How Regression Methods Impact Uncertainty Results

Dr. B. Jonov, Dr. S.P. Hu, A. Smith

21 June 2016
Tecolote Research PRT-210

Overview

- Background – error models and regressions (LOLS,MUPE,ZMPE)

- Comparison of LOLS, MUPE, and ZMPE regressions
  - Their strengths and weaknesses
  - Concerns about LOLS and the response to criticism

- Regressions and Uncertainty
  - LOLS uncertainty can be sound and justified
  - ZMPE does not have established uncertainty assignment process
  - Suggest a systematic approach to assign uncertainty to ZMPE CERs

- Examples and observations
  - The three regressions on the same data set
  - Similar point estimates but different uncertainty results
  - Which regression provides reliable uncertainty results?

Background

- Common notation and definitions
- Additive vs Multiplicative Error Models
- LOLS, MUPE, and ZMPE regressions

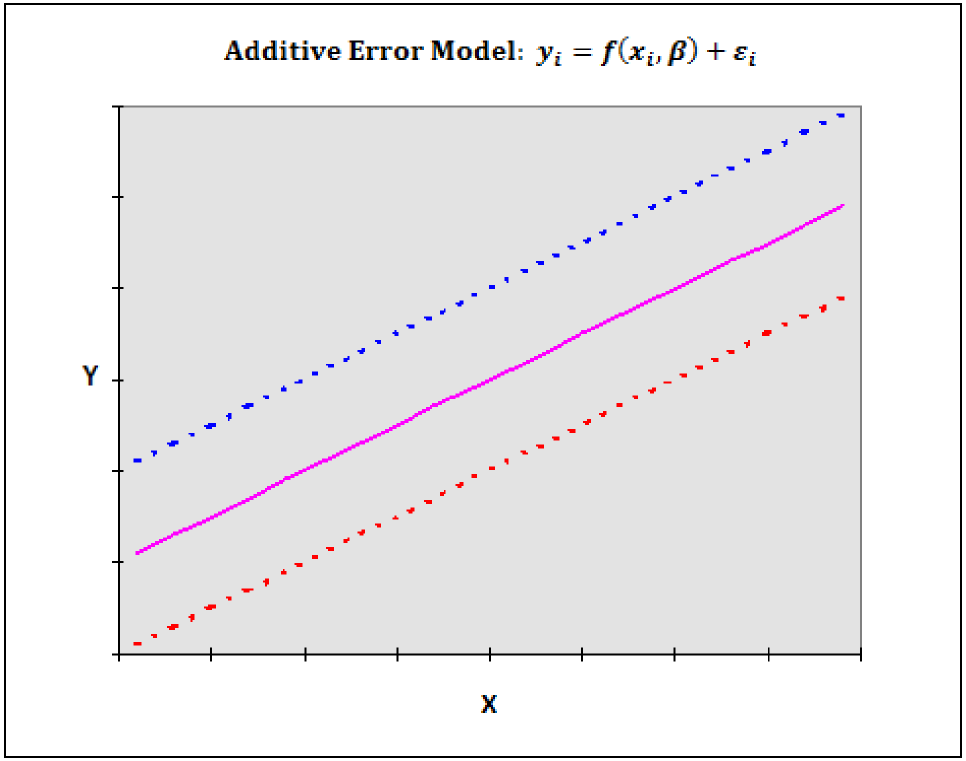
Notation & Definitions

- Given data set $(x_i, y_i)_{i=1}^{n}$
  - x_i - cost drivers, y_i - value of dependent variable (cost)
- Hypothetical equation $y = f(x, \beta)$
  - $\beta = (\beta_1, \dots, \beta_p)$ - unknown parameters
- Regression Results
  - $\hat{\beta} = (\hat{\beta}_1, \dots, \hat{\beta}_p)$ - regression estimates of the parameters
  - $\hat{y} = f(x, \hat{\beta})$ - predicted cost
- Goodness of fit measures (additive and multiplicative error models)
  - $SEE = \sqrt{\sum_{i=1}^{n} \frac{1}{n-p} (y_i - \hat{y}_i)^2}$ and $SPE = \sqrt{\sum_{i=1}^{n} \frac{1}{n-p} \left(\frac{y_i - \hat{y}_i}{\hat{y}_i}\right)^2}$

Additive Error Model $y_i = f(x_i, \beta) + \varepsilon_i$

- ε_i is the error of the cost at the i^{th} data point
- Error assumptions: mean 0 and variance σ^2
- Error is constant throughout the entire data range
- Minimize sum of squared errors

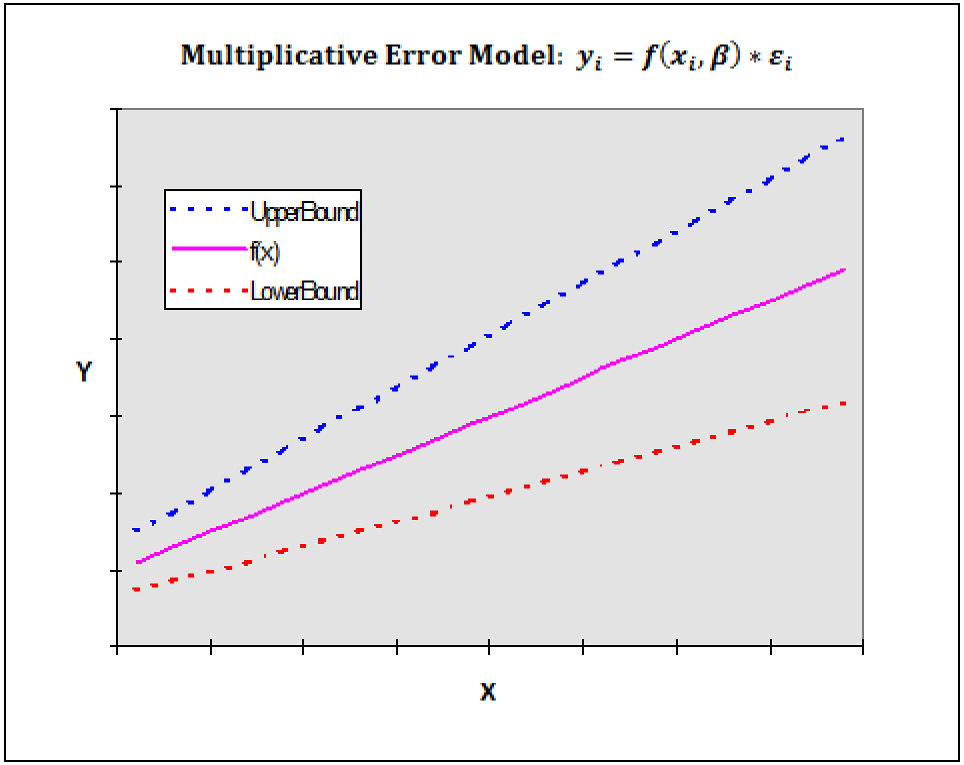
$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - f(x_i, \beta)\right)^2$$



Multiplicative Error Model $y_i = f(x_i, \beta) * \varepsilon_i$

- ε_i is the error of the cost at the i^{th} data point
- Error assumptions: mean 1 and variance σ^2 (MUPE & ZMPE)
- Error is proportional to magnitude of the equation
- Minimize sum of squared percent errors

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



LOLS Regression

- Log-linear multiplicative error model:

$$y_i = ax_i^b * \varepsilon_i \quad \text{with} \quad \varepsilon_i \sim LN(0, \sigma^2)$$

- Take the natural log: $\ln(y_i) = \ln(a) + b\ln(x_i) + \ln(\varepsilon_i)$

- This is now a linear additive model… $Y = A + B * X + E$

- Can use OLS regression to minimize $\sum(\ln \varepsilon_i)^2$

- Transform result back to unit space by taking exponents

MUPE Regression

- MUPE is iterative regression technique

- At k^{th} iteration, solve for β_k that minimizes :

$$\sum_{i=1}^{n} \left(\frac{y_i - f(x_i, \beta_k)}{f(x_i, \hat{\beta}_{k-1})} \right)^2$$

$\hat{\beta}_{k-1}$ is the coefficient solved in previous iterations

- Final solution $\hat{\beta}$ obtained when estimates change in successive iterations is within tolerance limit

ZMPE Regression

Minimize directly

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

Subject to the constraint:

$$\sum_{i=1}^{n}\frac{y_i - f(x_i, \beta)}{f(x_i, \beta)} = 0$$

Comparison of Regression Methods

- LOLS pros and cons
  - LOLS has been subject to criticism and academic concerns.
  - We address those concerns and defend LOLS
- MUPE and ZMPEs pros and cons

LOLS Pros

- **LOLS regression provides analytical solution for the coefficients**
  - OLS can be applied in log-space for log-linear equations
  - Bypasses MUPE and ZMPE issues such as consistency of estimates, dependence on input, method's convergence and stability
  - Linear optimization is less tedious and cumbersome than nonlinear
- **Sound and justified uncertainty assignment**
  - Conditions: log-normally distributed error term
  - PE is the median of log-normal distribution
    - Neither PE location nor distribution shape are known for ZMPE
  - Prediction intervals can be precisely generated
- **Large spectrum of goodness of fit measures**
  - Significance of coefficients can be established
  - Outliers can be detected
  - Model flaws can be exposed
  - ZMPE provides much limited goodness of fit measure.

Response to concerns about LOLS

Concern #1:

Minimizing $\sum(\ln y_i - \ln a - b \ln x_i)^2 = \sum(\ln \varepsilon_i)^2$

is not the same as minimizing $\sum\left(y_i - a x_i^b\right)^2 = \sum e_i^2$

Response to Concern #1:

- LOLS optimization was never intended to minimize $\sum e_i^2$
- LOLS optimizes squared percentage errors, not absolute error (unit space)
- Should not compare fit measures of models with different fit criteria (additive vs multiplicative model)

Response to concerns about LOLS (Cont.)

Concern #2:

The log-space error term $\ln(\varepsilon_i)$ is expressed in meaningless units (log-dollars instead of dollars)

Response to Concern #2:

- even in unit space, ε_i is never measured in dollars for a multiplicative error model;
- the error term ε_i represents the ratio of actual to hypothesized cost
- the error term $\ln(\varepsilon_i)$ does have a meaningful interpretation

$$\ln(\varepsilon_i) \approx \varepsilon_i - 1 = \frac{y_i - f(x_i, \beta)}{f(x_i, \beta)}$$

This is the percentage error term in unit space

Response to concerns about LOLS (Cont.)

Concern #3:

The LOLS process restricts the CER choice to log-linear forms such as $y = ax^b$

Response to Concern #3:

- True that OLS can not be applied to fixed cost equations $y = ax^b + c$ in log-space
- However, the model $y = (ax^b + c) * \varepsilon$ is still solvable in log-space…can use non-linear regression instead of OLS.
- The choice of CER and error model should be driven by technical grounds and logic, not by regression technique preference

Response to concerns about LOLS (Cont.)

Concern #4:

In unit space, the LOLS CER has a non-zero bias:

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

Response to Concern #4:

- The statement is true, but it is not a concern.
- A typical WBS is populated with a mix of point estimate types anyway (mode, mean, median, percentile). There is no compelling reason to convert all point estimates to the mean.
- Multiplicative correction factors (see PING or Goldberg factor) have been developed to remove the bias if necessary
- However, uncertainty assignment does not require adjusted CER results. Recognizing the PE is the median and defining one other PI point is enough to uniquely define the correct uncertainty distribution.

MUPE Pros & Cons

MUPE's Pros

- MUPE's estimator has zero percent bias (no transformations or corrections are applied to the CER result)
- For linear CERs, MUPE provides the best linear unbiased estimates solutions for the parameters
- For nonlinear CERs, MUPE gives consistent estimates for the parameters and mean of the equation
- The parameter estimates are the maximum likelihood estimators (MLE)
- A wider variety of goodness of fit measures than ZMPE (under the normality assumption)
- Statistical tools are available to provide prediction intervals

MUPE's Cons

- The MUPE regression relies on non-linear optimization (can be tedious and cumbersome)
- MUPE's iterative process does not always converge

ZMPE Pros & Cons

ZMPE's Pros

- Unbiased CER result is provided without the need of transformation or adjustment factors
- ZMPE's standard percent error is reported to be smaller than MUPE's SPE
  - May be overstated when considering ZMPE's impact on degrees of freedom
  - See S. Hu, 2015 "Generalized Degrees of Freedom(GDF)," for a discussion on how accounting for degrees of freedom will influence the ZMPE SPE

ZMPE's Cons

- Less reliable solution finding process
  - ZMPE's optimization fails to converge more often (trapped in local minima)
  - Less stable solutions because of sensitivity to starting point input
- Limited goodness of fit measure
  - Only SPE and R^2 are available
  - Insufficient to analyze coefficient significance levels and to detect model flaws (see Anderson, 2009, for heuristic approach).
- No established uncertainty assignment procedure
  - PE location is unknown
  - Distribution shape unknown
- Non-linear regression (tedious)

LOLS, MUPE and ZMPE Uncertainty

LOLS Uncertainty

LOLS CER Uncertainty is given by:

$$\hat{\boldsymbol{y}}_{\mathbf{0}} * \hat{\boldsymbol{\varepsilon}}_{\mathbf{0}} \quad \boldsymbol{where} \quad \hat{\boldsymbol{\varepsilon}}_{\mathbf{0}} \sim \boldsymbol{LN}\left(\mathbf{0}, \boldsymbol{\sigma}^{\mathbf{2}}\left[\mathbf{1} + \boldsymbol{\gamma}^{\mathbf{2}}(\boldsymbol{X}, \boldsymbol{x}_{\mathbf{0}})\right]\right)$$

- $\hat{y}_0$ is the LOLS predictor at $x = x_0$
- $\gamma^2(X, x_0) = \frac{1}{n} + \frac{\left(\ln(x_0) - \overline{\ln(x)}\right)^2}{\sum_{i=1}^{n}\left(\ln(x_i) - \overline{\ln(x)}\right)^2}$
  - Location of $\ln(x_0)$ relative to the mean $\overline{\ln(x)}$ of the cost drivers
- σ is approximated by LOLS' SEE in log-space

LOLS Uncertainty (Cont.)

To derive the uncertainty formula…

- Write the error model in log-space "friendly" format:

$$y_i = e^{\beta_0} x_i^{\beta_1} \varepsilon_i$$

- Take $\ln()$ of each side

$$\ln(y_i) = \beta_0 + \beta_1 \ln(x_i) + \ln(\varepsilon_i)$$

- Now we have a linear additive error model and can apply OLS

LOLS Uncertainty (Cont.)

Key ingredients in deriving LOLS CER uncertainty...

- The log-normal assumption of the error term in the model:

$$y_i = e^{\beta_0} x_i^{\beta_1} \varepsilon_i \quad \text{with} \quad \varepsilon_i \sim LN(0, \sigma^2)$$

- The coefficient estimates by OLS in log-space

$$\hat{\beta} = (X^T X)^{-1} X^T \ln(Y)$$

where

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & \ln(x_1) \\ \vdots & \vdots \\ 1 & \ln(x_n) \end{pmatrix}, \qquad \ln(Y) = \begin{pmatrix} \ln(y_1) \\ \vdots \\ \ln(y_n) \end{pmatrix}$$

MUPE & ZMPE Uncertainty

- MUPE CER uncertainty
  - It is analytical in nature
  - But it is an approximation process…relies on Taylor series linearization
  - Not a closed-form formula like LOLS uncertainty
  - Need statistical tools to obtain prediction intervals

- ZMPE CER Uncertainty
  - No established uncertainty assignment process
  - Shape of the uncertainty distribution is unknown
  - Location if the CER in the distribution is unknown

MUPE & ZMPE Uncertainty (Cont.)

Proposed method to assign CER Uncertainty

$$\hat{y}_0 \; * \text{ Fitted Distribution Curve of } \left\{ \frac{y_1}{\hat{y}_1}, \ldots, \frac{y_n}{\hat{y}_n} \right\}$$

- $\hat{y}_0$ - CER result at $x = x_0$
- Fitted curve that accounts for the location factor of x_0
- Informal but consistent and systematic process
- See S. Hu, 2013 "Fit, Rather Than Assume, a CER Error Distribution" for guidance on how to estimate a prediction interval from a distribution fitted on actual/predicted ratios

Examples

Examples

- Run the three regression methods on the same data sets
- Compare the corresponding point estimates
- Assign uncertainty to the CER result of each regression
- Compare the 80^{th} percentiles of the uncertainty

Example 1

- Generate data using

$$y_i = 0.07x_i^{1.8}\varepsilon_i \text{ where } \varepsilon_i \sim LN(0, \sigma = 0.34)$$

- Generated dataset

Observations	1	2	3	4	5	6	7
X – Cost Driver	7.9	8.2	9.8	11.5	16.4	19.7	23.6
Y – Observed Cost	1.6	3.2	2.3	5.1	7.5	16.3	14.5

Remark: *The error of the cost is log-normally distributed by construction.*

Example 1 (Cont.)

- Regression, point estimate, and uncertainty results

LOLS		MUPE		ZMPE	
$y = 0.038x^{1.936}$	SEE: 2.38	$y = 0.041x^{1.913}$	SEE: 2.386	$y = 0.046x^{1.869}$	SEE: 2.343
	SPE: 0.316		SPE: 0.302		SPE: 0.302
	CV: 0.329		CV: 0.33		CV: 0.324

	LOLS		MUPE		ZMPE	
	PE	80^{th} Ptile	PE	80^{th} Ptile	PE	80^{th} Ptile
$x_0 = 21$	13.8	18.4	14.1(2%)	18.6(1%)	13.8(0%)	21.8(18%)

- Observations
  - ZMPE and LOLS have identical point estimates
  - ZMPE's 80^{th} percentile is substantially different (+18%) from LOLS'

Example 2

- Given dataset

Obs.	1	2	3	4	5	6	7	8	9	10	11	12	13
X	40	50	75	75	75	100	100	240	250	300	550	670	780
Y	10	45	50	70	65	100	90	120	100	80	200	230	300

Remark: *The cost is not assumed to be generated from any specific hypothetical equation and error term distribution*

Example 2 (Cont.)

- Regression, point estimate, and uncertainty results

LOLS		MUPE		ZMPE	
$y = 2.059x^{0.7333}$	SEE: 27.19	$y = 3.047x^{0.67}$	SEE: 27.278	$y = 4.359x^{0.6}$	SEE: 30.19
	SPE: 0.392		SPE: 0.337		SPE: 0.329
	CV: 0.242		CV: 0.243		CV: 0.269

	LOLS		MUPE		ZMPE	
	PE	80^{th} Ptile	PE	80^{th} Ptile	PE	80^{th} Ptile
$x_0 = 500$	196	297	196(-0%)	259(-12%)	181(-7%)	244(-17%)

- Observations
  - MUPE and LOLS have identical point estimates.
  - Both MUPE's and ZMPE's 80^{th} percentiles differ significantly from LOLS' percentile value(by 12% and 17% correspondingly).

Example 2 (Cont.)

- If no evidence of log-normally distributed error, how to choose…
  - CER results analytically sound: Winner LOLS
    - LOLS CER results are analytical and stable
    - MUPE and ZMPE CER results are sensitive to starting position
  - Absence of Bias: Winner MUPE and ZMPE
    - LOLS CER results are biased, MUPE and ZMPE CER results are not
    - Uncertainty assignment, however, is not influenced by bias
  - Uncertainty assignment: No clear winner
    - LOLS, MUPE, and ZMPE uncertainty results are equally subjective
- Choose the regression method you prefer and use a tool like Distribution Finder to not only fit a distribution to the errors, but to calculate the prediction interval to be used in your uncertainty model

Conclusion

- Uncertainty results can differ substantially even when point estimates are not far apart… choose carefully
- When error is log-normally distributed…use LOLS
  - LOLS uncertainty results are sound and mathematically justified
  - Regression methods such as ZMPE do not have established uncertainty assignment procedure
- If no evidence of log-normally distributed error…take your pick of regression methods
  - Fit a distribution to the actual/predicted ratios
  - Calculate the prediction interval (see S. Hu. 2013)
  - This method can be used on LOLS as well if you are unsure of the uncertainty distribution shape
    - generally assumed to be lognormal with the point estimate at the median

References

- Anderson, T. P. 2009. A Distribution-Free Measure of Significance of CER Regression Fit Parameters Established Using General-Error Regression Methods. Journal of Cost Analysis and Parametrics Summer: 7 - 22.
- Book, S., "Significant Reasons to Eschew Log-Log OLS Regression when Deriving Estimating Relationships," 2012 ISPA/SCEA Joint Annual Conference, Orlando, FL, 26-29 June.
- Hu, S., "Fit, Rather Than Assume, a CER Error Distribution," 2013 ICEAA Annual Conference, New Orleans, LA, 18-21 June 2013
- Hu, S., "Generalized Degrees of Freedom(GDF)," 2015 ICEAA Annual Conference, San Diego, CA, June 9-12.
- Hu, Shu-Ping. 2005. "The Impact of Using Log CERs Outside the Data Range and PING Factor." Paper presented at Joint ISPA/SCEA International Conference, Denver, CO, June 14-17.
- Hu, S., "Prediction Interval Analysis for Nonlinear Equations," 2006 Annual SCEA International Conference, Tysons Corner, VA, 13-16 June 2006
- Hu, S. and A. Smith, "Why ZMPE When You Can MUPE," 6th Joint Annual ISPA/SCEA International Conference, New Orleans, LA, 12-15 June 2007.