUPE method are also the maximum likelihood estimates (MLE) of the parameters (by Goldberg, 2001)
- It relies on nonlinear regression technique to derive a solution.
- MUPE CERs do not always converge, especially with learning curves

What Can We Do When MUPE Fails?

Resort to Log-Error CERs!

Deriving the Log-Linear CER Correction Factor

At a given x value x_o, if $Y = e^{\alpha}x^{\beta}\varepsilon$, where $\varepsilon \sim LN(0, \sigma^2)$, then

- **$\ln(Y / x = x_o) = \alpha + \beta \ln(x_o) + \ln(\varepsilon) \sim N(\alpha + \beta\ln(x_o), \sigma^2)$**
- **$\ln(\hat{Y} / x = x_o) = a + b \ln(x_o) \sim N(\alpha + \beta\ln(x_o), r_o\sigma^2)$**
- **$\hat{Y} / (x = x_o) = e^{a}x_o^{b} \sim LN(\alpha + \beta\ln(x_o), r_o\sigma^2)$**

$$r_o = \frac{1}{n} + \frac{(\ln(x_o) - \overline{\ln(x)})^2}{SSx}$$

- **$E(\hat{Y} / x = x_o) = e^{\alpha}x_o^{\beta}\, e^{r_o\sigma^2/2} = f(x_o)\, e^{r_o\sigma^2/2}$**
- **$E(Y / x = x_o) = e^{\alpha}x_o^{\beta}\, e^{\sigma^2/2}$**

$$\text{Net CF for Mean at } \mathbf{x_o}: \quad E(Y / \mathbf{x} = \mathbf{x_o}) \Big/ E(\hat{Y} / \mathbf{x} = \mathbf{x_o}) = e^{(1 - r_o)\frac{\sigma^2}{2}}$$

Goldberger's Method

- **Net Correction Factor (NFC):** $\exp((1-r_o)\frac{\sigma^2}{2})$
  - Downward bias (mean ~ median): $e^{\sigma^2/2}$
  - Upward bias (median): $e^{r_o\sigma^2/2}$

$$r_o = \frac{1}{n} + \frac{(\ln(x_o) - \overline{\ln(x)})^2}{\sum_i (\ln(x_i) - \overline{\ln(x)})^2}$$

However, our regression analysis provides SEE rather than the true standard deviation (σ) in log space. If we use SEE in NFC, we will **overestimate** the true value!

- **Goldberger's Factor:** $GF = g((1-r_o)\frac{s^2}{2}) \cong \exp((1-r_o)\frac{s^2}{2})$

$$g(t) = 1 + \frac{t}{1!} + \frac{(n-p)t^2}{2!(n-p+2)} + \frac{(n-p)^2 t^3}{3!(n-p+2)(n-p+4)} + \ldots$$

(p = total number of estimated coefficients)

- $E(g(as^2)) = e^{a\sigma^2}$ ⇨ **E(GF) = NCF**

Contrasting Goldberger & PING Factors

- **Goldberger's Factor:** $$GF = g((1-r_o)\frac{s^2}{2}) \cong \exp((1-r_o)\frac{s^2}{2})$$
  - A **variable** factor, which must be evaluated point by point $$r_o = \frac{1}{n} + \frac{(\ln(x_o) - \overline{\ln(x)})^2}{SSx}$$
  - It can become very cumbersome with multiple drivers in the CER
- **To avoid evaluating r_o point by point, we suggest using mean leverage value (p/n) to approximate r_o: $p/n \approx r_o$**
- **PING Factor:** $$PF = g((1-\frac{p}{n})\frac{s^2}{2}) \cong \exp((1-\frac{p}{n})\frac{s^2}{2})$$
  - A **constant** factor, which is used to adjust the level of the entire function
  - p = total number of estimated coefficients
  - n = sample size
- **Simple approximation sufficient for most cases:** $$PF \cong e^{(1-\frac{p}{n})\frac{s^2}{2}}$$

Comparing Goldberger and PING Factors

Consider the Ratio of GF to PF:

$$\exp((1-r_o)\frac{s^2}{2}) \Big/ \exp((1-\frac{p}{n})\frac{s^2}{2}) = \exp((\frac{p}{n}-r_o)\frac{s^2}{2})$$

- **The PING Factor is sufficiently close to the theoretical unbiased Goldberger's Factor within the data range.**

  (p/n = E(r_o) if x_o is from the data matrix)

- **What if the prediction is made outside the database?**
- **$r_o = \ln(x_o)(X'X)^{-1}\ln(x_o)^t$** (X =design matrix in log space)

$$r_o = \frac{1}{n} + \frac{(\ln(x_o) - \overline{\ln(x)})^2}{SSx}$$

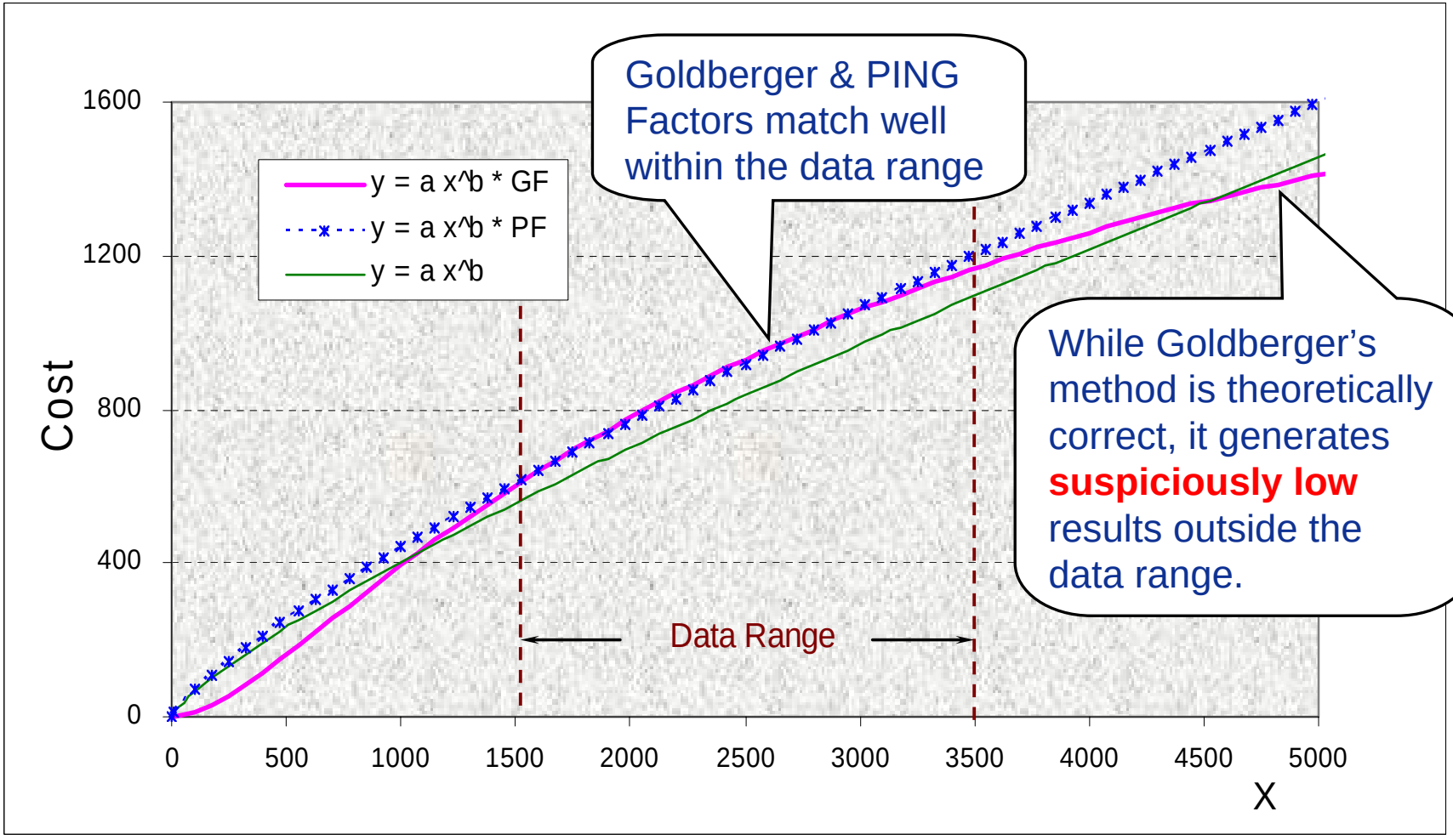
(in one-independent variable model)

- **The value of r_o can be larger than one if it is evaluated outside the data range. Goldberger's factor is *less than one* in this situation.**

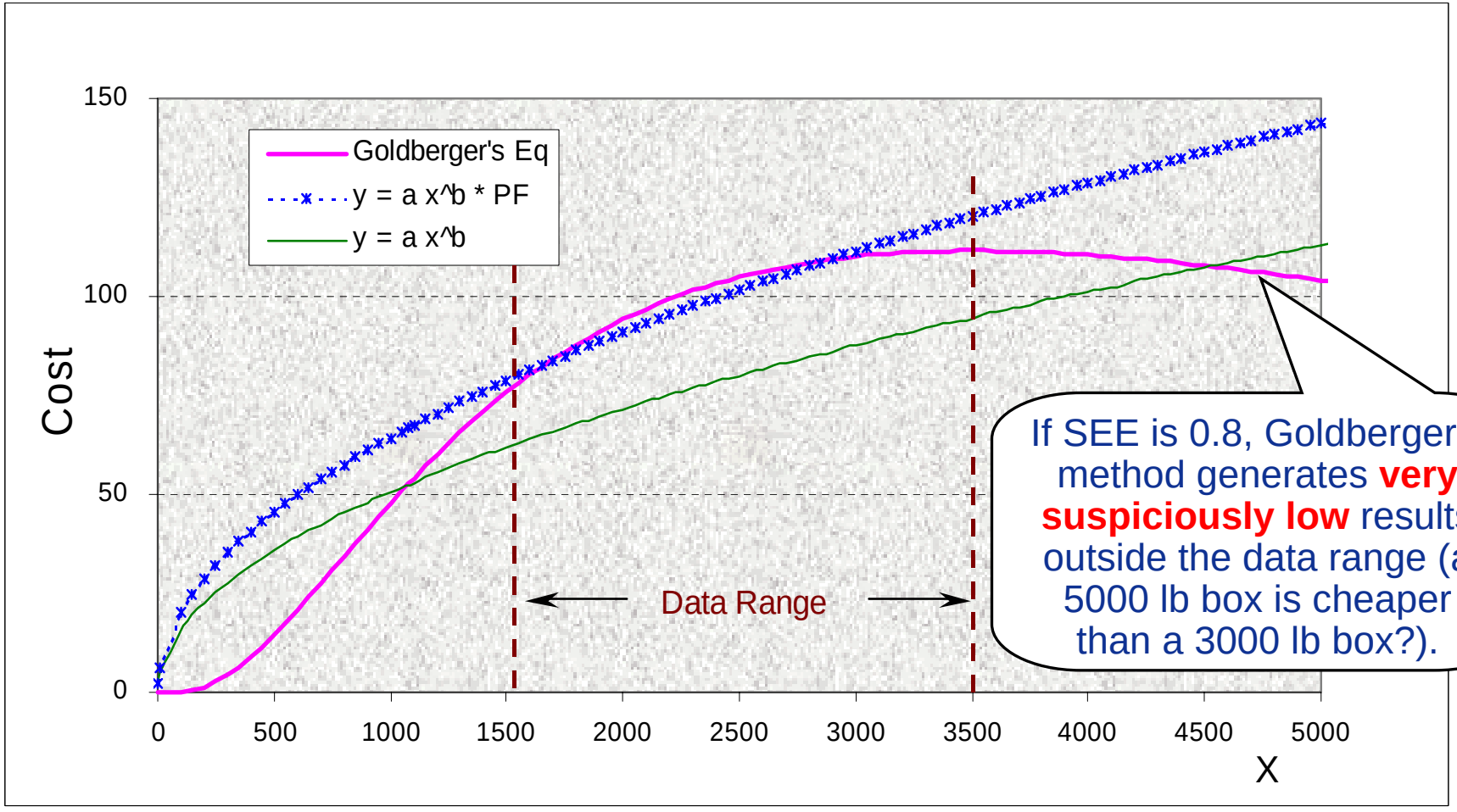
Compare Goldberger & PING When SEE = 0.5

$$GF = \exp((1 - \frac{1}{n} - \frac{(\ln(x_o) - \overline{\ln(x)})^2}{SSx})\frac{\sigma^2}{2})$$

$$PF = \exp((1 - \frac{2}{n})\frac{\sigma^2}{2})$$



Compare Goldberger & PING When SEE = 0.8



Comparing a Log-Linear CER with the **Goldberger** and **PING Factor** Equations Using 0.8 as SEE in Log Space

Conclusions – MUPE vs. Log-Error CERs?

- **Given the multiplicative model $y = f(x)\varepsilon$, the decision to develop a MUPE or a log-error CER should be based upon the error term assumption**
  - Choose MUPE if you believe the error term (ε) distribution has a mean of **one** and variance of σ^2
  - Choose log-error model if you believe ε follows a log-normal distribution with a mean of zero and variance σ^2 in log space, i.e., $\varepsilon \sim LN(0, \sigma^2)$
- **The choice of the CER functional form (linear, non-linear) should not drive the error term assumption (additive, multiplicative) and vice versa**

Conclusions – Log-Error CERs Have Their Advantages

- **If the hypothesized equation is log-linear, e.g., $y = ax^b\varepsilon$, then the regression can be done in log space linearly under the logarithmic transformation**
  - This process is an OLS in log space and all the goodness-of-fit measures can be evaluated in that space
  - The above advantage does not exist if the CER has a non-linear functional form in unit space, which cannot be linearized in log space
- **$SEE_L \cong$ CoV at a given x value**
  - The *standard error of estimate* in log space (SEE_L) can be regarded as the coefficient of variation (CoV) in unit space at a given x value. (proof follows from a Taylor series expansion)
- **Log-errors (log(y) – log(f(x)) can be viewed approximately as the MUPE percentage error**

Conclusions – The Bad News About Using Log-Error CERs

- **It involves a two-step process:**
  - Perform the curve fitting in log space
  - Transform the results back to unit space
- **We need to derive a correction factor (by Goldberger's method or the PING Factor) to adjust the unit space CER result to obtain an unbiased estimate**
- **We must be extremely cautious when the future prediction lies outside the data range**
  - Goldberger's Factor may generate counter intuitive results outside the data range, and this is more pronounced as SEE increases
  - The PING Factor may be more suitable than Goldberger's Factor in this situation

Conclusions – Correction Factors

- Goldberger's Factor and the PING Factor generally match each other very closely within the data range.
- Goldberger's Factor is a variable factor, which must be evaluated point by point and multiplied to the log-error CER result to obtain the theoretical mean in unit space.
- The PING Factor is a constant factor, which is used to adjust the level of the entire function. For most cases: $PF = e^{(1 - \frac{p}{n})\frac{s^2}{2}}$
- A common misuse of Goldberger's Factor is to adjust the intercept, and then use it for the entire equation. This practice generates an equation that underestimates the majority of the data points and should be avoided.
- Goldberger's Factor should be used with caution when predicting outside the data range because this factor may be considerably less than one. The PING Factor may be more suitable in this situation.

TECOLOTE RESEARCH, INC.

Bridging Engineering and Economics
Since 1973

Backup Slides

MPE and MUPE Methods

- **Two possible ways to perform the optimization for the weighted least squares using the predicted values**
  - MPE ⇨ high bias due to simultaneous minimization

$$Minimize \quad \sum_{i=1}^{n} \left(\frac{y_i - f(x_i)}{f(x_i)} \right)^2$$

  - MUPE ⇨ bias eliminated

$$Minimize \quad \sum_{i=1}^{n} \left(\frac{y_i - f(x_i)}{f_{k-1}(x_i)} \right)^2$$

where k is the iteration number