Regression for CERs with Multiplicative Errors

Modern Techniques

An unsophisticated forecaster uses statistics as a drunken man uses lamp-posts - for support rather than for illumination.
Andrew Lang

Acknowledgments

- ICEAA is indebted to TASC, Inc., for the development and maintenance of the Cost Estimating Body of Knowledge (CEBoK®)
- ICEAA is also indebted to Technomics, Inc., for the independent review and maintenance of CEBoK®
- ICEAA is also indebted to the following individuals who have made significant contributions to the development, review, and maintenance of CostPROF and CEBoK ®
- For what concern this specific module on modern techniques of regression for CERs with multiplicative errors, ICEAA is also indebted to:
  - Raymond P. Covert former Technical Director and Chief Practitioner, Cost and Schedule Analysis, MCR, LLC, now President Covarus LLC, and
  - Timothy P. Anderson, Director, NASA Program Assessments, The Aerospace Corporation

Acknowledgments

The subject matter of this training session was first organized in tutorial form under the title “Statistical Foundations of CER Development” for a March 2004 presentation to the management and support staff (including FFRDC and contractor) of the NRO Cost Group (now NRO CAIG), Chantilly VA. The presenters would like to thank the NRO CAIG, under the direction of Keith Robertson, for financing the original preparation of the tutorial. They would also like to thank his support-staff colleagues for the vigorous intellectual debate that traditionally precedes, follows, and improves such discussions of methodology at the NRO CAIG. Especially, they would like to thank the original author of this tutorial, the late Dr. Stephen A. Book. Steve’s leadership was the driving force behind many of the costing techniques we use on a daily basis today. Thank you Steve! We miss you!

Contents

Contents

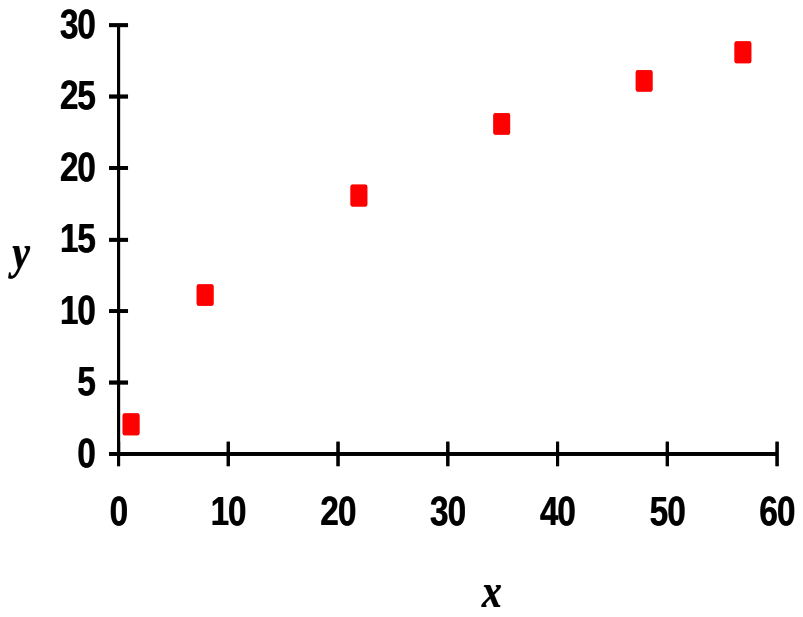
Mathematical Formulation of CERs

- y = Cost
  x = Technical Parameter (Cost Driver)
- Factor CER: $y = ax$
  Linear CER: $y = a + bx$
  "Nonlinear" CERs: $y = ax^b$
  $y = ab^x$
  $y = a + bx^c$
- a, b, c are Constant Coefficients Derived from Historical Data
- This Tutorial Will Discuss the Case of Only One Cost Driver per CER – For Multiple Cost Drivers, the Concepts are the Same, but the Statistics are More Complicated

Ideal Historical Data (after Normalization)

COST	TECHNICAL PARAMETER
y_1	x_1
y_2	x_2
y_3	x_3
.	.
.	.
.	.
y_n	x_n

SCATTERGRAM



Contents

Traditional Linear Regression

- Linear CER Additive-Error Model

$$y = a + bx + \varepsilon$$

(Actual Cost = Estimated Cost + Error of Estimation)

- Ordinary Least-Squares (OLS) Regression Minimizes Sum of Squared Errors
  - Actual cost for data point i is y_i
  - Estimated cost for data point i is $a + bx_i$
  - Error of estimation for data point i is $\varepsilon_i = y_i - (a + bx_i)$
  - Choose values for a and b that minimize $\sum (y_i - a - bx_i)^2 = \sum \varepsilon_i^2$
  - Resulting estimates are unbiased
- OLS Solution:

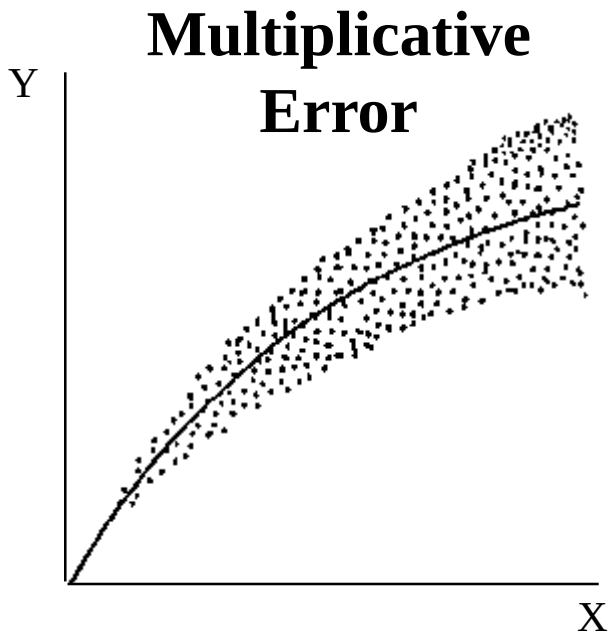
$$b = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{n \sum x_i^2 - (\sum x_i)^2}$$

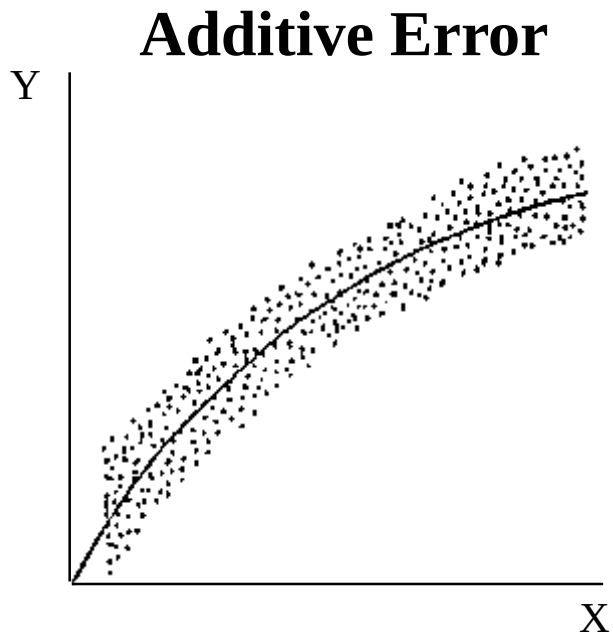
$$a = \frac{\sum y_i - b \sum x_i}{n}$$

OLS Standard Error of the Estimate

- A "One-Sigma"-type Error Bound on Error Implicit in OLS CERs
- SEE = $\sqrt{\frac{1}{n-k}\sum_{i=1}^{n}(y_i - a - bx_i)^2}$ Dollars
  - $y = a+bx$ is the CER that Expresses Cost (y) in Terms of a Cost-Driving Technical or Programmatic Parameter (x)
  - n is the Number of Data Points Used to Derive the CER
  - k is the Number of Parameters in the Algebraic Expression for the CER, e.g., $k = 2$ for the CER $y = a+bx$
- Standard Error is an OLS CER Quality Metric

But, There are Alternative Error Specifications





Reference: **H.L. Eskew and K.S. Lawler, "Correct and Incorrect Error Specifications in Statistical Cost Models," *Journal of Cost Analysis*, Spring 1994, page 107.**

Contents

- CER Development
  - Ordinary Least Squares (OLS)
  - Log-transformed OLS
    - Statistical Issues
    - Quality Metrics
- General-Error Regression
  - Iteratively Reweighted Least Squares (IRLS)
    - Factor CERs
    - Linear CERs
    - Triad CERs
  - Quality Metrics
  - Zero Percentage Bias, Minimum Percentage Error (ZMPE)
- Summary

Some History: The 18th Century Approach to Nonlinear* Regression

- Consider the Nonlinear Power Model $y = ax^b$
- Take Logarithms of Both Sides:

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

- Determine a and b to Predict $\log y$:
  - Assume Additive-Error Model

$$\log y = \log a + b \log x + E,$$

where $E = \log y - (\log a + b \log x)$ is Error of Estimation in Predicting Logarithm of Cost

  - Choose Values for a and b that Minimize

$$\sum(\log y_i - \log a - b \log x_i)^2 = \sum E_i^2$$

- Logarithmic Transformation of Nonlinear Form Into Linear Form Permits Use of OLS Mathematics to Solve Nonlinear Problem
- Excel "Trend Line" Function and Other Common "Quickie" Approaches Use This Technique

*For the purposes of this discussion, a "non-linear" model refers to a power function

Nonlinear OLS-Based Solution

- Predict $\log y = \log a + b \log x = A + b \log x$ where

$$b = \frac{n\sum_{i=1}^{n}(\log y_i)(\log x_i) - \sum_{i=1}^{n}(\log y_i)\sum_{i=1}^{n}(\log x_i)}{n\sum_{i=1}^{n}(\log x_i)^2 - \left(\sum_{i=1}^{n}\log x_i\right)^2}$$

$$A = \log a = \frac{\sum_{i=1}^{n}\log y_i - b\left(\sum_{i=1}^{n}\log x_i\right)}{n}$$

$$(a = 10^{\log a} = 10^{A})$$

- Standard Error of Estimate is reported as

$$SEE = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\left(\log y_i - \log a - b\log x_i\right)^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}E_i^2}$$

A Word on Error of Estimation

- In Order for Logarithms to Work, Nonlinear CER Model Must be "Multiplicative-Error" Model

$$y = ax^b \varepsilon$$

Because Applying Logarithms Yields Additive-Error Model for Predicting Logarithm of Cost

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

- Setting $E_i = \log \varepsilon_i$, what is Actually Minimized is

$$\sum(\log \varepsilon_i)^2 = \sum E_i^2 = \sum (\log y_i - \log a - b \log x_i)^2$$

Instead of $\sum \varepsilon_i^2$

Disconnects Between OLS and Log-OLS Regression

- Bad: Minimizing $\sum (\log \varepsilon_i)^2$ not Same as Minimizing $\sum \varepsilon_i^2$ (as in Traditional Linear Regression)
  - a and b Values Turn out to be Different
  - Error of Estimating Logarithm of Cost is Minimized
  - Error Expressed in Meaningless Units ("log dollars")
- Worse: Standard Error in Linear Case $\sqrt{\frac{1}{n-2}\sum \varepsilon_i^2}$ Cannot be Compared with Standard Error in Nonlinear Case to see which Functional Form is the Better Estimator
- Worst: Log-OLS Nonlinear CERs Must Have Multiplicative Error; OLS Linear CERs Must Have Additive Error $\sqrt{\frac{1}{n-2}\sum (\log \varepsilon_i)^2}$

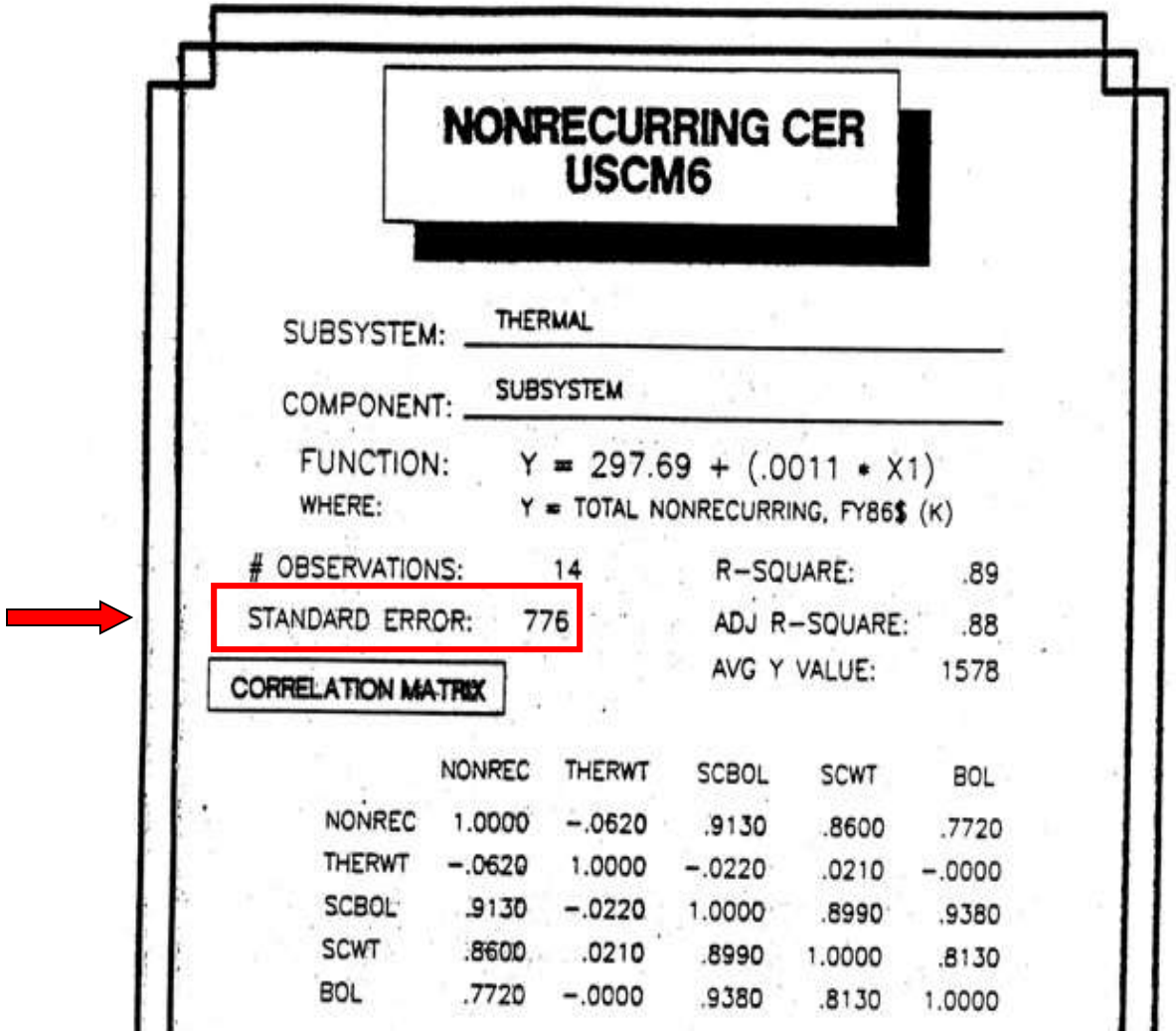
Even Worse

- Capability of Logarithmic-Transformation Method is Limited by the Mathematical Properties of Logarithms
- Major Casualty of This Situation is not Having Access to the Nonlinear Functional Form

$$y = a + bx^c$$

- Therefore, Using the OLS Method for Linear Regression and the OLS-Based Logarithmic-Transformation Method for Nonlinear Regression Puts Us in Another Self-Contradictory Situation:
  - We Allow Linear CERs to Have a Nonzero Fixed-Cost Component, Namely a, but …
  - We Require Nonlinear CERs to Pass Through the (0,0) Point

USCM6 Linear CER with Statistics



NONRECURRING CER
USCM6

SUBSYSTEM: THERMAL

COMPONENT: SUBSYSTEM

FUNCTION: Y = 297.69 + (.0011 * X1)

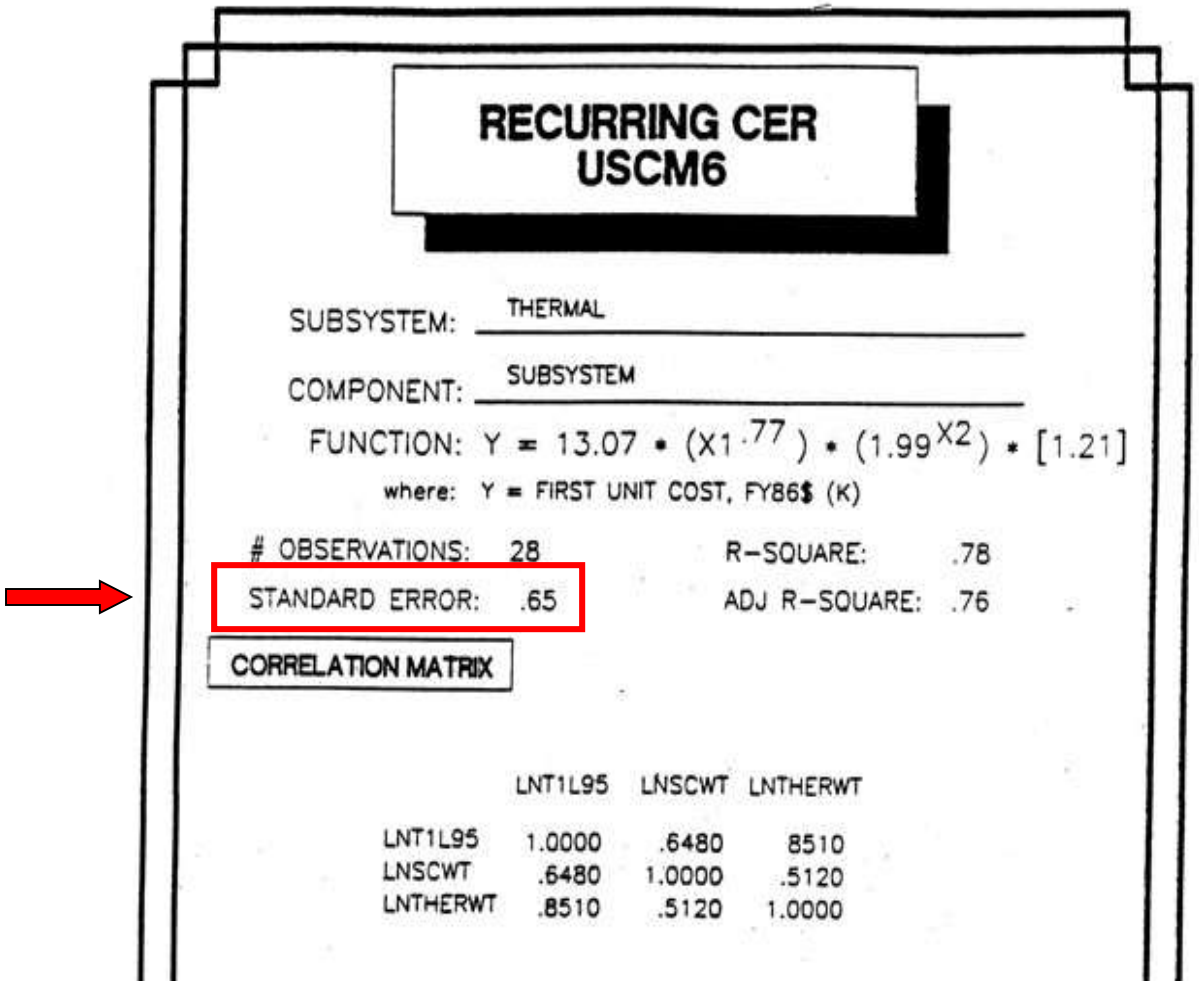
WHERE: Y = TOTAL NONRECURRING, FY86$ (K)

# OBSERVATIONS:	14	R-SQUARE:	.89
STANDARD ERROR:	776	ADJ R-SQUARE:	.88
		AVG Y VALUE:	1578

CORRELATION MATRIX

	NONREC	THERWT	SCBOL	SCWT	BOL
NONREC	1.0000	-.0620	.9130	.8600	.7720
THERWT	-.0620	1.0000	-.0220	.0210	-.0000
SCBOL	.9130	-.0220	1.0000	.8990	.9380
SCWT	.8600	.0210	.8990	1.0000	.8130
BOL	.7720	-.0000	.9380	.8130	1.0000

USCM6 Nonlinear CER with Statistics



RECURRING CER
USCM6

SUBSYSTEM: THERMAL

COMPONENT: SUBSYSTEM

FUNCTION: $Y = 13.07 \cdot (X1^{.77}) \cdot (1.99^{X2}) \cdot [1.21]$

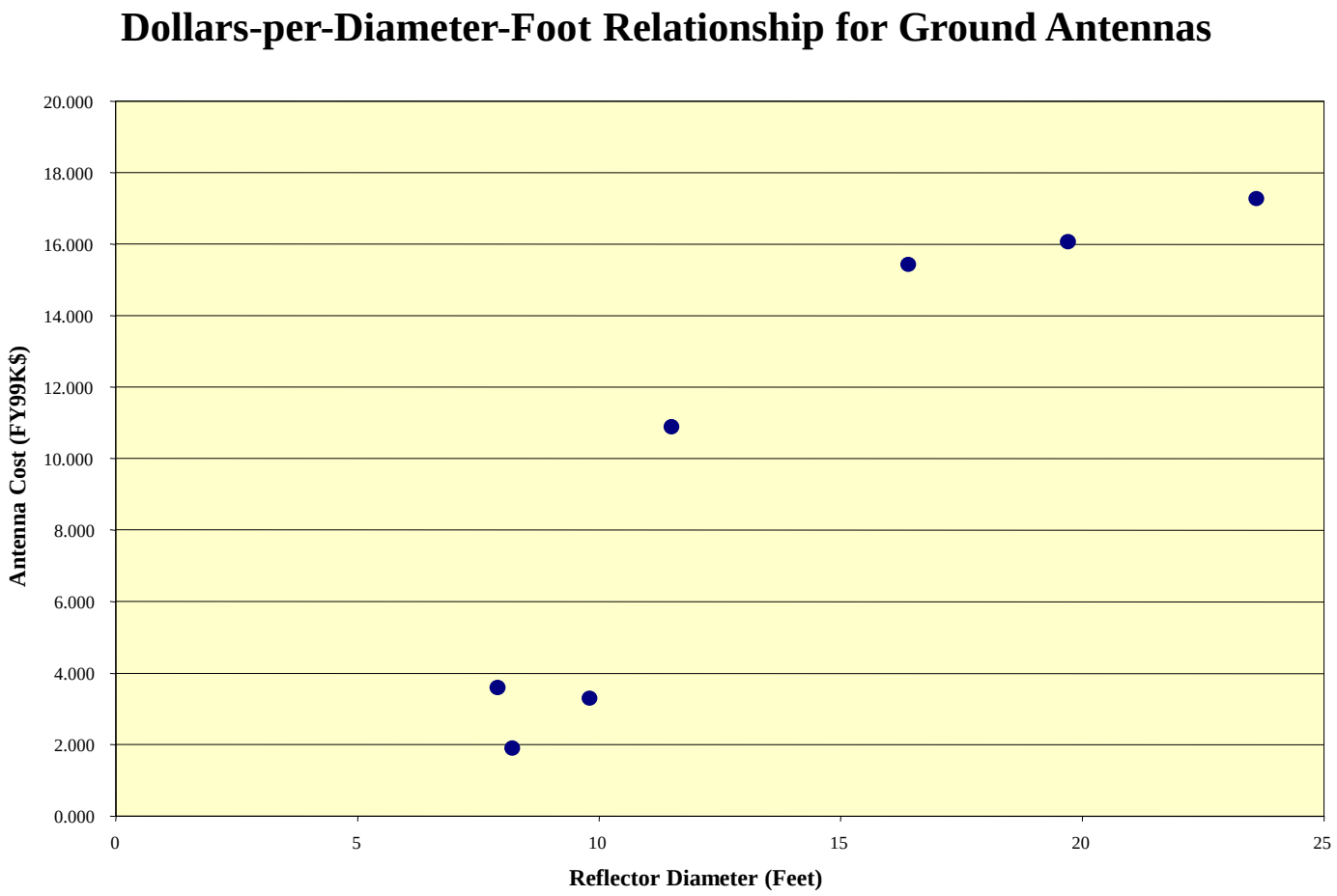
where: Y = FIRST UNIT COST, FY86$ (K)

# OBSERVATIONS:	28	R-SQUARE:	.78
STANDARD ERROR:	.65	ADJ R-SQUARE:	.76

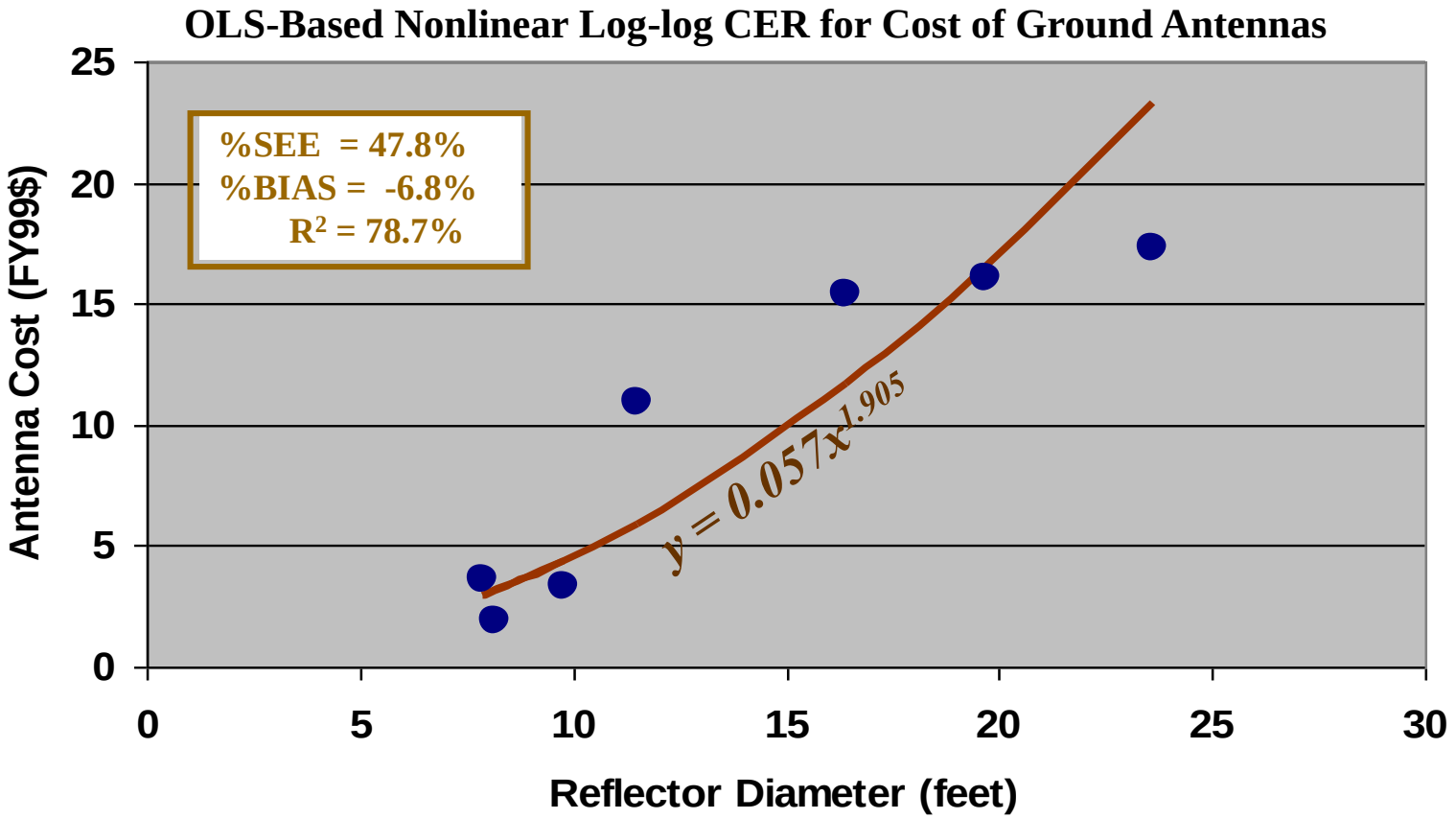
CORRELATION MATRIX

	LNT1L95	LNSCWT	LNTHERWT
LNT1L95	1.0000	.6480	8510
LNSCWT	.6480	1.0000	.5120
LNTHERWT	.8510	.5120	1.0000

CER Example: Antenna Cost vs. Reflector Diameter



OLS Nonlinear CER and Its Quality Metrics (Compared with Data Set on Previous Chart)



Note: CER $y = ax^b$ has opposite concavity from data set, due to its necessity to pass through (0,0) point.

Contents

- CER Development
  - Ordinary Least Squares (OLS)
  - Log-transformed OLS
    - Statistical Issues
    - Quality Metrics
- General-Error Regression
  - Iteratively Reweighted Least Squares (IRLS)
    - Factor CERs
    - Linear CERs
    - Triad CERs
  - Quality Metrics
  - Zero Percentage Bias, Minimum Percentage Error (ZMPE)
- Summary

General Multiplicative-Error Model Eliminates All These Problems

- No Logarithms
  - Functional Forms Predict Cost, Not Logarithm of Cost
  - Standard Errors Can Be Compared and Ranked in Magnitude for All Functional Forms
  - Error Model (Additive or Multiplicative) Can Be Chosen Independently of Functional Form

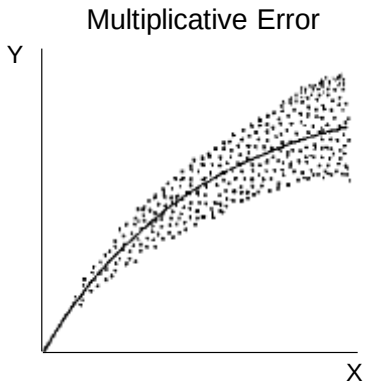
- Take Advantage of Modern Computing Capability
  - Least-squares Minimization Problem Does Not Have to Be Solved Explicitly (to Get Formulas for a and b as in the Linear Additive-error Case)
  - Sequential-search Techniques Based on Newton's Method or Simplex Method Are Used to Find Error-minimizing Values of a and b
  - All Functional Forms Can Be Considered, Even

$$y = a + bx^c$$

Multiplicative-Error Model

- Actual Cost Equals Estimate times Error

$$y = f(x) \times \varepsilon$$



- Error is Ratio of Actual to Estimate

$$\varepsilon = \frac{y}{f(x)} = \frac{\textbf{Actual}}{\textbf{Estimate}}$$

- Minimum Percentage Error (MPE) CERs: Choose f(x)'s Coefficients so that Sum of Squared Percentage Errors

$$\sum (\varepsilon_i - 1)^2 = \sum \left[\frac{y_i - f(x_i)}{f(x_i)} \right]^2$$

  is as Small as Possible

- Actual Cost = Estimate ± Percentage of Estimate

Percentage Standard Error of the Estimate

- "One-Sigma"-type Error Bound that Characterizes Multiplicative-Error CERs $y = f(x)$
- %SEE = $\sqrt{\frac{1}{n-k}\sum_{i=1}^{n}\left[\frac{y_i - f(x_i)}{f(x_i)}\right]^2}$ x 100%
  - $y = f(x)$ is the CER that Expresses Cost (y) in Terms of a Technical or Programmatic Cost Driver (x)
  - n is the Number of Data Points Used to Derive the CER
  - k is the Number of Parameters in the CER's Algebraic Expression, e.g., $k = 2$ for the CER $y = ax^b$
- Percentage Standard Error is a CER Quality Metric

Multiplicative-Error Regression

- Using $f(x) = a + bx^c$ for Purposes of Illustration . . .
- Multiplicative-Error Model

$$y = \left(a + bx^c\right) \times \varepsilon \quad \text{where} \quad \varepsilon = \frac{y}{a + bx^c}$$

- For Best Results, ε Should be as Close to One as Possible
  - Choose a, b, c so that $\sum \left(\varepsilon_i - 1\right)^2 = \sum \left(\frac{y_i - a - bx_i^c}{a + bx_i^c} \right)^2$ is as Small as Possible
  - Apply Computation-Intensive Techniques of Numerical Analysis

Percentage Bias

- Sample Percentage Bias = $\frac{1}{n}\Sigma\left[\frac{f(x_i)-y_i}{f(x_i)}\right]$
  MPE Procedure Tends to Produce Estimates that are Biased Upward
  - This is Bad?
  - Bias Seems to Occur Because $\Sigma\left(\frac{y_i-f(x_i)}{f(x_i)}\right)^2$ will be Smaller if the $f(x)$ Values are Larger
- Tecolote Recommended that MPE be Replaced by "Iteratively Reweighted Least Squares" (IRLS), a Standard Statistical Method that Eliminates Percentage Bias in Minimum-Percentage-Error Estimates
- Percentage Bias is a CER Quality Metric

Contents

- CER Development
  - Ordinary Least Squares (OLS)
  - Log-transformed OLS
    - Statistical Issues
    - Quality Metrics
- General-Error Regression
  - Iteratively Reweighted Least Squares (IRLS)
    - Factor CERs
    - Linear CERs
    - Triad CERs
  - Quality Metrics
  - Zero Percentage Bias, Minimum Percentage Error (ZMPE)
- Summary

Iteratively Reweighted Least Squares

- USCM Refers to IRLS as “Minimum Unbiased Percentage Error” (MUPE)
- Solve by Generating Sequences of Coefficients (beginning with “initial guess” a_0, b_0, c_0)
  - $a_1, a_2, a_3, \ldots$
  - $b_1, b_2, b_3, \ldots$
  - $c_1, c_2, c_3, \ldots$
- Given a_j, b_j, c_j calculate $a = a_{j+1}, b = b_{j+1}, c = c_{j+1}$ by minimizing

$$\sum \left[\frac{y_i - a - bx^c}{a_j + b_j x^{c_j}} \right]^2$$

- a, b, c = Respective Limits of Sequences if Sequences Converge
  - IRLS CER is $y = a + bx^c$
  - Not the same coefficient values as MPE CER

R^2 Between Estimates and Actuals*

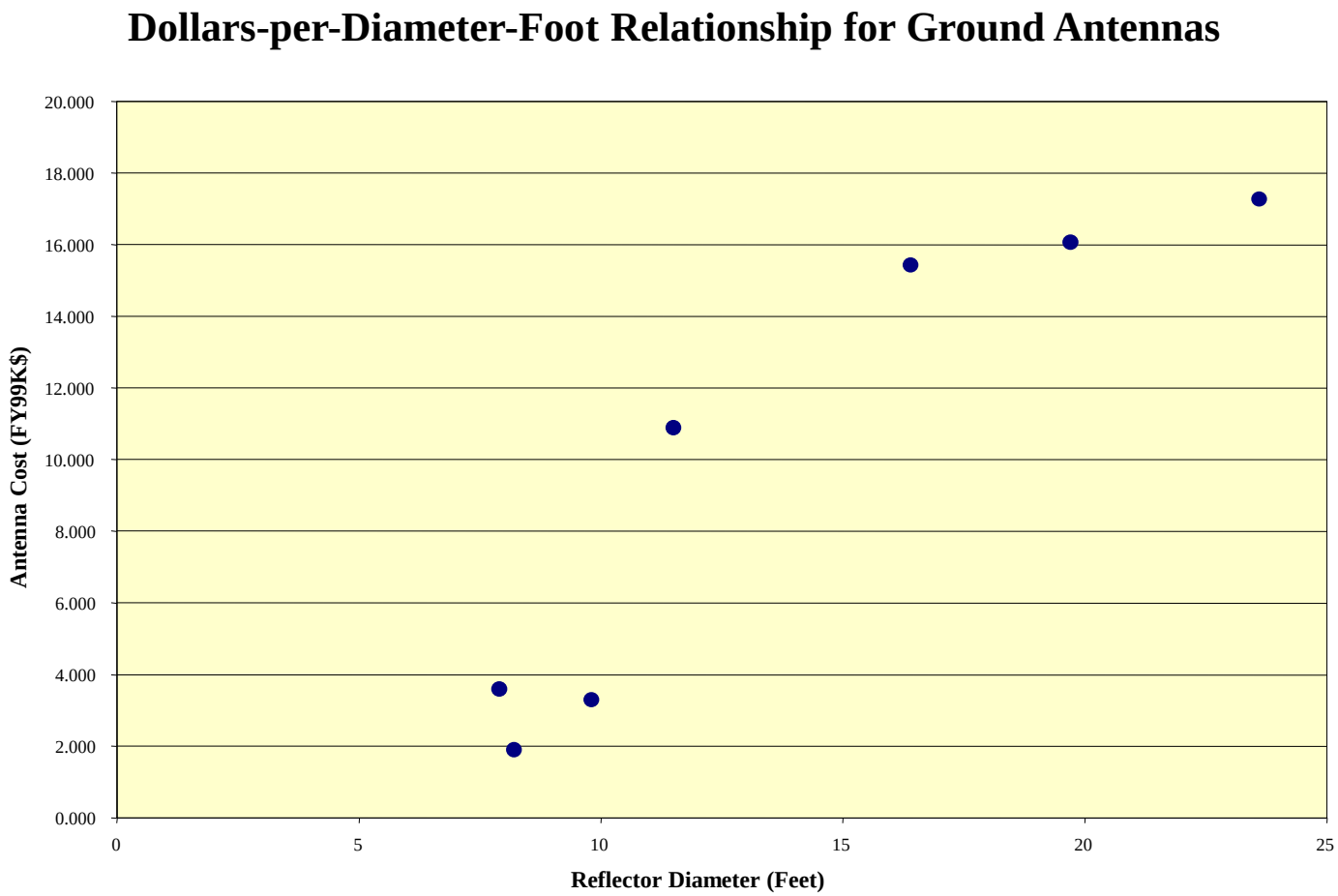
- The term R is sometimes used to describe the Correlation Between Actuals (x values) and Estimates (y values), Measuring the Extent to which the Relationship between x and y is Linear
  - R Does not Depend on Specific Coefficients of Relationship
  - R^2 = Proportion of Variation in Estimates (y) that is Attributable, through a OLS Linear Relationship, to Variations in Actuals (x)
  - Larger (closer to 1.00) Values of R^2 Indicate Better Linear Fit
- If CER is "Good", Estimates Should be Pretty Close to Actuals, i.e., the (Actual, Estimate) = (x,y) Points Should Lie Along Straight Line

  $y = x$
- R^2 ("Pearson's Correlation Squared") is a CER Quality Metric

$$R^2 = \frac{\left[n\sum_{k=1}^{n} x_k y_k - \sum_{k=1}^{n} x_k \sum_{k=1}^{n} y_k \right]^2}{\left[n\sum_{k=1}^{n} x_k{}^2 - \left(\sum_{k=1}^{n} x_k \right)^2 \right] \left[n\sum_{k=1}^{n} y_k{}^2 - \left(\sum_{k=1}^{n} y_k \right)^2 \right]}$$

*Note: this is different than the classical definition of R^2, which arises in OLS, and is defined as 1-SSE/SST.

CER Example: Antenna Cost vs. Reflector Diameter



Case 1: Calculating Multiplicative-Error "Factor" CER $y = ax \times \varepsilon$ Using IRLS Method

- Choose Initial Guess a_0 (from looking at the dollars-per-pound scattergram)
- Proceed from a_0 Through the Sequence a_1, a_2, a_3,… by Successively Minimizing the IRLS "Objective Function"

$$F(a_{j+1}) = \sum_{i=1}^{n} \left(\frac{y_i - a_{j+1} x_i}{a_j x_i} \right)^2$$

$$F'(a_{j+1}) = \frac{d}{da_{j+1}} \sum_{i=1}^{n} \left(\frac{y_i - a_{j+1} x_i}{a_j x_i} \right)^2 = \sum_{i=1}^{n} 2 \left(\frac{y_i - a_{j+1} x_i}{a_j x_i} \right) \left(\frac{-x_i}{a_j x_i} \right) = -\frac{2}{a_j} \sum_{i=1}^{n} \left(\frac{y_i - a_{j+1} x_i}{a_j x_i} \right)$$

$$= -\frac{2}{a_j^2} \left(\sum_{i=1}^{n} \frac{y_i}{x_i} - n a_{j+1} \right) = 0 \text{ when } a_{j+1} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{x_i}$$

- Note that a_{j+1} Does not Depend on j, so the IRLS Factor CER is $y = ax$, where a is Calculated Using the Formula for a_{j+1} Above

Note: IRLS-Derived Factor CER Has Zero Percentage Bias

- Percentage Bias =

$$PB(a) = \sum_{i=1}^{n}\left(\frac{y_i - ax_i}{ax_i}\right) = \sum_{i=1}^{n}\left(\frac{y_i}{ax_i} - 1\right) = \frac{1}{a}\sum_{i=1}^{n}\left(\frac{y_i}{x_i}\right) - n$$

- We Insert the IRLS Solution for a, namely

- And then we Obtain $$a = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{x_i}$$

$$PB(a) = \frac{1}{a}\sum_{i=1}^{n}\left(\frac{y_i}{x_i}\right) - n = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i}{x_i}\right)}\sum_{i=1}^{n}\left(\frac{y_i}{x_i}\right) - n = n - n = 0$$

Dollars-per-Foot Data Base for IRLS Factor CER for Ground Antennas

Antenna Cost vs. Reflector Diameter Data			IRLS Factor CER and Quality Metrics					
	Diameter (x)	Cost (y)						
	(Feet)	FY99$K	y/x	n	a	ESTy = ax	%SE	%BIAS
	9.8	3.300	0.337	7	0.637	6.244	0.222	-0.472
	7.9	3.595	0.455	7	0.637	5.034	0.082	-0.286
	8.2	1.900	0.232	7	0.637	5.225	0.405	-0.636
	11.5	10.900	0.948	7	0.637	7.328	0.238	0.488
	16.4	15.434	0.941	7	0.637	10.450	0.227	0.477
	19.7	16.074	0.816	7	0.637	12.553	0.079	0.281
	23.6	17.274	0.732	7	0.637	15.038	0.022	0.149
Sums	97.1	68.477	4.460			Totals:	1.275	0.000

EST = Estimated
FY = Fiscal Year
SE = Squared Error

Dollars-per-Foot IRLS Factor CER

- Based on Computations on the Historical Data …

$$a = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{x_i} = \frac{1}{7}(4.460) = 0.637$$

- Multiplicative-Error CER

$$y = 0.637x \quad (= 637\ \$/\text{diam-ft})$$

- Standard Error of the Estimate (%SEE)

$$Standard\ Error = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{y_i - ax_i}{ax_i}\right)^2}$$

$$= \sqrt{\frac{1}{7-1}(1.275)} = 0.461$$

(Average 46.1% Across Data Range)

Calculation of R^2 Quality Metric for IRLS Factor CER

n	Actual y	ESTy	x2	y2	xy
7	3.300	6.244	10.890	38.993	20.607
	3.595	5.034	12.924	25.339	18.096
	1.900	5.225	3.610	27.300	9.927
	10.900	7.328	118.810	53.695	79.872
	15.434	10.450	238.208	109.200	161.284
	16.074	12.553	258.373	157.568	201.771
	17.274	15.038	298.391	226.131	259.760
Sums	**68.477**	**61.871**	**941.207**	**638.225**	**751.316**

R-Squared =	**0.861**

R2 Numer Term =	**1022.476**
R2 Denom1 Term =	**1899.349**
R2 Denom2 Term =	**639.561**

R^2 Between Actuals and Estimates

Actual Cost (y)	Estimated Cost (ax)
3.300	7.382
3.595	5.951
1.900	6.177
10.900	8.662
15.434	12.353
16.074	14.839
17.274	17.777

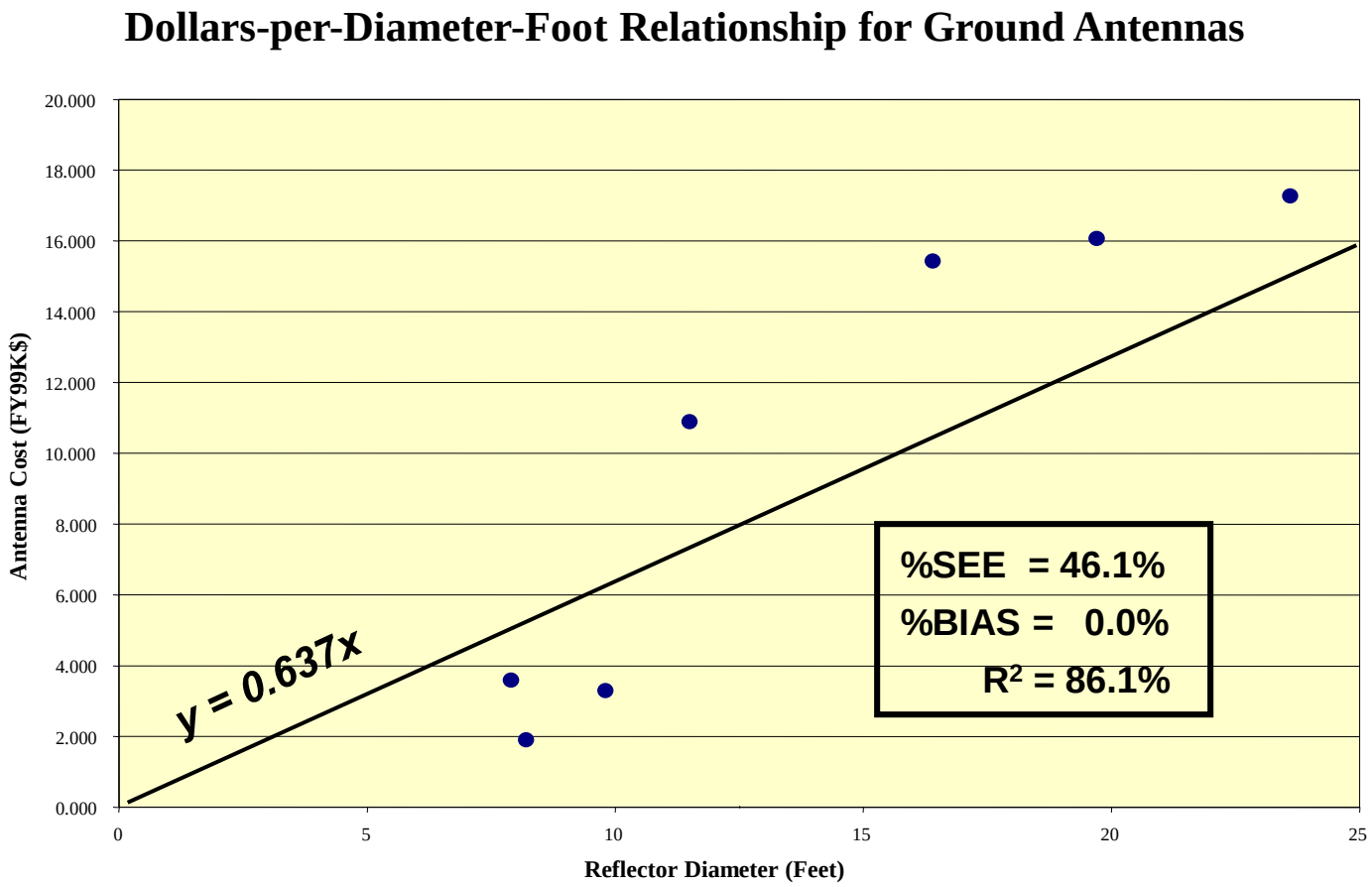
- $$R^2 = \frac{\left[n\sum_{k=1}^{n} x_k y_k - \sum_{k=1}^{n} x_k \sum_{k=1}^{n} y_k\right]^2}{\left[n\sum_{k=1}^{n} x_k^{\,2} - \left(\sum_{k=1}^{n} x_k\right)^2\right]\left[n\sum_{k=1}^{n} y_k^{\,2} - \left(\sum_{k=1}^{n} y_k\right)^2\right]}$$

$$= \frac{(1022.476)^2}{[1899.349][639.561]} = 0.861$$

$$= 86.1\%$$

  - **Therefore the R^2 Quality Metric for this CER is 86.1% (Perfect Fit is 100%)**

IRLS Factor CER and Its Quality Metrics Superimposed on Data Base



Case 2: Determining Multiplicative-Error Linear CER $y = (a+bx) \times \varepsilon$ Using IRLS Method

- Choose Initial Guess a_0+b_0x (by looking at the dollars-per-pound scattergram)
- Proceed from a_0 and b_0 Through the Sequences a_1, a_2, a_3,… and b_1, b_2, b_3,… by Successively Minimizing the IRLS "Objective Function"

$$F(a_{j+1}, b_{j+1}) = \sum_{i=1}^{n} \left(\frac{y_i - a_{j+1} - b_{j+1} x_i}{a_j + b_j x_i} \right)^2$$

Calculating the IRLS Coefficients a_{j+1}, b_{j+1}

- Use Partial Derivatives to Minimize the Objective Function and Find the Optimal Values of a_{j+1} and b_{j+1}

$$\frac{\partial}{\partial a_{j+1}} F(a_{j+1}, b_{j+1}) = \frac{\partial}{\partial a_{j+1}} \sum_{i=1}^{n} \left(\frac{y_i - a_{j+1} - b_{j+1} x_i}{a_j + b_j x_i} \right)^2 = \sum_{i=1}^{n} 2 \left(\frac{y_i - a_{j+1} - b_{j+1} x_i}{a_j + b_j x_i} \right) \left(\frac{-1}{a_j + b_j x_i} \right)$$

$$= -2 \sum_{i=1}^{n} \frac{y_i - a_{j+1} - b_{j+1} x_i}{(a_j + b_j x_i)^2}$$

$$= -2 \left(\sum_{i=1}^{n} \frac{y_i}{(a_j + b_j x_i)^2} - a_{j+1} \sum_{i=1}^{n} \frac{1}{(a_j + b_j x_i)^2} - b_{j+1} \sum_{i=1}^{n} \frac{x_i}{(a_j + b_j x_i)^2} \right)$$

$$\frac{\partial}{\partial b_{j+1}} F(a_{j+1}, b_{j+1}) = \frac{\partial}{\partial b_{j+1}} \sum_{i=1}^{n} \left(\frac{y_i - a_{j+1} - b_{j+1} x_i}{a_j + b_j x_i} \right)^2 = \sum_{i=1}^{n} 2 \left(\frac{y_i - a_{j+1} - b_{j+1} x_i}{a_j + b_j x_i} \right) \left(\frac{-x_i}{a_j + b_j x_i} \right)$$

$$= -2 \sum_{i=1}^{n} \frac{x_i y_i - a_{j+1} x_i - b_{j+1} x_i^2}{(a_j + b_j x_i)^2}$$

$$= -2 \left(\sum_{i=1}^{n} \frac{x_i y_i}{(a_j + b_j x_i)^2} - a_{j+1} \sum_{i=1}^{n} \frac{x_i}{(a_j + b_j x_i)^2} - b_{j+1} \sum_{i=1}^{n} \frac{x_i^2}{(a_j + b_j x_i)^2} \right)$$

Solving the Equations for a_{j+1} and b_{j+1}

- Set Both Partial Derivatives to Zero and Solve the Resulting Simultaneous Equations

$$\sum_{i=1}^{n} \frac{y_i}{(a_j + b_j x_i)^2} = a_{j+1} \sum_{i=1}^{n} \frac{1}{(a_j + b_j x_i)^2} + b_{j+1} \sum_{i=1}^{n} \frac{x_i}{(a_j + b_j x_i)^2}$$

$$\sum_{i=1}^{n} \frac{x_i y_i}{(a_j + b_j x_i)^2} = a_{j+1} \sum_{i=1}^{n} \frac{x_i}{(a_j + b_j x_i)^2} + b_{j+1} \sum_{i=1}^{n} \frac{x_i^2}{(a_j + b_j x_i)^2}$$

- Solving these Equations Yields the IRLS-Optimal Values of the Coefficients at Stage j+1

The Resulting Values of a_{j+1} and b_{j+1}

$$b_{j+1} = \frac{\left(\sum_{i=1}^{n} \frac{x_i y_i}{(a_j + b_j x_i)^2}\right)\left(\sum_{i=1}^{n} \frac{1}{(a_j + b_j x_i)^2}\right) - \left(\sum_{i=1}^{n} \frac{x_i}{(a_j + b_j x_i)^2}\right)\left(\sum_{i=1}^{n} \frac{y_i}{(a_j + b_j x_i)^2}\right)}{\left(\sum_{i=1}^{n} \frac{x_i^2}{(a_j + b_j x_i)^2}\right)\left(\sum_{i=1}^{n} \frac{1}{(a_j + b_j x_i)^2}\right) - \left(\sum_{i=1}^{n} \frac{x_i}{(a_j + b_j x_i)^2}\right)^2}$$

$$a_{j+1} = \frac{\sum_{i=1}^{n} \frac{y_i}{(a_j + b_j x_i)^2} - b_{j+1} \sum_{i=1}^{n} \frac{x_i}{(a_j + b_j x_i)^2}}{\sum_{i=1}^{n} \frac{1}{(a_j + b_j x_i)^2}}$$

Note: IRLS-Derived Linear CER Has Zero Percentage Bias

- Percentage Bias =

$$PB(a,b) = \sum_{i=1}^{n}\left(\frac{y_i - a - bx_i}{a + bx_i}\right) = \sum_{i=1}^{n}\left(\frac{y_i}{a + bx_i} - 1\right) = \sum_{i=1}^{n}\left(\frac{y_i}{a + bx_i}\right) - n$$

- Unlike the Factor CER Case, the Algebra Cannot be Worked Out, because there are no "Closed" Expressions for a and b
- However, Percentage Bias Can be Calculated Numerically in any Particular Case and Turns Out to be Zero

Diameter-vs.-Cost Data Base and Computations for IRLS Linear CER

Antenna Cost vs. Reflector Diameter Data											
	Diameter (x)	Cost (y)									
n	(Feet)	FY99$K	x2	xy	a+bx	(a+bx)2	1/(a+bx)2	x/(a+bx)2	y/(a+bx)2	x2/(a+bx)2	xy/(a+bx)2
7	9.8	3.300	96.040	32.340	5.055	25.551	0.039	0.384	0.129	3.759	1.266
7	7.9	3.595	62.410	28.401	2.706	7.325	0.137	1.079	0.491	8.521	3.877
7	8.2	1.900	67.240	15.580	3.077	9.469	0.106	0.866	0.201	7.101	1.645
7	11.5	10.900	132.250	125.350	7.156	51.208	0.020	0.225	0.213	2.583	2.448
7	16.4	15.434	268.960	253.118	13.212	174.568	0.006	0.094	0.088	1.541	1.450
7	19.7	16.074	388.090	316.658	17.291	298.986	0.003	0.066	0.054	1.298	1.059
7	23.6	17.274	556.960	407.666	22.112	488.923	0.002	0.048	0.035	1.139	0.834
Sums	97.1	68.477	1571.950	1179.112	70.610	1056.029	0.312	2.761	1.211	25.941	12.579

EST = Estimated
FY = Fiscal Year
SE = Squared Error

Diameter-vs.-Cost Data Base and IRLS Linear CER Statistics

Antenna Cost vs. Reflector Diameter Data			IRLS Linear CER and Quality Metrics				
n	Diameter (x) (Feet)	Cost (y) FY99$K	b	a	ESTy = a+bx	%SE	%BIAS
7	9.8	3.300	1.236	-7.058	5.055	0.121	-0.347
7	7.9	3.595	1.236	-7.058	2.707	0.108	0.328
7	8.2	1.900	1.236	-7.058	3.078	0.146	-0.383
7	11.5	10.900	1.236	-7.058	7.157	0.274	0.523
7	16.4	15.434	1.236	-7.058	13.213	0.028	0.168
7	19.7	16.074	1.236	-7.058	17.292	0.005	-0.070
7	23.6	17.274	1.236	-7.058	22.113	0.048	-0.219
Sums	97.1	68.477			Totals:	0.729	0.000

EST = Estimated
FY = Fiscal Year
SE = Squared Error

Diameter-vs.-Cost IRLS Linear CER

- Based on Computations on the Historical Data …

$$a = -7.058; \quad b = 1.236$$

- Multiplicative-Error CER

$$y = -7.058 + 1.236x$$

- Standard Error of the Estimate (%SEE)

$$Standard\ Error = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\left(\frac{y_i - a - bx_i}{a + bx_i}\right)^2}$$

$$= \sqrt{\frac{1}{7-2}(0.729)} = 0.382$$

(Average 38.2% Across Data Range)

Calculation of R^2 Quality Metric for IRLS Linear CER

Antenna Cost vs. Reflector Diameter Data

n	ESTy (x) FY99$K	Cost (y) FY99$K	x2	y2	xy
7	5.055	3.300	25.557	10.890	16.683
	2.707	3.595	7.328	12.924	9.732
	3.078	1.900	9.473	3.610	5.848
	7.157	10.900	51.218	118.810	78.008
	13.213	15.434	174.588	238.208	203.932
	17.292	16.074	299.015	258.373	277.952
	22.113	17.274	488.964	298.391	381.972
Sums	**70.615**	**68.477**	**1056.143**	**941.207**	**974.126**

Num R = 1983.405

Den1 R2 = 2406.568

Den2 R2 = 1899.349

R2 = 0.861

R^2 Between Actuals and Estimates

ESTy (x)	Cost (y)
5.055	3.300
2.707	3.595
3.078	1.900
7.157	10.900
13.213	15.434
17.292	16.074
22.113	17.274

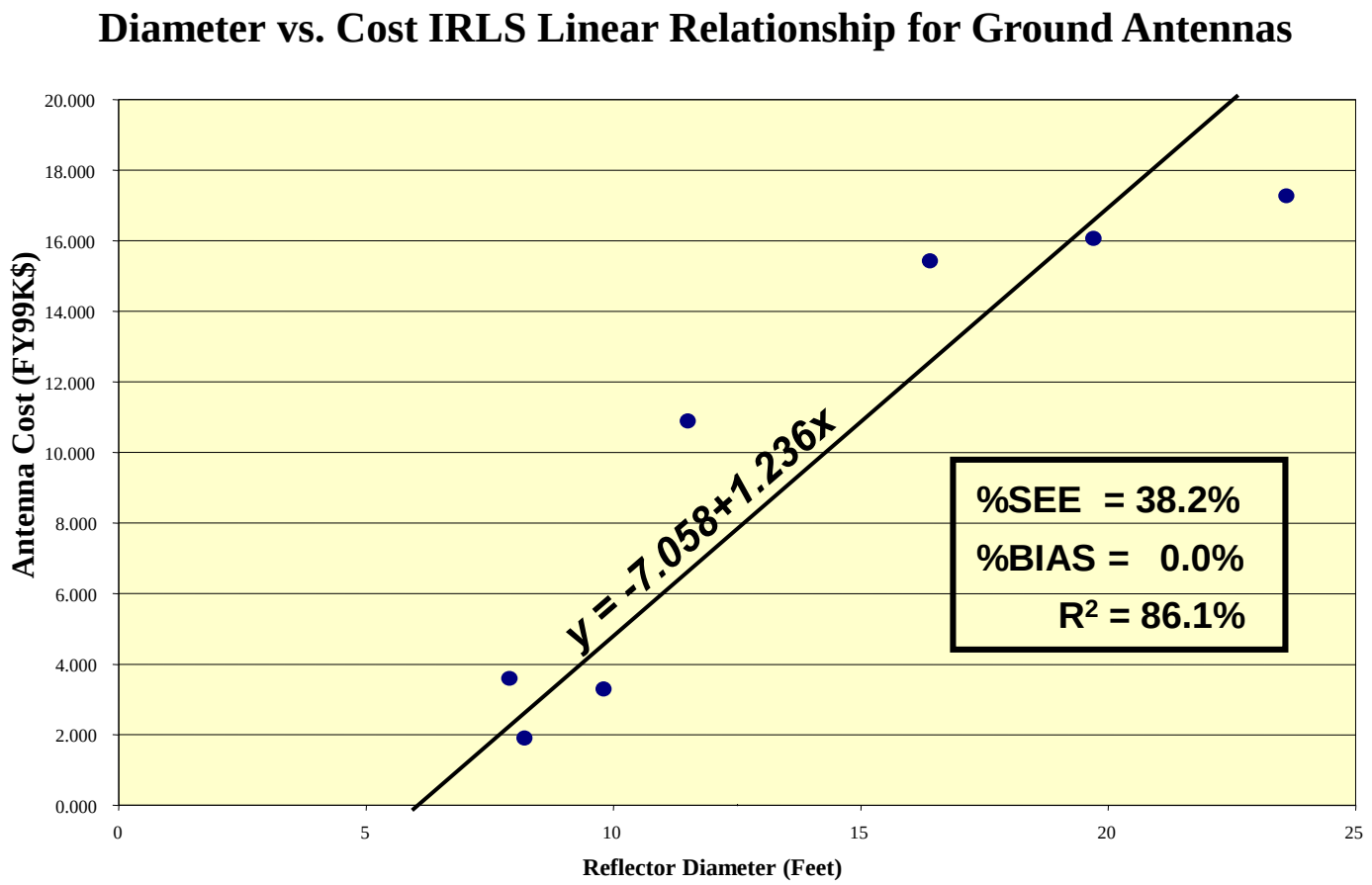
- $$R^2 = \frac{\left[n\sum_{k=1}^{n} x_k y_k - \sum_{k=1}^{n} x_k \sum_{k=1}^{n} y_k\right]^2}{\left[n\sum_{k=1}^{n} x_k^{\,2} - \left(\sum_{k=1}^{n} x_k\right)^2\right]\left[n\sum_{k=1}^{n} y_k^{\,2} - \left(\sum_{k=1}^{n} y_k\right)^2\right]}$$

$$= \frac{(1983.405)^2}{[2406.568][1899.349]} = 0.861$$

$$= 86.1\%$$

- Therefore the R^2 Quality Metric for this CER is 86.1% (Perfect Fit is 100%)

IRLS Linear CER and Its Quality Metrics Superimposed on Data Base



Case 3: Determining Multiplicative-Error Triad CER $y = (a+bx^c) \times \varepsilon$ using IRLS Method

- Choose Initial Guess $a_0 + b_0 x^{c_0}$ (from looking at the dollars-per-pound scatter-gram)
- Proceed from a_0, b_0, and c_0 Through the Sequences a_1, a_2, a_3,… , b_1, b_2, b_3,… , and c_1, c_2, c_3,… by Successively Minimizing the IRLS "Objective Function"

$$F(a_{j+1}, b_{j+1}, c_{j+1}) = \sum_{i=1}^{n} \left(\frac{y_i - a_{j+1} - b_{j+1} x_i^c}{a_j + b_j x_i^c} \right)^2$$

Attempt to Calculate IRLS Triad Coefficients a_{j+1}, b_{j+1}, c_{j+1}

- Use Partial Derivatives (with respect to a_{j+1}, b_{j+1}, and c_{j+1}) as Before to Minimize Objective Function and Establish Simultaneous Equations for Optimal Values of a_{j+1}, b_{j+1}, c_{j+1}
- Unfortunately, These Simultaneous Equations are not Linear and Cannot be Solved Explicitly to Get Algebraic Expressions for a_{j+1}, b_{j+1}, c_{j+1}
- "Optimal" Numerical Values of a_{j+1}, b_{j+1}, and c_{j+1} Must be Found by Iterative Techniques, such as those Used by Excel's "Solver" Routine

Note: IRLS-Derived Triad CER Has Zero Percentage Bias

- Percentage Bias =

$$PB(a,b) = \sum_{i=1}^{n}\left(\frac{y_i - a - bx_i^c}{a + bx_i^c}\right)$$

$$= \sum_{i=1}^{n}\left(\frac{y_i}{a + bx_i^c} - 1\right) = \sum_{i=1}^{n}\left(\frac{y_i}{a + bx_i^c}\right) - n$$

- Algebra again cannot be worked out as in the Factor CER Case, because there are no "closed" Expressions for a, b, and c

- Again, though, Percentage Bias will turn out to be Zero in any specific case (when you make the specific computations)

Triad CER Successive Iterative Parameter Solutions

Initial Guesses		IRLS Sol.	#0	#1	#2	#3	#4	#5	#6	#7
a =	-28.484	-28.48426	-10.000	-47.995	-25.775	-29.567	-28.102	-28.486	-28.484	-28.48426
b =	13.507	13.50731	10.000	29.126	11.624	14.289	13.234	13.506	13.508	13.50731
c =	0.404	0.40369	0.500	0.266	0.433	0.393	0.408	0.404	0.404	0.40369

Note: Iterative Process Converges to Solution in 7 Steps, Starting from Initial "Guess" #0.

Diameter-vs.-Cost Data Base, with IRLS Triad CER Statistics

Antenna Cost vs. Reflector Diameter Data			IRLS Triad CER and Quality Metrics					
n	Diameter (x) (Feet)	Cost (y) FY99$K	c	b	a	ESTy = a+bx^c	%SE	%BIAS
7	9.8	3.300	0.404	13.507	-28.484	5.456	0.156	0.395
7	7.9	3.595	0.404	13.507	-28.484	2.628	0.135	-0.368
7	8.2	1.900	0.404	13.507	-28.484	3.099	0.150	0.387
7	11.5	10.900	0.404	13.507	-28.484	7.720	0.170	-0.412
7	16.4	15.434	0.404	13.507	-28.484	13.297	0.026	-0.161
7	19.7	16.074	0.404	13.507	-28.484	16.507	0.001	0.026
7	23.6	17.274	0.404	13.507	-28.484	19.910	0.018	0.132
Sums	97.1	68.477				Totals:	0.655	0.000

EST = Estimated
FY = Fiscal Year
SE = Squared Error

Example: Diameter vs. Cost Triad CER

- Based on Computations using the Historical Data …

$$a = -28.484; \quad b = 13.507; \quad c = 0.404$$

- Multiplicative-Error CER

$$y = -28.484 + 13.507x^{0.404}$$

- Standard Error of the Estimate (%SEE)

$$Standard\ Error = \sqrt{\frac{1}{n-3}\sum_{i=1}^{n}\left(\frac{y_i - a - bx_i^c}{a + bx_i^c}\right)^2}$$

$$= \sqrt{\frac{1}{7-3}(0.655)} = 0.405$$

(Average 40.5% Across Data Range)

Calculation of R^2 Quality Metric for IRLS Triad CER

Antenna Cost vs. Reflector Diameter Data

n	ESTy (x) FY99$K	Cost (y) FY99$K	x2	y2	xy
7	5.456	3.300	29.766	10.890	18.004
	2.628	3.595	6.905	12.924	9.447
	3.099	1.900	9.606	3.610	5.889
	7.720	10.900	59.596	118.810	84.146
	13.297	15.434	176.817	238.208	205.230
	16.507	16.074	272.476	258.373	265.331
	19.910	17.274	396.409	298.391	343.926
Sums	**68.617**	**68.477**	**951.575**	**941.207**	**931.972**

Num R = 1825.126
Den1 R2 = 1952.741
Den2 R2 = 1899.349
R2 = 0.898

R^2 Between Actuals and Estimates

ESTy (x)	Cost (y)
5.456	3.300
2.628	3.595
3.099	1.900
7.720	10.900
13.297	15.434
16.507	16.074
19.910	17.274

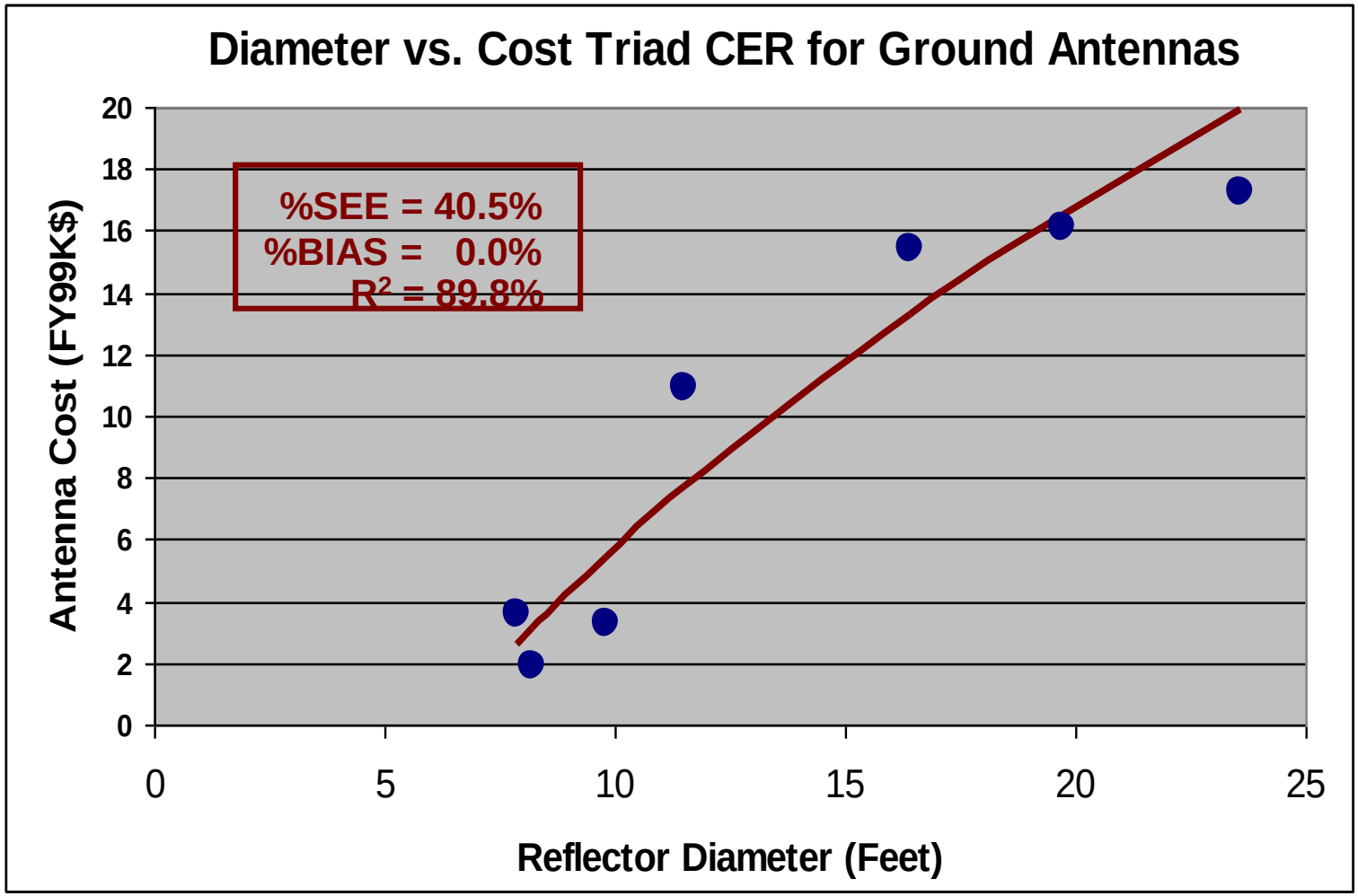
- $$R^2 = \frac{\left[n\sum_{k=1}^{n} x_k y_k - \sum_{k=1}^{n} x_k \sum_{k=1}^{n} y_k \right]^2}{\left[n\sum_{k=1}^{n} x_k^{\,2} - \left(\sum_{k=1}^{n} x_k \right)^2 \right] \left[n\sum_{k=1}^{n} y_k^{\,2} - \left(\sum_{k=1}^{n} y_k \right)^2 \right]}$$

$$= \frac{(1825.126)^2}{[1952.741][1899.349]} = 0.898$$

$$= 89.8\%$$

- Therefore the R^2 Quality Metric for this CER is 89.8% (Perfect Fit is 100%)

Triad CER and Its Quality Metrics (Compare with Data Base on Earlier Chart)



Contents

- CER Development
  - Ordinary Least Squares (OLS)
  - Log-transformed OLS
    - Statistical Issues
    - Quality Metrics
- General-Error Regression
  - Iteratively Reweighted Least Squares (IRLS)
    - Factor CERs
    - Linear CERs
    - Triad CERs
  - Quality Metrics
  - Zero Percentage Bias, Minimum Percentage Error (ZMPE)
- Summary

Zero-Percentage-Bias, Minimum-Percentage-Error CERs

- In 1998, the "Zero Percentage Bias – Minimum Percentage Error" (ZMPE) Technique was Proposed* to Yield CERs Guaranteed to Have Minimum Possible Percentage Error among all Unbiased CERs for a Given Data Set that Have the Functional Form being Considered

- ZMPE Pursues the Minimum-Percentage-Error Goal Directly
  - Computes Minimum-Percentage-Error CER, Subject to Constraint that Percentage Bias be Exactly Zero
  - CERs Derived Using "Constrained Optimization" – Another Capability of *Excel Solver*
  - But, caution must be taken when using Excel

*** Book, S. and Lao, N, "Minimum-Percentage-Error Regression under Zero-Bias Constraints", *Proceedings of the Fourth Annual U.S. Army Conference on Applied Statistics, 21-23 October 1998*, U.S. Army Research Laboratory, Report No. ARL-SR-84, November 1999, pages 47-56.**

ZMPE Mathematics

Using the Triad Case as an Example, ZMPE Minimizes

$$F(a,b,c) = \sum_{k=1}^{n} \left(\frac{y_k - a - bx_k^c}{a + bx_k^c} \right)^2,$$

Subject to the Constraint

$$\%Bias(a,b,c) = \sum_{k=1}^{n} \left(\frac{a + bx_k^c - y_k}{a + bx_k^c} \right) = 0$$

ZMPE Diameter vs. Cost Triad CER

- Based on Computations on the Historical Data …

$$a = -236.11; \quad b = 212.42; \quad c = 0.06$$

- Multiplicative-Error CER

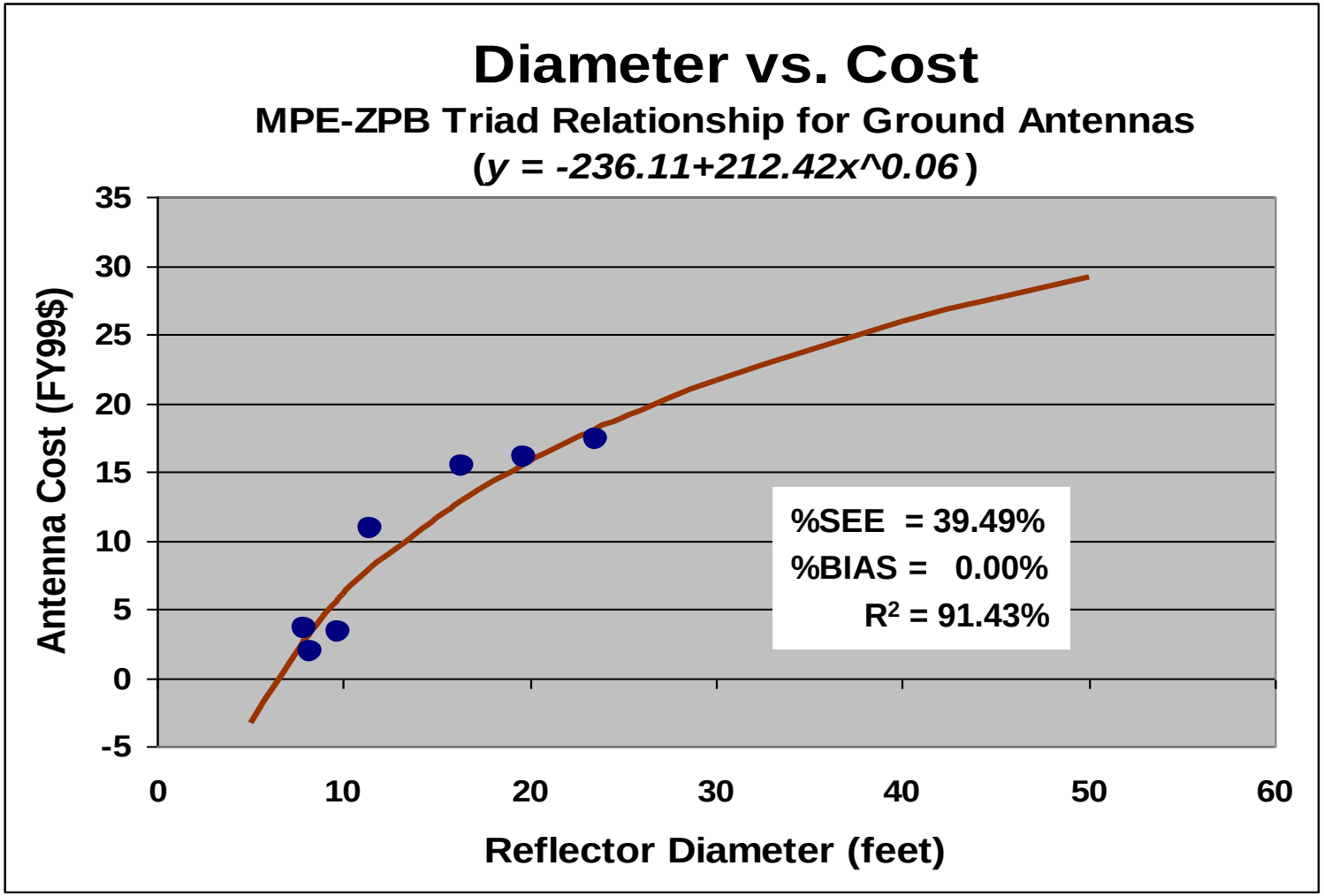
$$y = -236.11 + 212.42x^{0.06}$$

- Standard Error of the Estimate (%SEE)

$$\text{Standard Error} = \sqrt{\frac{1}{n-3}\sum_{i=1}^{n}\left(\frac{y_i - a - bx_i^c}{a + bx_i^c}\right)^2} = \sqrt{\frac{1}{7-3}(0.6236)} = 0.3949$$

("Average" 39.49% Across Data Range)

ZMPE Triad CER, Quality Metrics Superimposed on Data Base



General Comments

- Note that the ZMPE CER's Quality Metrics are "Better" than are the IRLS CER's Quality Metrics for Data Set of the Example – This is a General Phenomenon
- A Presentation by Dr. S.A. Book to the 2006 ISPA International Conference in Bellevue WA (2006) Showed that ZMPE and MUPE CERs Derived from the Same Data Set are Different and that the ZMPE CER has Smaller Percentage Standard Error than does the IRLS/MUPE CER
- A Monograph (2003) by Dr. M.S. Goldberg and A.E. Tuow, then of IDA, Pointed Out that IRLS (aka "MUPE") CERs are Actually not Unbiased When Based on Small Samples
- A Paper (1998) by Dr. S.A. Book and N.Y. Lao Recommended the Use of Constrained Optimization to Minimize Percentage Standard Error, Subject to the Constraint that the CER's Percentage Sample Bias is Zero

Multiplicative-Error CER Facts

- Minimum-Percentage-Error (MPE)
  - Multiplicative-Error CERs of Arbitrary Functional Form
  - CERs that Minimize Percentage Error of the Estimate
  - CERs that are Biased High
  - CERs with "Heuristic" Statistical Properties*
- Iteratively Reweighted Least Squares (IRLS)
  - Multiplicative-Error CERs of Arbitrary Functional Form
  - CERs that are Unbiased
  - CERs with "Good" Statistical Properties*
  - CERs that Maximize the "Quasi-Likelihood"
- Zero-Percentage-Bias/Minimum-Percentage-Error (ZMPE) CERs
  - Multiplicative-Error CERs of Arbitrary Functional Form
  - CERs that Minimize Percentage Error of the Estimate Subject to being Unbiased
  - CERs that are Unbiased
  - CERs with "Heuristic" Statistical Properties*

*** The term "statistical properties" refers to hypothesis tests, confidence intervals, and their associated accoutrements such as t and F values.**

Contents

Summary

- CERs are Derived by Applying Statistical Analysis to Cost Data Bases Reflecting Historical Cost Experience
  - Multiplicative-Error Regression Frequently More Appropriate than Additive-Error Regression for CERs
  - IRLS (aka “MUPE”) and ZMPE Allow CERs of All Appropriate Forms to be Derived
- CER Quality Metrics Support Credibility of Estimates Derived from Multiplicative-Error CERs
  - Percentage Standard Error of the Estimate
  - Percentage Bias
  - Pearson’s Correlation Squared between Estimates and Actuals

References

Covert, R. and Anderson,T., Modern Techniques of Regression for CERs with Multiplicative Errors, 2012 SCEA/ISPA Joint Annual Conference & Training Workshop Orlando, FL 26-29 June 2012

Book, S. and Lao, N, "Minimum-Percentage-Error Regression under Zero-Bias Constraints", *Proceedings of the Fourth Annual U.S. Army Conference on Applied Statistics, 21-23 October 1998*, U.S. Army Research Laboratory, Report No. ARL-SR-84, November 1999, pages 47-56.

Book, S., "IRLS/MUPE CERs Are Not MPE-ZPB CERs," International Society of Parametric Analysts, 28th Annual Conference, Seattle WA, 23-26 May 2006.

Goldberg, Matthew S. and Tuow, Anduin E., "Statistical Considerations in Estimating Learning Curves and Multiplicative CERs," 32nd Annual DoD Cost Analysis Symposium, Williamsburg VA, 2-5 February 1999, 44 charts.

Goldberg, Matthew S. and Touw, Anduin E., *Statistical Methods for Learning Curves and Cost Analysis*, Institute for Operations Research and the Management Sciences (INFORMS) Topics in Operations Research Series, 2003, 196 pages.

Nelder, J. A., "Weighted Regression, Quantal Response Data, and Inverse Polynomials," *Biometrics*, Vol. 24 (1968), pages 979-985.

Wedderburn, R.W.M., "Quasi-likelihood Functions, Generalized Linear Models, and the Gauss-Newton Method," *Biometrika*, Vol. 61, Number 3 (1974), pages 439-447.

United States Air Force Space and Missile Systems Center, *Unmanned Space Vehicle Cost Model, Sixth Edition* (USCM6), 1988.