

New Research in General Error Regression Model (GERM) Significance Testing

K. Cincotta

Andrew Busick

*Presented at the Society of Cost Estimating and
Analysis (SCEA) Conference*

June 8-11, 2010

San Diego, CA



Acknowledgments

- Dr. Stephen Book, for inspiring the research and for critical feedback
- Tim Anderson, for further inspiration and for sharing the original GERM data set
- Technomics, for sponsorship of research

Quotable

“It would be so nice if something made sense for a change.”

- Lewis Carroll. *Alice's Adventures in Wonderland* (1865)

Outline

- Background
- The Problem
- The SIG Solution
- The Bias Adjuster Solution
- The Bootstrapping Solution
- Comparison of Results
- Conclusions
- Ideas for Further Research

Background: GERM

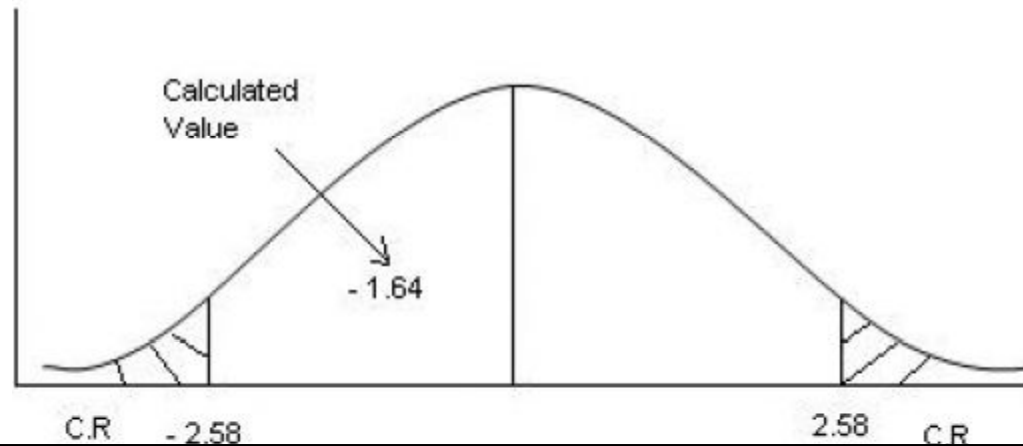
- **General Error Regression Models (GERM)** are regression models in which no particular assumption is made about the distribution of the error term
- However, we still need to specify whether additive or multiplicative error
- We also need to specify a *penalty function*, e.g.:
 - Sum of squared errors (weighted or unweighted)
 - Sum of squared percent errors
 - Average absolute percent error with zero bias constraint, etc.

Background: Significance Testing

- **Significance testing** refers to a set of methods designed to determine whether results have occurred by chance¹
- In classical significance testing, the p-value is the probability that the *true* value of the parameter being estimated is zero
- Significance testing *adds value* to a cost estimating relationship (CER):
 - Allows for a process of nullification, wherein insignificant variables can be dropped
 - Increases degrees of freedom/explanatory power of regression
 - May remove bias in estimated coefficients of significant variables
 - Leads decision-makers down the “right path” when making cost-design tradeoffs:
 - The relationship between cost and an insignificant variable, by definition, cannot be expected to hold in the future
 - True cost drivers are statistically significant and can be managed as such

Significance Testing: Example

T Statistics

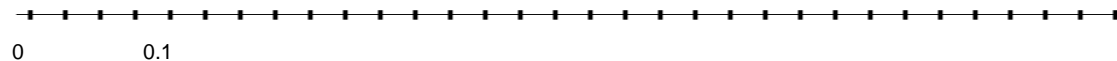


Significant

Insignificant

Significant

P Values (assuming 90% significance test)



Significant

Insignificant

The Problem

- Traditional significance testing is not possible in GERM
 - We want a p-value
 - Need a t-stat to get a p-value
 - $T \text{ stat} = \text{parameter estimate} / \text{standard error}$
 - We don't know the standard error of the parameter estimate
- To remedy this, one author ² proposes the “SIG Test”
- We will explore this, and two other methods

The SIG Test (Multiplicative Error Case)

- Run the full model to generate estimated coefficients
- Set all inputs to their means; record the CER output and its standard error (SE) or standard percent error (SPE)
- “Nullify” a variable; record the CER output & SPE

$$\text{SIG}_{\text{mean}} = \% \text{ change in CER output} = \frac{[f(x)_{\text{reduced}} - f(x)_{\text{full}}]}{f(x)_{\text{full}}}$$

$$\text{SIG}_{\text{SPE}} = \% \text{ change in CER SPE} = \frac{(\text{SPE}_{\text{reduced}} - \text{SPE}_{\text{full}})}{\text{SPE}_{\text{full}}}$$

$$\text{SIG}_{\text{total}} = |\text{SIG}_{\text{mean}}| + \text{SIG}_{\text{SPE}}$$

As a starting point, the parameter is deemed “significant” if $\text{SIG}_{\text{total}} > .1$

Inconvenient Truths: SIG_{mean}

- $SIG_{\text{mean}} = 0$ in all cases in a linear (OLS) framework
 - OLS predictions always pass through (μ_x, μ_y)
 - Implies that if your GERM model happens to have a linear or nearly linear relationship, “significance” cannot be tested using SIG_{mean}
 - But a *general* error method should work in the *general* case, a minimum criterion for which is working in this specific case
- SIG_{mean} isn't related to CER performance and therefore is not measuring the same thing as a t-test
 - It measures movement in the output of the CER, but these outputs are never linked to the dependent variable. What if the CER “moves” in the wrong direction?

Inconvenient Truths: SIG_{SPE}

- SIG_{SPE} does measure CER performance, because it relies upon comparison of standard percent errors, with and without the variable in question
- Unfortunately, it does not accurately replicate significance test results in the (testable) OLS case
- It is not clear how to handle negative values:
 - “Note that there should be no need to take the absolute value of SIG_{SE} or SIG_{SPE} . If either...were negative, it would mean that the variance of the reduced CER was less than the variance of the full CER. However, this case should not happen if the optimization is done correctly. If a lower variance solution of the CER were available...then the optimization should have found that solution.” (Anderson (2009))
 - Yet we have found numerous examples where SIG_{SPE} is legitimately negative (degrees of freedom change). This has been noticed before ³
 - Adding absolute value bars does not fix the problem, because then highly insignificant variables (ones with $SIG_{SPE} < 0$) get rewarded!

Inconvenient Truths: SIG_{total}

- $SIG_{total} = |SIG_{mean}| + SIG_{SPE}$
 - Each component has its malcontents
 - We're adding a “how much does the mean move” metric to a CER performance metric
 - Percentages are being summed where the denominators are different. This has been previously called “innumerate,” and in so doing, one is said to fall victim to Simpson's Paradox ⁴
 - We do not feel so harshly about the method; in fact, it can be quite useful. But is there something better?

The “Bias Adjuster” Method

- Consider the GERM equation

$$Y = a + bX^cW^dQ^{efType}$$

- Regress/optimize as usual
- Recognize that a is simply a bias adjuster, and would not be needed if the zero-intercept model were unbiased
- Subtract a from both sides and take logs:
$$\ln(Y-a) = \ln(b) + c\ln(X) + d\ln(W) + e\ln(Q) + Type \ln(f)$$
- This gives an unbiased estimate of $\ln(Y-a)$
- This equation is linear, so p-values can be calculated
- Use these p-values as proxies for p-values associated with the original fit parameters

The Bootstrapping Method

- Recall the original problem: we don't have reliable estimates of the standard errors around regression fit parameters derived in a GERM framework
- *“Bootstrapping is the practice of estimating properties of an estimator (such as its variance) by measuring those properties when sampling from an approximating distribution... It may also be used for constructing hypothesis tests. It is often used as an alternative to inference based on parametric assumptions when ...parametric inference is impossible.”*⁵
- This is exactly the problem at hand: if we could estimate the standard error of a fit parameter, we could also estimate its significance using standard t-tests/p-value analysis
- Bootstrapping as a technique to generate confidence intervals and significance is *not* new to the cost community⁶

5. [http://en.wikipedia.org/wiki/Bootstrapping_\(statistics\)](http://en.wikipedia.org/wiki/Bootstrapping_(statistics))

6. Book, Dr. Stephen A. *Prediction Bounds for General Error Regression CERs*. 39th Annual DoDCAS: Williamsburg, VA (2006)

The Bootstrapping Method Applied to GERM

- Perform the GERM regression/optimization as usual
- Generate n error terms
- Assume that the data are independent of their error terms
 - This does not violate GERM because we still have no assumption about the *distribution* of those error terms
- This gives a population of n^n possible combinations of data with errors to sample from
- Sample, with replacement, from this population to create 30 new data sets of 30 points each
- Estimate each coefficient 30 times
- The standard deviation of the 30 estimates approximates the standard error of the estimated coefficient
- Approximate t stat =
$$\text{average}(\text{estimated coefficients}) / \text{stdev}(\text{estimated coefficients})$$
- This is a high-level summary; much more details provided elsewhere in this conference ⁷

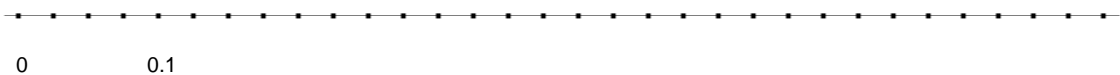
7. Feldman, Daniel I. *Testing for the Significance of Cost Drivers Using Bootstrap Sampling*. ISPA/SCEA Joint Conference: San Diego, CA (2010)

Comparison of Results: Ground Rules

- Assess all three methods in two cases:
 - An additive error, linear model in which OLS assumptions hold (but are not *assumed* to hold when performing the regression) that has exact parameter values, and exact p-values associated with each parameter
 - A multiplicative error, GERM model in which no particular assumptions hold (except that the error terms and data are uncorrelated), which also has exact parameter values, but no exact p-values
- Assess how faithfully each method replicates actual significance testing in both cases
- Because the SIG method does not produce a p-value, assume a 90% significance level and use estimated p-value =

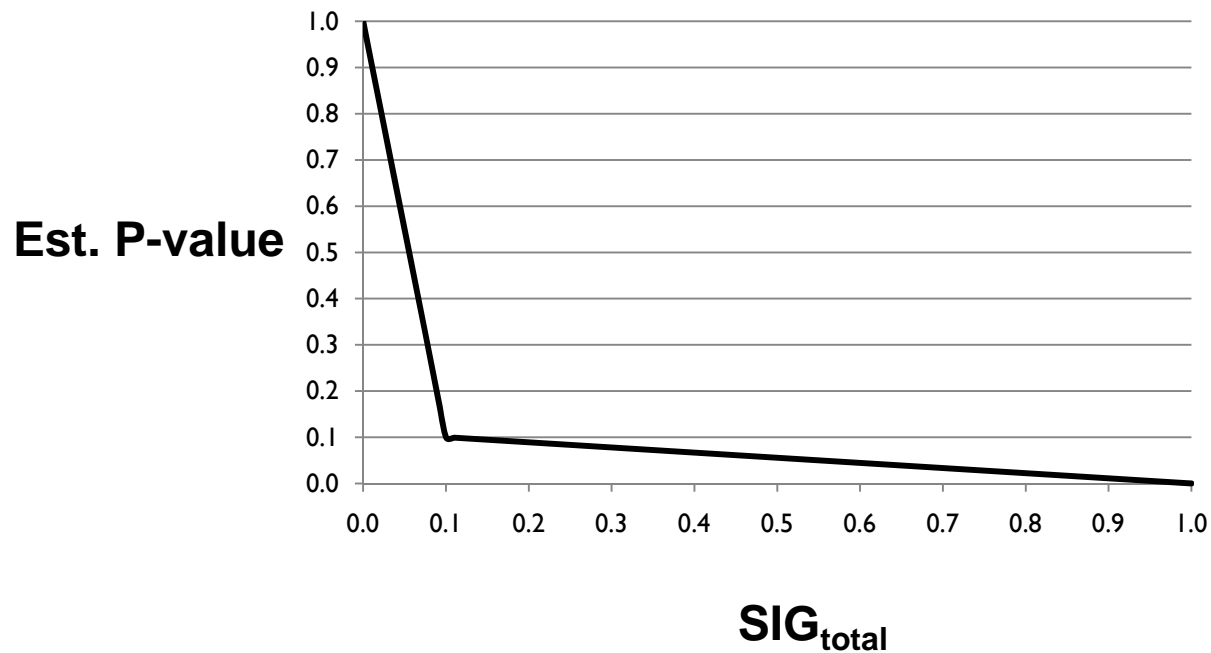
$$0.1 - (\text{SIG}_{\text{total}} - 0.1)/0.9 \text{ for } \text{SIG}_{\text{total}} \geq 0.1 \text{ (significant case)}$$

$$0.1 + (0.1 - \text{SIG}_{\text{total}})/0.1 \text{ for } \text{SIG}_{\text{total}} \leq 0.1 \text{ (insignificant case)}$$



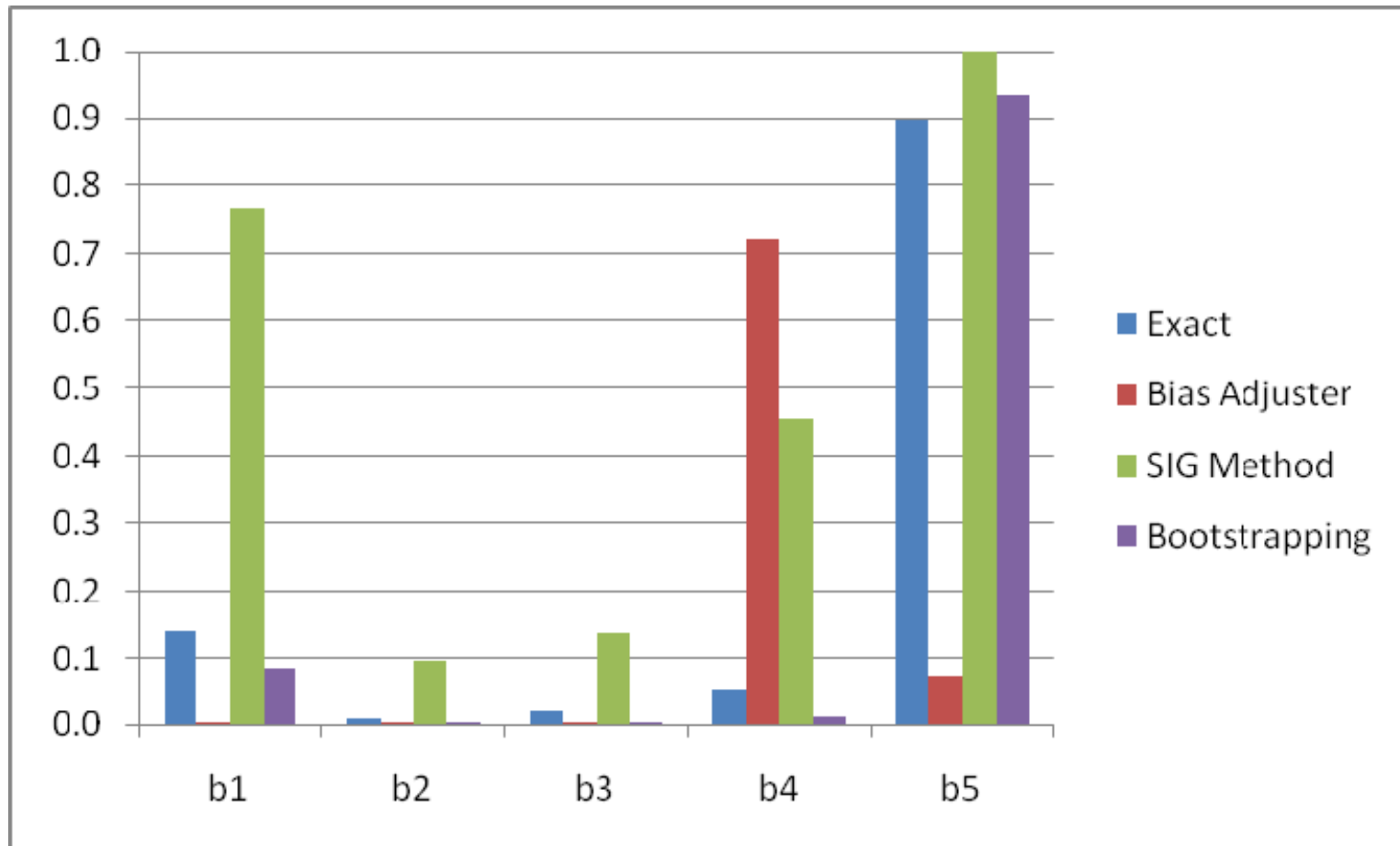
Significant	Insignificant
$\text{SIG}_{\text{total}} > .1$	$\text{SIG}_{\text{total}} < .1$
Est. p value < .1	Est. p value > .1

Graphical Depiction of Mapping of SIG_{total} to Estimated P-Values

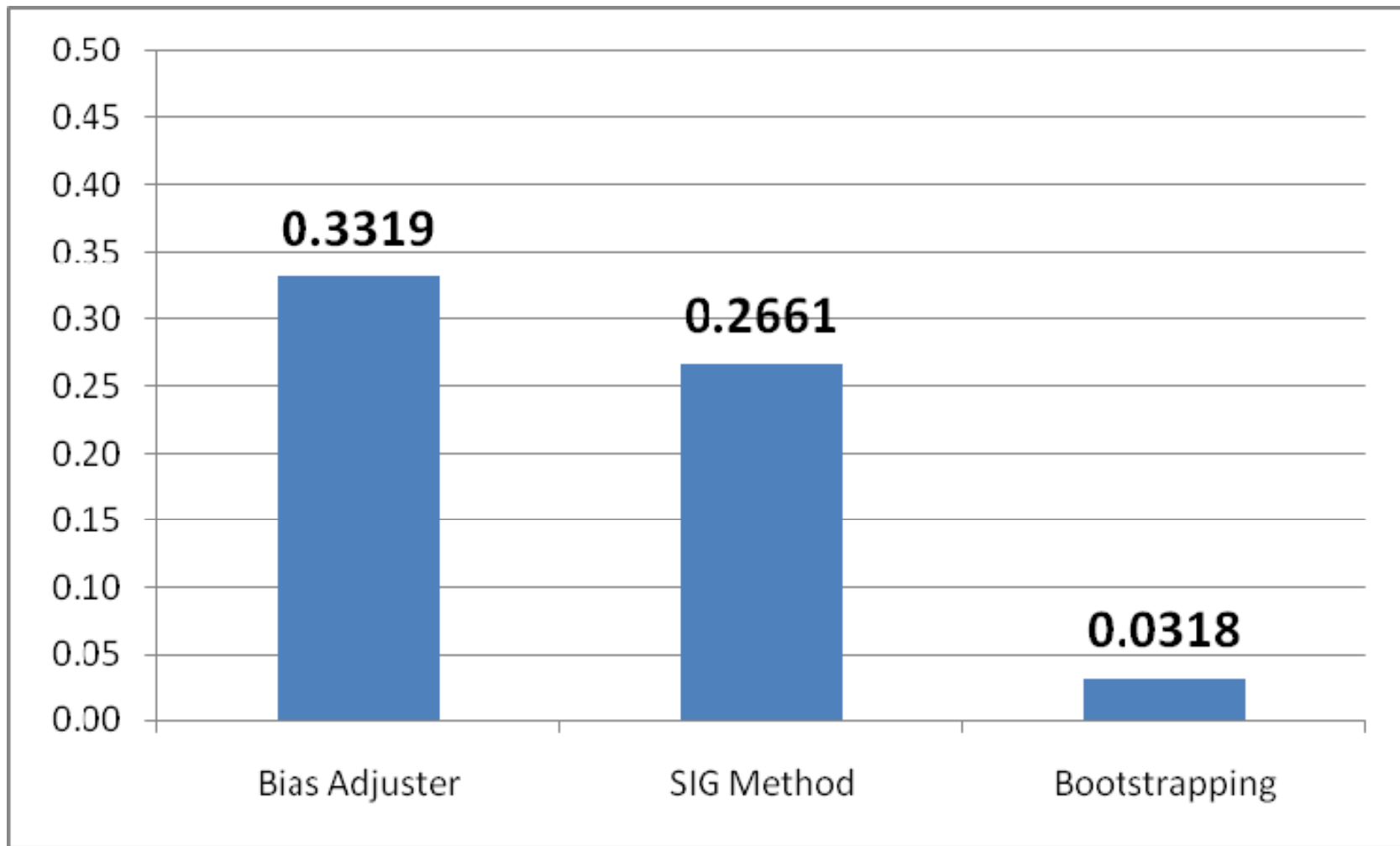


Comparison of Results: P-Values (OLS Case)

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + \varepsilon$$



Comparison of Results: Average Absolute Error in P-Value (OLS Case)

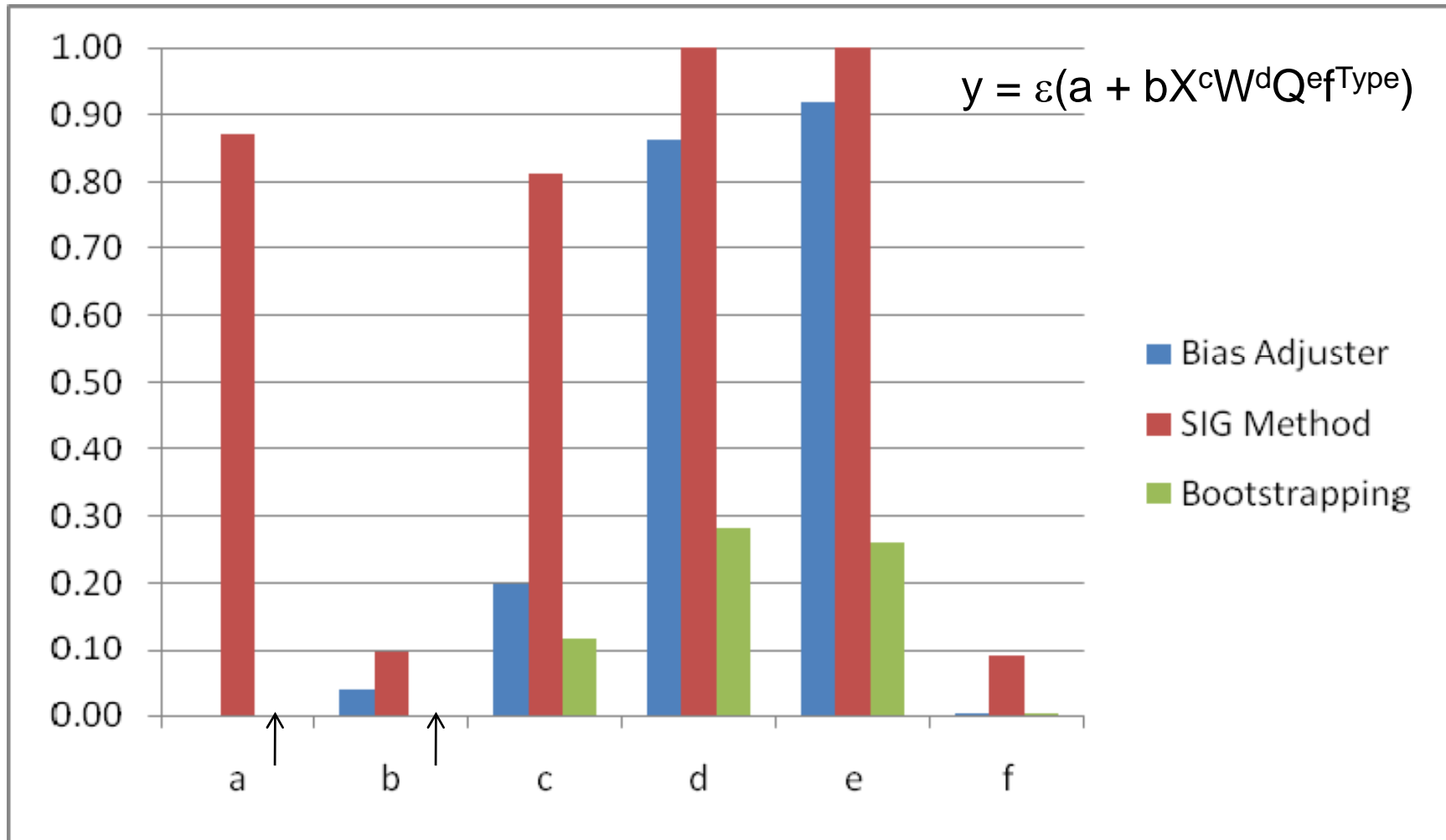


Comparison of Results: Conclusions as to Significance (OLS Case)

	b1	b2	b3	b4	b5	# correct
Exact	Not Significant	Significant	Significant	Significant	Not Significant	–
Bias Adj	Significant	Significant	Significant	Not Significant	Significant	2
SIG Method	Not Significant	Significant	Not Significant	Not Significant	Not Significant	3
Bootstrapping	Significant	Significant	Significant	Significant	Not Significant	4

Note: When the data were originally generated, only b5 was set to zero (i.e. b1 is nonzero). When compared to *those* “exact” values, the Bootstrapping method draws the correct inference in all five cases, because it draws the correct inference about b1.

Comparison of Results: P-Values (GERM with Multiplicative Error)

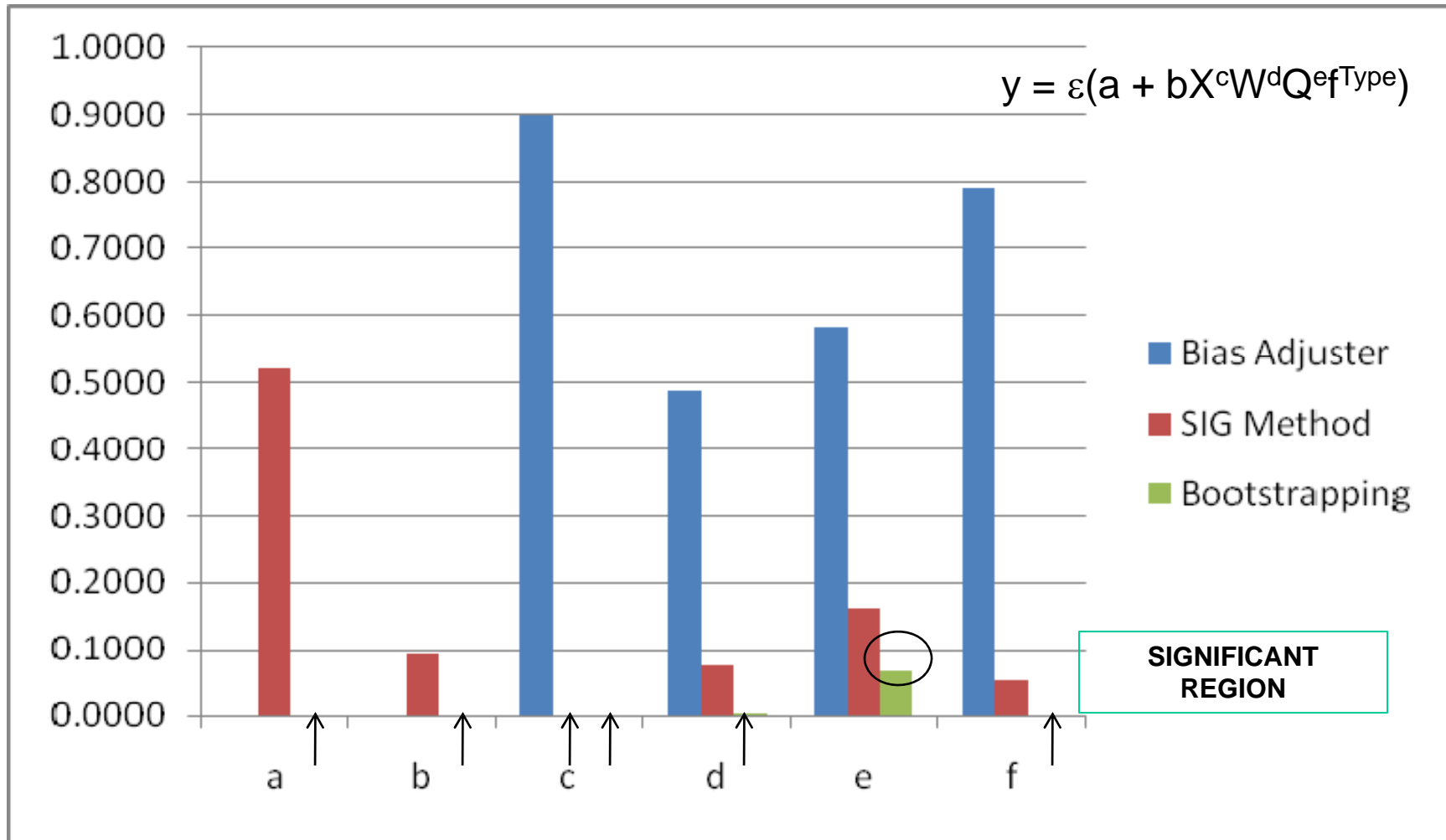


Note: When the data were generated, only “e” was set to zero. So nominally, **lower is better in all cases, except with “e,” where higher is better.** There are bars for all methods in all columns, except Bias Adjuster for “a.”

Comparison of Results (GERM with Multiplicative Error): Which Method Comes “Closest to Reality”?

Parameter	"True" P Value	Bias Adjuster	SIG Method	Bootstrapping
a	0.0000	--		X
b	0.0000			X
c	0.0000			X
d	0.0000			X
e	1.0000		X	
f	0.0000			X

Comparison of P-Value Results: Applied to Original “SIG Method” Data Set



Note: The SIG method paper's *example* of an insignificant parameter (e) tests as *significant* under Bootstrapping.

Analysis of What Happened

- Fits are almost always tighter in log space, so Bias Adjuster tends to *systematically overestimate* significance and *underestimate* p-values
- Removing a variable (even a significant one) rarely “changes things” by 10%, so the SIG test tends to systematically underestimate significance and overestimate p-values
 - Overestimation of p-values is compounded by “steep slope” of our curve mapping SIG scores to p values in insignificant region, and “flat slope” in significant region
- Bootstrapping is the “goldilocks” solution
 - Not too high, not too low: just right
 - Does not appear to contain systematic bias

Conclusions

- In the (testable) OLS case, bootstrapping does a far better job of approximating “true” significance than the “Bias Adjuster” and “SIG” methods
 - Much lower average absolute error
 - Draws correct inference as to significance more often
- While an exact test in the GERM case is not possible, Bootstrapping results show promise
 - Comes “closest to reality” in 5 of 6 cases
 - Draws different conclusions when applied to “SIG Method” data
- The SIG Method does incrementally better than the Bias Adjuster Method, but Bootstrapping provides superior results

Ideas for Future Research

- Better method of converting SIG_{total} to estimated p-values using SIG Method
- Method of deriving estimated p-values for additive intercept term in Bias Adjuster Method
- Investigate whether study findings are replicated with:
 - Other data sets
 - Cases where regressors are mostly *insignificant*
 - Using other calibration points (e.g. SIG_{median} , SIG_{80th} instead of SIG_{mean})
- User-friendly Excel routines that automate the Bootstrapping process (we have a great start)
- Other discussion

References

- Anderson, Tim. *A Distribution-Free Measure of the Significance of CER Regression Fit Parameters Established Using General Error Regression Methods*. *Journal of Cost Analysis and Parametrics*. Volume 2, Issue 1 (Summer, 2009)
- Book, Dr. Stephen A.
 - *Modern Techniques of Multiplicative Error Regression*. IPSA/SCEA Training Workshop: St. Louis, MO (2009).
 - *Prediction Bounds for General Error Regression CERs*. 39th Annual DoDCAS: Williamsburg, VA (2006)
 - and Lao, Norman. *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), pp. 47–56.
 - and Young, Phillip. *General Error Regression for Developing Cost Estimating Relationships*. *The Journal of Cost Analysis* (Fall, 1997)
- Davison, A. C. and Hinkley, D. *Bootstrap Methods and their Application*. 8th ed. Cambridge, MA: Cambridge Series in Statistical and Probabilistic Mathematics (2006)
- Feldman, Daniel I. *Testing for the Significance of Cost Drivers Using Bootstrap Sampling*. ISPA/SCEA Joint Conference: San Diego, CA (2010)
- Hulkower, Neal. *Numeracy for Cost Analysts: Doing the Right Math, Getting the Math Right*. MCR: McLean, VA (2008)
- [http://en.wikipedia.org/wiki/Bootstrapping_\(statistics\)](http://en.wikipedia.org/wiki/Bootstrapping_(statistics))

Backup: OLS Model Used

- $y = a + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + b_5x_5 + \varepsilon$
- $\varepsilon \sim N(0,500)$
- All x 's $\sim U(0,100)$

Fit Parameter	Exact Value	Exact P Value
a	5	0.6860
b1	6	0.1408
b2	7	0.0096
b3	8	0.0218
b4	9	0.0535
b5	0	0.8985

Backup: GERM Model Used

- $y = a + \varepsilon(bX^cW^dQ^{efType})$
- Independent variables distributed as follows:

Variable	Distribution	Mean	StDev
X	Normal	108	10.8
W	Normal	12	1.2
Q	Normal	9.9	0.99
Type	Bernoulli	0.5	0.5

- ε distributed lognormally as follows:

Parameter	Log Space	Unit Space Pop	Unit Space Sample
Mean	0	1.0560	1.0777
Variance	0.1089	0.1283	0.1155
StDev	0.33	0.3582	0.3399
Median	0	1	1.0180