Regression and CER Development

4th Annual Workshop Canada
Ottawa, ON

Tuesday, May 1st, 2018

Abstract

The hallmark technique of cost analysis is the Parametric approach of using regression to develop cost estimating relationships, or CERs. CERs are mathematical equations with statistical properties that analysts use to forecast cost and to quantify the uncertainty around those forecasts.

This session will review the foundational concepts of costs drivers and correlation, and explore the various model forms commonly used in CERs. It will present an end-to-end example of CER development, starting with a dataset; using ordinary least squares (OLS) regression to fit an equation and analysis of variance (ANOVA) to characterize goodness of fit; assessing the validity of assumptions, appropriateness of model form, and impact of influential data points; and establishing a prediction interval (PI) for use in risk and uncertainty analysis.

By attending this session, you will be inspired to look for patterns in your data and opportunities to apply this rigorous but accessible statistical technique. You will learn about software tools to enable the process, with a focus on the cost estimator's *lingua franca*, Microsoft Excel. You will gain a better appreciation of the Parametric approach (many data points) as constituting the end of spectrum beginning with Expert Opinion (zero data points) and Analogy (one data point), and its applicability at multiple levels of the estimate and for different life-cycle phases.

This session draws heavily on Module 8 Regression Analysis of the Cost Estimating Body of Knowledge (CEBoK®). It also uses engaging Excel demonstrations to better illustrate and develop intuition for the underlying mathematics, and lively personal datasets from the presenter's avocation as a cruciverbalist.

Acknowledgments

- This presentation is a purposeful “reboot”
- It draws heavily from:
  - Cost Estimating Body of Knowledge (CEBoK®)
    - Your organization should have a license!
  - Joint Agency Cost Estimating Relationship (CER) Development Handbook (dtd 09 Feb 2018)
    - https://www.ncca.navy.mil/references.cfm OTHER REFERENCES & LINKS
    - Shout out to principal author Adam James!
  - Teaching and practice of Regression and CER Development
    - Including Defense Acquisition University (DAU)

Outline

- Cost Estimating Techniques Review
- Cost Drivers and Influence Diagrams
- Least Squares Best Fit (LSBF)
- Analysis of Variance (ANOVA)
- Residual Analysis
  - Functional Forms and Model Assumptions
- Statistical Significance
- Confidence and Prediction Intervals
- Nonlinear and Multivariate

The Question

- How much does an aircraft carrier cost?
  - Or a Joint Support Ship?!
- How long does it take to develop the next-generation ground control software
  - Or a payroll system?
- What is the level of effort needed to conduct Border Security?
- How long does it take me to solve a crossword puzzle?

Expert Opinion – The Last Refuge of the Scoundrel

$n = 0$

- The Delphi Method aka "Round Table"
  - "…everybody's got one!"
- The Subject Matter Expert (SME)
  - Teaching Pigs to Sing
- Not to be confused with Expert Judgment
  - Informs the application of the legitimate techniques
- Not to be confused with Expert Testimony
  - AKA Anecdotal Actuals ("I recall…" vs. "I reckon…")

- The answer: Five and a half minutes [5:30]

Don't be too lazy to Google!

The Technical Baseline

- What: A daily 15x15 American-style crossword puzzle published in the *Washington Post* (L.A. Times syndicate)
- How: Paper and pen(cil)
- When: Day of publication (usually), various times of day
- Who: Peter Braxton (Top 50 ACPT finisher)
- Where: Er, never mind…
- Why: For fun, to learn and keep the brain sharp

What cases are excluded?

Analogy – We Have Data!

$n = 1$

- I solved the puzzle by James Sajdak on Friday, April 6th, 2018, finishing at 12:12 p.m. EDT
  - The theme was puns, first words rhyming with EAR
- How would we adjust this data point for:
  - A different author? editor? publisher?
  - A different day of the week? time of day?
  - A different size puzzle?
  - A different solving method (e.g., online)?

- The answer: 9:26

Estimating outside the range of the data!

Extrapolation from Actuals – We *Really* Have Data!

$n = N$
$k = 1$
$p = 0$

- When in doubt, use the mean
- Average of 320 daily solve times = 6:03
  - CV 24.2%
- Median of 320 daily solve times = 5:36
- Average of 51 Friday solve times = 7:19
  - CV 12.7%
- Confidence Interval (CI) for solve times
  - [5:54, 6:13] daily, [7:03, 7:34] Friday
- Prediction Interval (PI) for solve times
  - [3:10, 8:56] daily, [5:26, 9:12] Friday

Which is the right question?!

Cost Drivers – Diversity Is Good!

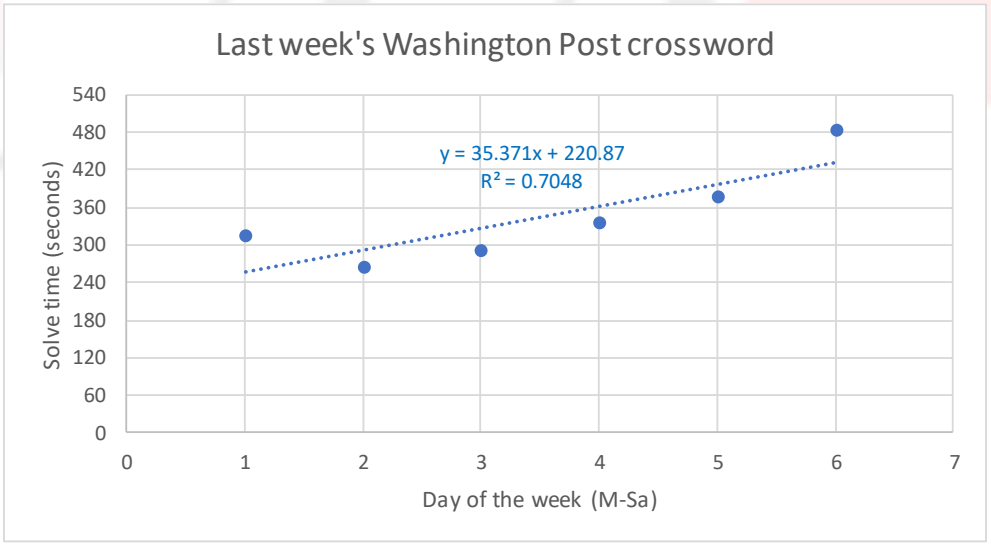
- Ask the Engineers – Expert Judgment
- (Scatter) plot your data – "The gift of sight"
- Draw an Influence Diagram
- Round up the Usual Suspects (including Proxies)
  - Size: Linear Dimension, Square Dimension, Number of Entries
  - Complexity: Author, Theme, **Day of the Week**, Density (of entry squares)
  - Capability: Time of Day, Caffeine Consumption, Solve Mode
  - Quantity: Puzzle Sequence

Don't forget to ask – What will you know ahead of time?

OLS Assumptions

$n = 6$
$k = 2$
$p = 1$

- Ordinary Least Squares (OLS)
  1. Independence of Errors
  2. Normality of Errors
  3. Homoscedasticity
  4. Linearity



"All models are wrong. Some models are useful."

Correlation, Causation, and Significance

- Non-parametric statistics is an under-utilized set of techniques in cost analysis
- Spearman's Rho and Kendall's Tau indicate statistical significance for correlation
  - Based solely on respective *rank order* for paired data

Pearson's R-squared is not the be-all end-all of Regression

The Myth of Homogeneity – Avoid *Reductio ad Absurdum*

- Cost Drivers should be…
  - …as *different* as possible (cf. the hat matrix)
- Indicator Variables (aka Dummy Variables) should be…
  - …*different*, enabling combining sub-populations will similar cost driver response
- Cost data may be…
  - …*different*, as long as they are normalized to be comparable
- All remaining characteristics should be as *similar* as possible

Parametric (CER) approach seeks to use *all* the data

Least Squares Best Fit – A Classic Optimization Problem

- LSBF provides an elegant closed-form solution for the OLS regression coefficients
- Slope is the quotient of the standard deviations of X and Y, scaled by their correlation

$$b = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^{\,2} - n\overline{X}^{\,2}} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}$$

- Y-intercept is determined by the fact that the regression line goes through the mean point

$$Y = a + bX \Rightarrow a = Y - bX$$

Have you heard the one about the three statisticians who went duck hunting?!

Regression Formats in MS Excel

- Data Analysis ToolPak provides a Regression macro
- LINEST(Ys, Xs) provides "live output" for *most* statistics

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.8395							
R Square	0.7048							
Adjusted R Square	0.6310							
Standard Error	47.8812							
Observations	6							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	21894.9143	21894.9143	9.5502	0.0366			
Residual	4	9170.4190	2292.6048					
Total	5	31065.3333						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	220.8667	44.5749	4.9550	0.0077	97.1068	344.6265	97.1068	344.6265
X Variable 1	35.3714	11.4458	3.0903	0.0366	3.5928	67.1500	3.5928	67.1500

	slope	**intercept**	
coeff	35.3714	220.8667	**coeff**
std err	11.4458	44.5749	**std err**
R2	0.7048	47.8812	**SEE**
F	9.5502	4	**d.f.**
SSR	21894.9143	9170.4190	**SSE**
MSR	21894.9143	2292.6048	**MSE**
t-stat	3.0903	4.9550	**t-stat**
p-value	0.0366	0.0077	0.0366

The cost analyst's *lingua franca*

Interpretation of Regression Coefficients

- Slope: the marginal contribution of the cost driver (X) to cost (Y)

35 seconds per day of the week

- Y-intercept: don't do it!
  - Consequence of the mean point – regression line goes through the heart of the data
  - Temptation to think of it as fixed cost

  - How long does it take to solve a crossword puzzle the day before Monday?!

3:41 metaphysical minimum solve time

What is Purchased Services for *Der Fliegende Holländer*?!

Goodness of Fit – Coefficient of Determination

- R^2 is perhaps the most intuitive and widely-used measure of Goodness of Fit
  - Graphical interpretation from best-fit slope
  - "Linear" version of non-parametric rank correlation
  - Percentage of *explained variation*

70%

- There is no magical threshold value for R^2
  - However, there is a little-known significance test for R^2 against the Beta distribution!
- Can be <u>adjusted</u> to "penalize" for d.o.f.
  - Enables comparison between models with different numbers of explanatory variables

Low R^2 means Y is "just not that into" X

Goodness of Fit – Analysis of Variance (ANOVA)

$n = 6$
$k = 2$
$p = 1$

- Pythagorean metrics – sums of squares
- Mean Squares are Sums of Squares divided by their respective degrees of freedom
- SST (MST) = Total uncertainty in model
  - Deviations between *Actual* and *Mean*

31065 / 5 = 6213

- SSE (MSE) = Error remaining in model
  - Deviations between *Predicted* and *Actual*

21895 / 1 = 21895

- SSR (MSR) = explained by Regression
  - Deviations between *Predicted* and *Mean*

9170 / 4 = 2293

MST is just the Variance of Y!

Uncertainty of Coefficients – The "Bounce" and the "Wiggle"

- Each coefficient has an associated standard error
  - SE vs. SSE vs. SEE (coming soon!)
- Normality of errors implies normality of parameter estimates
  - Distribution of *estimated* about *true*
  - Best estimate of standard deviation
- Can be used to derive:
  - Confidence Intervals for the coefficients
  - Significance Tests (*t*-stats and *p*-values) for the coefficients

Slope: 35.4 +/- 11.4

Intercept: 220.9 +/- 44.6

No fit is perfect – unless we have "birds on a wire"

Standard Error of the Estimate – The "Noise"

- SEE is just the square root of MSE
  - Best estimate of the standard deviation of the homoscedastic error term
  - Divided by the mean of Y to produce the coefficient of variation (CV) of the regression

47.9 seconds

- The whole point of regression is to reduce the uncertainty in the prediction
  - Relative to the "naïve" estimate (mean of Y)
  - We can relate reduction in CV to the Adjusted R^2!

78.8 seconds

$$CV_{redux} = \frac{CV_{new}}{CV_{old}} = \sqrt{\frac{n-1}{(n-1)-k}}\sqrt{\frac{SSE}{SST}} = \sqrt{1-R_a^2}$$

The Signal and the Noise

Hypothesis Testing – Bringing Rigor to Regression Analysis

- Null hypothesis: "Innocent until proven guilty"
- Significance level: "Beyond a reasonable doubt"
  - Type I error: "Convicting an innocent man"
  - Type II error: "Letting a guilty man go free"
- Critical value: determined by H_0 and alpha 2.78
- Test statistic: "the preponderance of evidence" 3.09
- There are two equivalent ways to adjudicate a hypothesis test
  - Does the test statistic exceed the critical value?
  - Is the p-value less than the significance level

0.0366 Jurisprudence for CERs! 0.05

Statistical Significance in Regression

- Significance of coefficients
  - Parameter estimate divided by its standard error
  - (Student's) *t*-distribution
    - Thank you, William Sealy Gossett of Guinness Brewing Company!
  - "Is this a good cost driver?"
- Significance of model
  - MSR divided by MSE
  - F distribution
    - Thank you, Ronald Aylmer Fisher!
  - "Is this a good CER?"

3.0903 > 2.7764

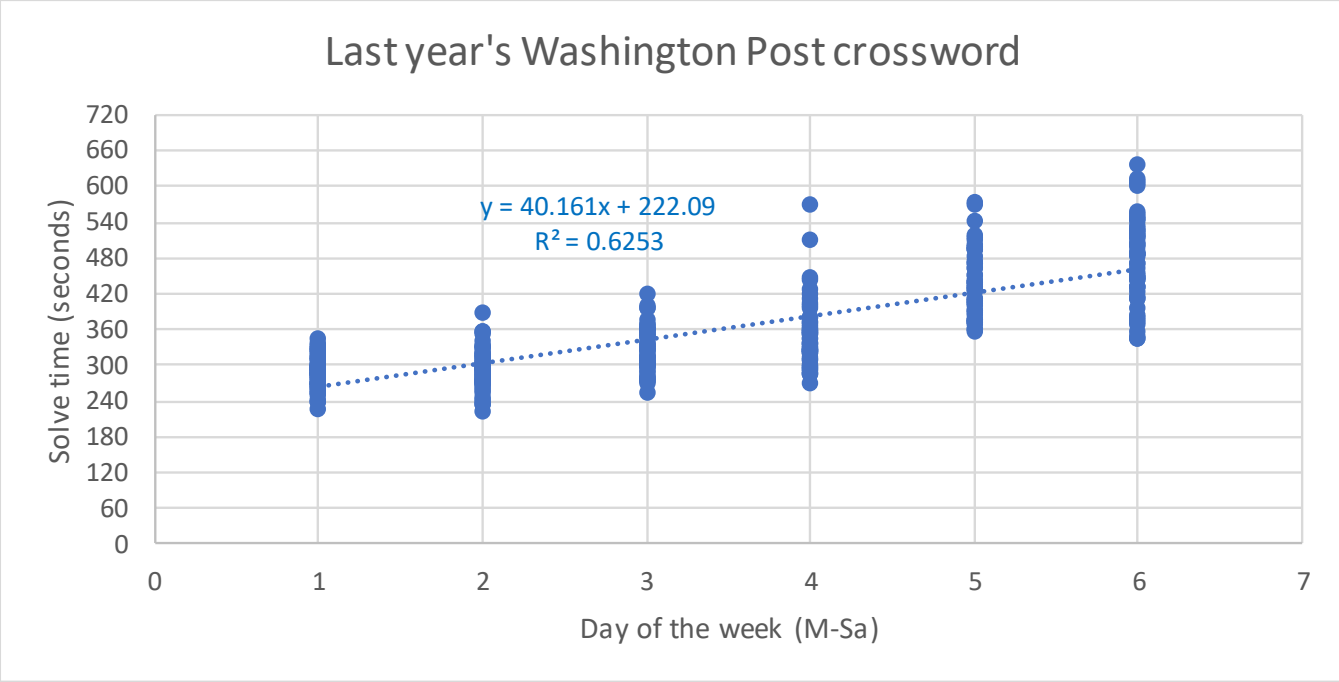
0.0366 < 0.5000

9.5502 > 7.7086

When do we worry about the Y-intercept?

The Power of Parametrics

- What happens when we accumulate more data?



Look where that Analogy data point was!

Influential Data Points (IDPs)

- Outliers with respect to X (leverage)
  - Puzzle 5 at ACPT
- Outliers with respect to Y
  - "That puzzle took forever!"
- Outliers with respect to the regression (Cook's distance)
  - "That puzzle took a lot longer than expected!"

Was it the "Wrath of Klahn" or just a bad day?

Residual Analysis – Regression Diagnostics

- Raw residuals
  - Deviation between *Actual* and *Predicted*
- Internally-studentized (aka standardized) residuals
  - Raw residual divided by approximate standard error
- Externally-studentized residuals
  - Raw residual divided by standard error
  - Follows a *t*-distribution (under OLS assumptions)

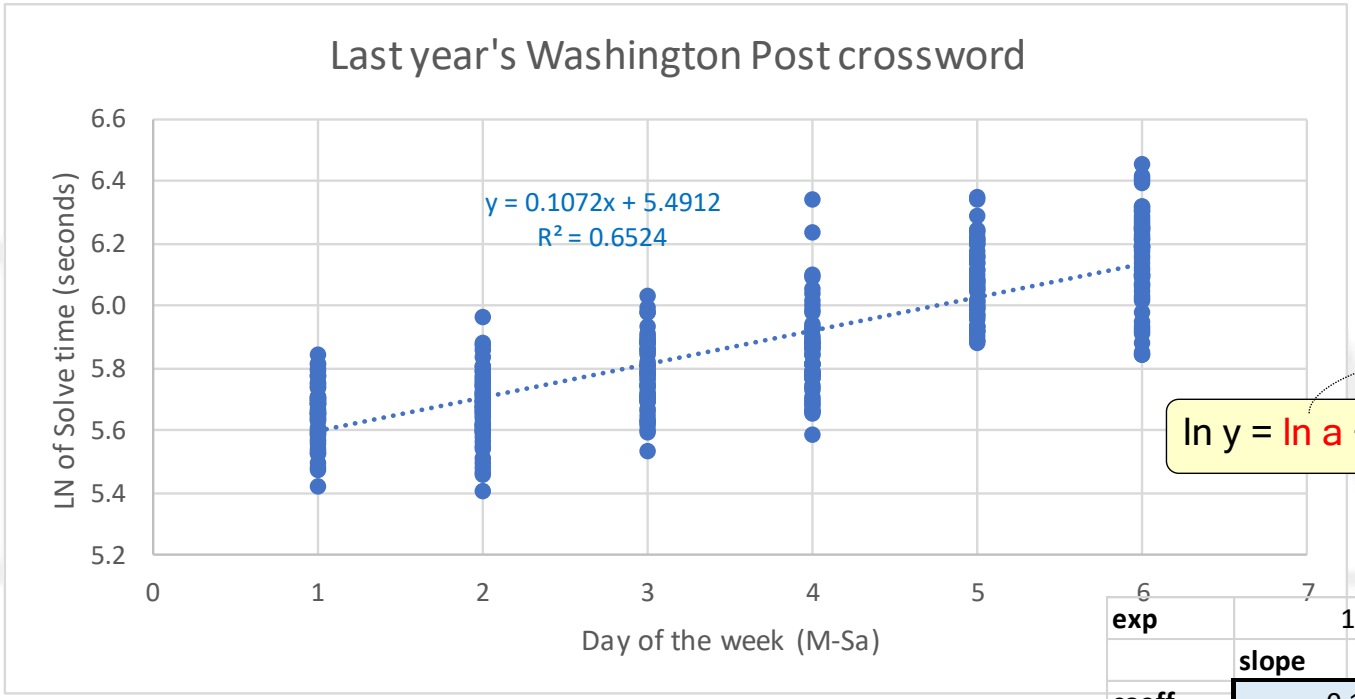
Something looks funny here…

Functional Forms and Transformations

- Linear (Y vs. X)
  - Constant change in Y per change in X ("Rise over run")
- Power (log Y vs. log X)
  - Constant *percent* change in Y per *percent* change in X
  - Learning curve: percent reduction for every doubling
- Exponential (log Y vs. X)
  - Constant *percent* change in Y per change in X
- Logarithmic (Y vs. log X)
  - Constant change in Y per *percent* change in X

Engineering logic (*a priori*) and statistical evidence (*a posteriori*)

Exponential Model Form – Example



$$\ln y = \ln a + b\,x \quad \Leftrightarrow \quad y = a\,e^{b\,x}$$

$a = e^{\ln a}$

$b = b$

exp		1.113	242.550	exp
		slope	intercept	
coeff		0.1072	5.4912	coeff
std err		0.0044	0.0172	std err
R2		0.6524	0.1355	SEE
F		596.8161	318	d.f.
SSR		10.9625	5.8411	SSE
MSR		10.9625	0.0184	MSE
t-stat		24.4298	319.4977	t-stat
p-value		0.0000	0.0000	0.0000

- Interpretation:
  - Base solve time is 4:03
  - Increases by 11.3% each day
    - Including Monday!

Exponential models are rare in cost analysis outside of inflation

Confidence Intervals in Regression – Bridging Cost and Risk

- Confidence Interval general format:
  - Best guess (mean) plus or minus half-width
  - Half-width = multiples [driven by confidence] of standard error [driven by data]
- Confidence Interval of coefficients
- Confidence Interval (CI) of regression
  - What is the true mean Friday solve time? [6:55, 7:11]
- Prediction Interval (PI) of regression
  - What will my solve time be *next* Friday? [5:17, 8:49]

$$\hat{Y} \pm t_{\alpha/2,df} \times SEE \sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum X^2 - n\bar{X}^2}} \qquad \hat{Y} \pm t_{\alpha/2,df} \times SEE \sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{\sum X^2 - n\bar{X}^2}}$$

How do these compare to previous CI and PI?

Indicator Variables – Combining and Stratifying Datasets

$n = n$
$k = 3$
$p = 2$

- Indicator Variable (aka Dummy Variable)
  - 0 = base case (regular daily puzzle)
  - 1 = alternate case (Bob Klahn puzzle)
- Coefficient provides an “adder” (or “multiplier”)
  - Alternate case relative to base case
- Assumes same basic behavior relative to primary cost driver(s)

Which is the right question?!

CER Validation – Yet More Rigor!

- Trying out a bunch of different cost driver combinations and functional forms compounds the likelihood of Type I errors
  - Walking around estimating with a “bad” CER
  - Bonferroni’s *p* is a useful indicator
- Ideally, we randomly divide our dataset in two
  - Training Set: Develop CERs
  - Validation Set: Test CERs
- A dearth of data often means the cost analyst can ill afford this approach

Thank you, John Tukey!

Big Data – the *n* vs. *p* problem

- In traditional data analysis, we have:
  - Large *n* (lots of data)
  - Small *p* (handful of explanatory variables)
  - Preserves degrees of freedom (*n* – *p*)
- In "Big Data," the situation is different:
  - Large *n* (lots of users)
  - Larger *p* (endless data about those users)
- As just noted, cost analysts often live in the world of "Small Data"
- That being said, Big Data is coming
  - Keep your head on a swivel!

Errors, Blunders, and Lies: How to Tell the Difference, David S. Salsburg, CRC Press, 2017

Road Ahead

- Collect and analyze data – It's what we do!
- Invite me back to *really* teach you how to do Regression
- Questions?