The Signal and the Noise in Cost Estimating

Christian Smart, Ph.D., CCEA
Director, Cost Estimating and Analysis
Missile Defense Agency
christian.smart@mda.mil

Introduction (1 of 3)

- **We seek to extract signal and eliminate noise when building models**
- **There are many pitfalls in this process - leads to confusion of signal with noise ("overfitting")**
- **Overfitting is a common problem that interferes with the attempt to develop accurate predictions**
- **Natural tendency to want to explain all historical variation in cost**
- **Leads to inclusion of too many variables in cost models**
- **Real data is messy – includes both repeatable phenomena, as well as random events that cannot be reliably predicted with regularity**
- **For example, a labor strike may have caused an increase in the cost of a historic program in your data set**

Introduction (2 of 3)

- **The most common form of overfitting is including too many variables in your model, but there are others:**
  - **Trying numerous different equation forms**
  - **Trying different types of models**
  - **Etc.**
- **Overfitting problem is worse when you have a small amount of data**
- **Ways to combat overfitting include:**
  - **Limit the number of variables**
  - **Split data into training and testing sets**
  - **Perform cross-validation**

Introduction (3 of 3)

- **Two other considerations we discuss are:**
  - **Normalizing data *before* you model it leads to more noise in the data**
  - **Using bootstrapping to estimate standard errors for small data sets does not fully reflect the residuals you will see going forward**

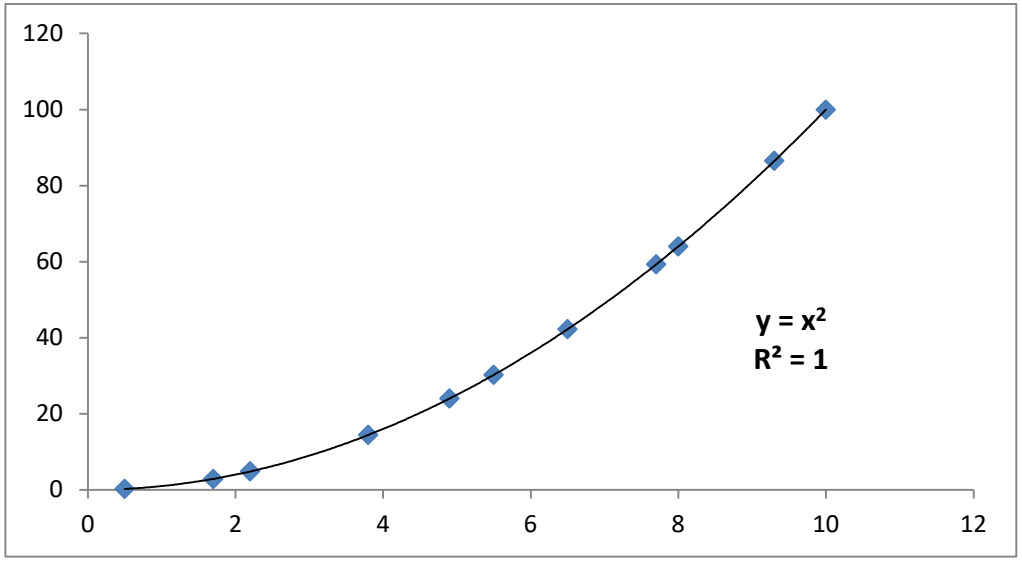
John von Neumann – "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk."

Overfitting

- **Nate Silver in *The Signal and the Noise*, calls overfitting the "the most important scientific problem you've never heard of."**
- **Overfitting – confusing noise with signal**
- **If the fit to the historical data is too loose, it is underfit; if too tight, it is overfit**
- **Overfitting much more common in practice than underfitting**
- **Overfitting is appealing because the fit statistics look great – high R^2s, low standard errors, etc.**

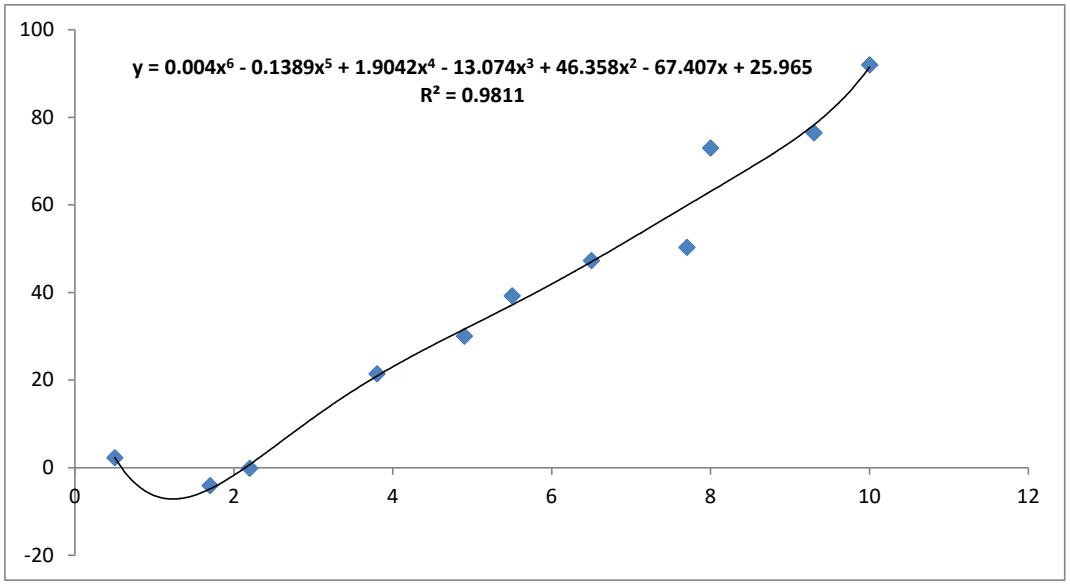
Example of Overfitting (1 of 3)

- **To understand overfitting we illustrate with a simple example**
- **Start with pure signal:**



Example of Overfitting (2 of 3)

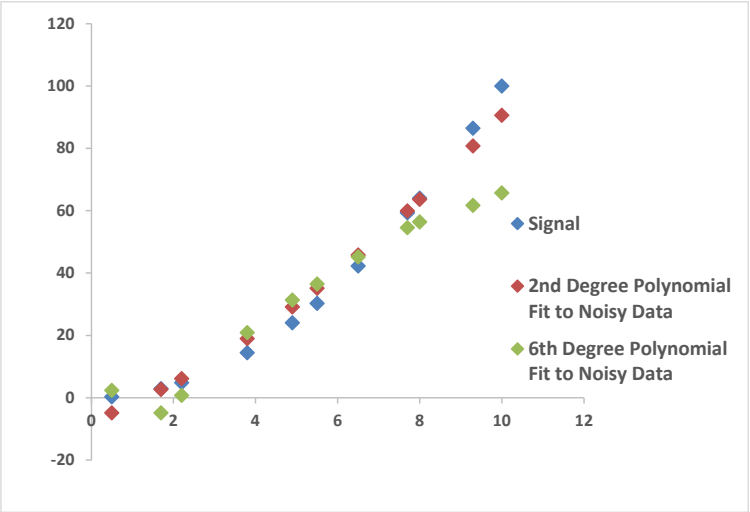
- **Then randomly add noise:**



- **We fit the noisy data with a sixth-degree polynomial with R^2 = 98%**

Example of Overfitting (3 of 3)

- **Compare the sixth-degree polynomial fit and a simpler second-degree polynomial fit to the true underlying signal**



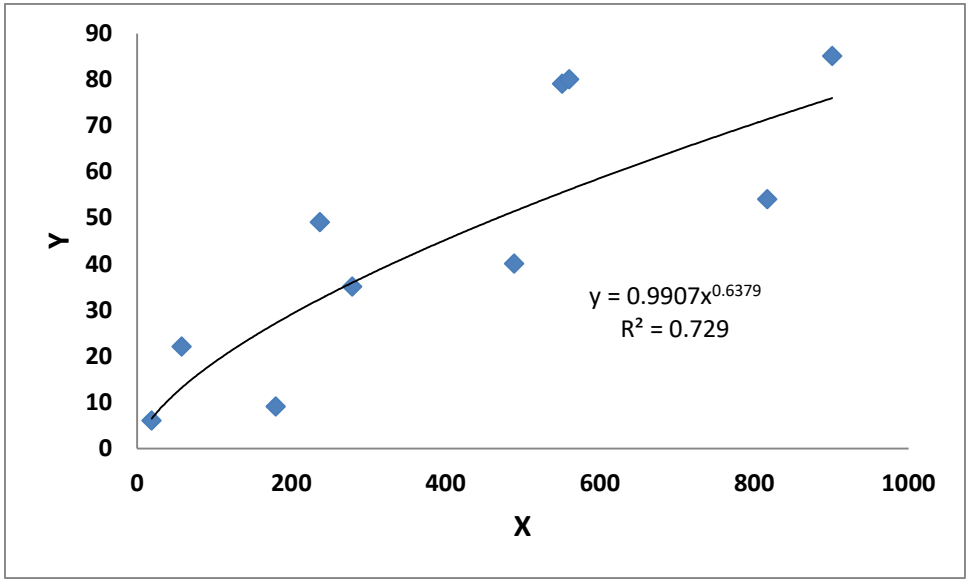
- **Simpler polynomial is a better representation of the underlying signal – closer fit on 10 of 11 data points, and Pearson's R2 between actuals and estimates is 98% for simpler fit vs. 88% for the sixth-degree polynomial**

Noise and Overfitting

- **If we only had access to the true signal, we would not be misled by the noisy data**
- **We can only see the noisy data**
- **Nassim Taleb in *Fooled by Randomness* - in the real world, we have to work by induction, which means we have to infer the structure from the available evidence**
- **Most likely to overfit a model when:**
  - **data are limited**
  - **data are noisy**
  - **when your understanding of the fundamental relationships is poor**

Small Data Sets

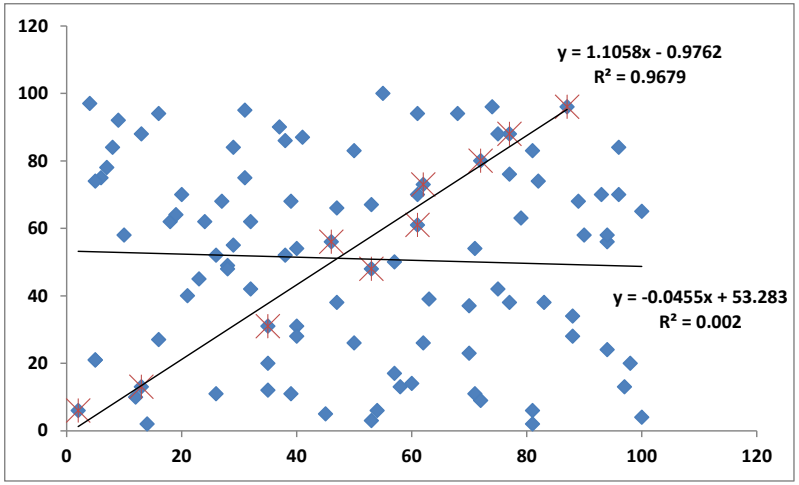
- **There is a strong correlation between the variables displayed in the graph**



- **It's only 10 data points but surely there is a strong connection between these two variables, right?**

Random Data

- **The awful truth is that the data displayed on the graph in the preceding chart is randomly generated**
- **Such patterns are easy to find in small data sets**
- **What we typically see is a small sample of a larger population, even when there is no correlation in the population it is easy to find small samples with a clear pattern**



Small Data and Number of Variables

- **Adding more randomly generated variables in a small data set allows for even better ostensible fits in simple linear regression**

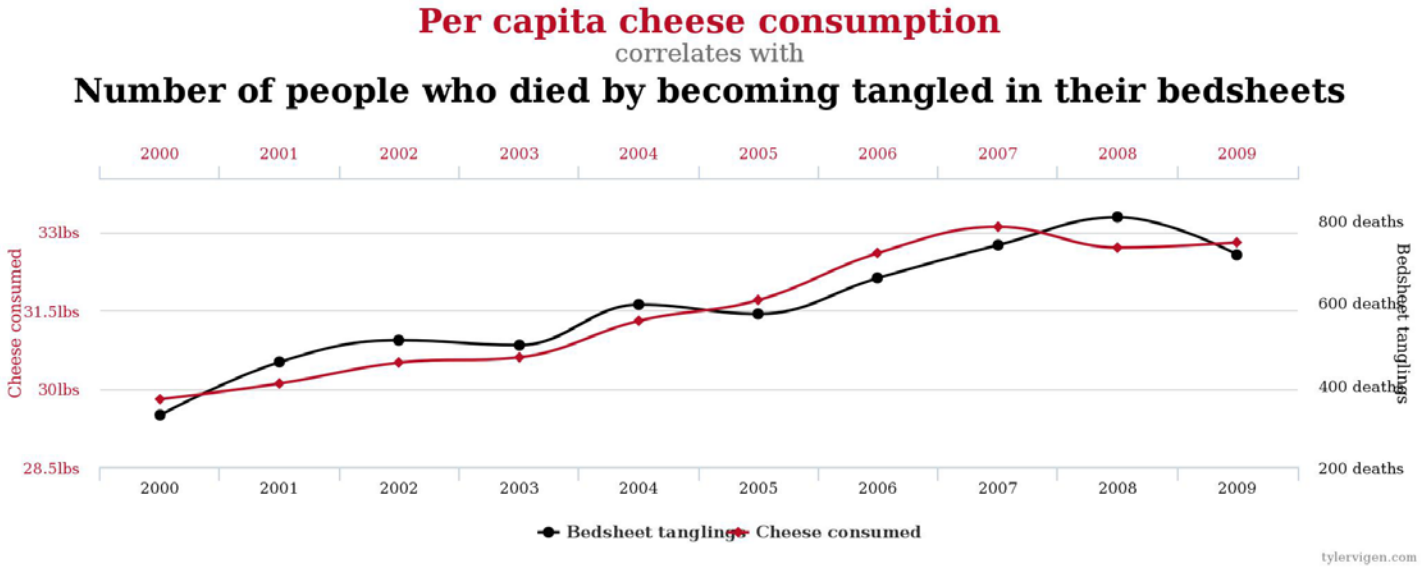
Y	X_1	X_2	X_3	X_4
36	328	482	3	18%
51	124	351	2	70%
41	210	264	3	40%
17	822	99	2	27%
5	255	373	1	92%
11	554	457	7	32%
98	373	24	8	25%
35	551	350	3	6%
46	180	80	9	74%
70	88	250	3	45%

Variables	R^2	SE
X_1	17%	87%
X_1, X_2	44%	76%
X_1, X_2, X_3	53%	75%
X_1, X_2, X_3, X_4	92%	39%

- **Adding four variables increases the R^2 in this set of randomly generated data to 92%!**

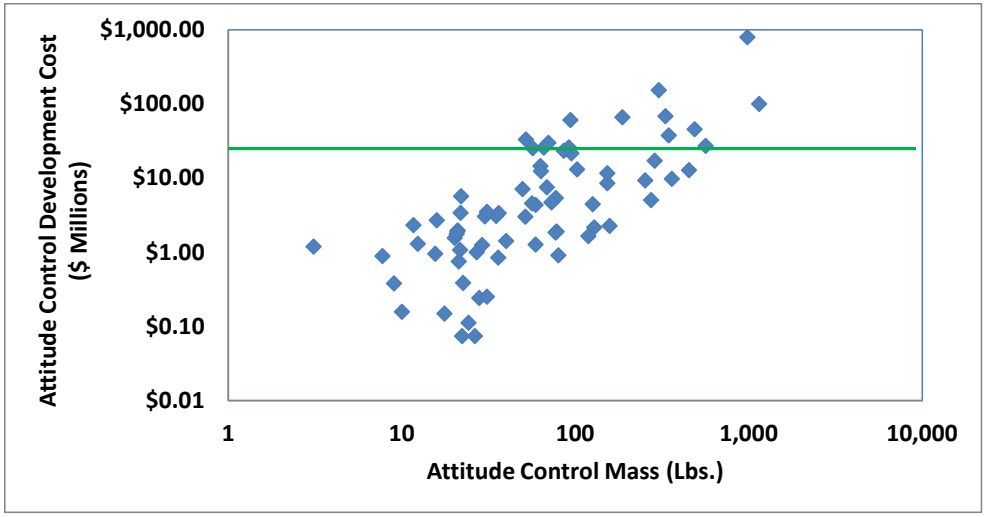
Spurious Correlations

- **Correlations between variables that have no connection are referred to as "spurious correlations"**
- **It is easy to find spurious correlations for small data sets**
- **Tyler Vigen has built a website and has written a book on the subject**



Practical Example

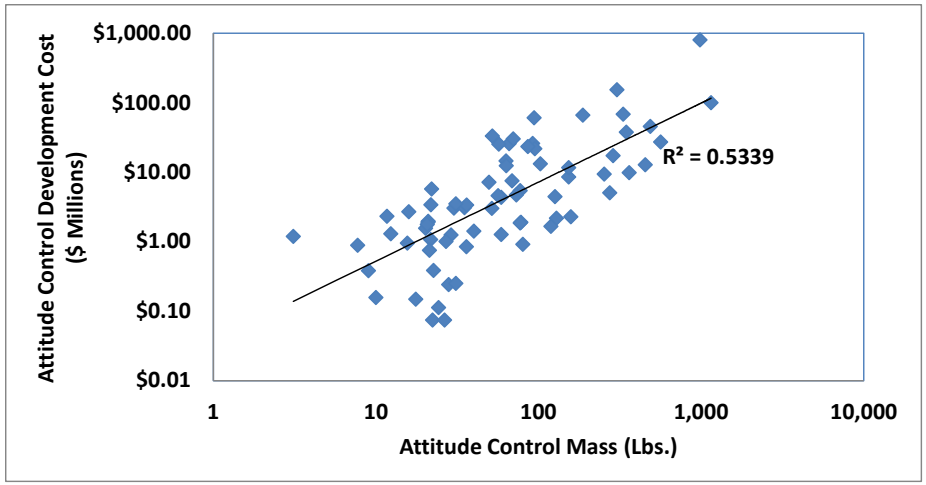
- **Historical data points for 72 NASA and AF satellites and spacecraft**
- **Simple model would be to use the mean of the data set**



- **The average, represented by the green line, is too simple, it is underfit to the data**

A Better Model

- **A regression model based on weight is a step in the right direction**



- **Weight is not truly a cost "driver" but is a good proxy for program scope**
- **Decent R^2 but can we do better?**

How Can We Improve the Model?

- **One option for improving the simple weight-based model would be to add variables and do multivariate regression**
- **One way to do this is *stepwise regression***
- **Popular, traditional method touted by textbooks such as Draper and Smith's *Applied Regression Analysis* and has been implemented in MINITAB**
- **One issue is that each time we sift through the data we lose degrees of freedom**
- **Studies indicate the 30-70% of the variables included in stepwise regressions are *pure noise* (Babyak 2004)**

Number of Variables

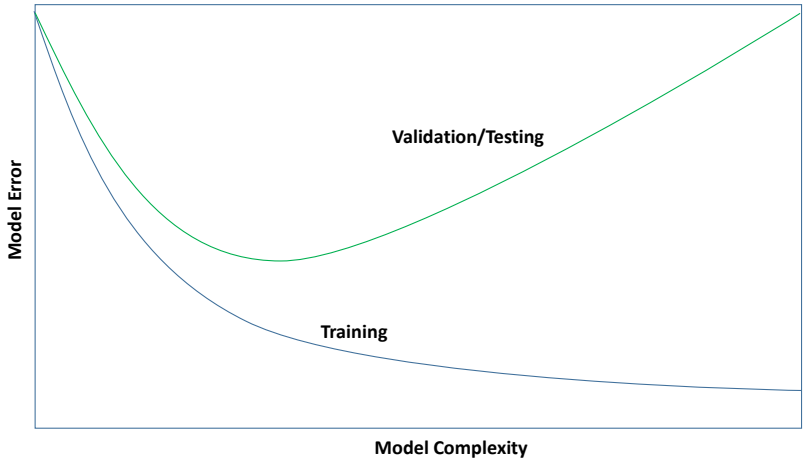
- **A typical statistical rule of thumb is 50 data points plus one variable for every 10 data points after that**
  - **For one variable 50+10*1 = 60 data points are needed**
  - **For two variables 70 data points are required**
  - **For three variables 80 data points; and so on**
- **This rule of thumb is based on simulations of randomly generated data**
- **Reflects the idea of parsimony in modeling**
- **There are not many data sets for government programs that have 50 applicable data points - need to turn to alternate methods, such a Bayesian regression (Smart 2014)**
- **In our example the rule of thumb would allow two variables**

Other Forms of Overfitting

- **In addition to adding variables to a regression model it is also tempting to try out many different types of models, equation forms, and ways to model**
- **Each attempt to improve the fit of a model, such as trying out different model forms, reduces the number of degrees of freedom available to us so this also leads to overfitting**
- **If you try out enough different model forms, variables, and ways of generating models (neural networks vs. regression), you will eventually find a really good fit that is only a good fit due to randomness/luck**

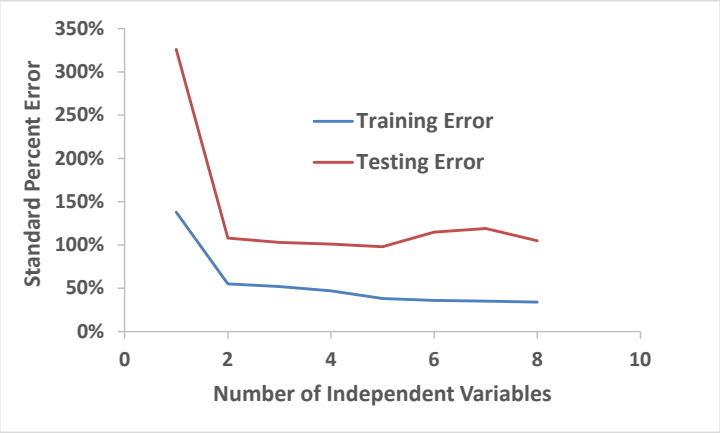
Training and Testing

- One way to reduce overfitting and increase the predictive accuracy of a cost model is to split the data into a training set and a testing set
- One widely used rule of thumb is to split one-third of the available data for validation, and use the other two-thirds for training
- We train the model on the training set and test the goodness of fit on the testing set



Back to the Example

- **Revisiting our attitude control example, we split it into 48 data points for training and 24 for testing**



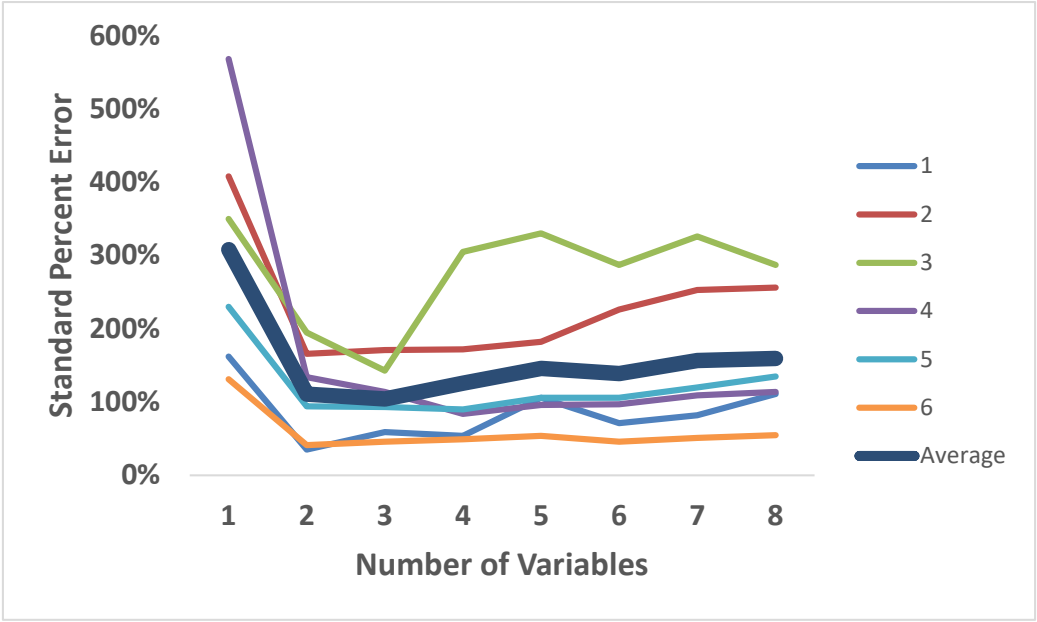
- **We measure the standard percent error for both the training set and the testing set but only use the data in the training set to fit the model**
- **The sweet spot in this case is five variables**
- **At this point, the standard percent error for the testing set is 98%, still much higher than the training error at 38%**

Cross-Validation

- **An alternative when the data set is small is to perform cross-validation instead of separating the data into training and testing sets**
- **Split the data set into multiple partitions and do the testing over multiple small partitions and average the results**
- **Advantage - you can save more of the data for training**
- **For the attitude control subsystem example, we use six-fold cross validation**
- **Split the data into six sets of 12 data points, fit the model on 60 data points, and validate on 12**
- **Do this process six times and average the results over the six validation sets**

Cross-Validation Example

- **The figure shows the standard percent error for each fold as a function of the number of variables, along with the overall average**
- **The average error reaches a minimum on the third variable**



The Final Model

- **Once cross-validation has helped you decide to not use more than three variables in your regression model you can go back and fit the final model using all the data - keep in mind that the predictive accuracy in practice will be worse than the fit on the sample space**
- **Another option is to notice that for each set of variables, six-fold cross validation has produced six different models**
- **One option would be to average the coefficients from the six different models and use the average model to make predictions**
- **This is a simple form of bagging, a powerful technique for variance reduction**

Normalization and Noise-ification

- **Normalizing data is the process of manipulating raw data to make it comparable**
- **While intended for just comparing data points, it is typically the case that estimators model with normalized data, rather than raw data**
- **Normalization is a source of noise if normalized data are used in modeling**
- **Examples of this include learning, test hardware, and inflation**
- **We begin with a set of data, we then apply some type of linear or nonlinear transformation, and then run a regression on this transformed data**
- **To get back to the original data we then have to apply the transformation in reverse**

Inflation (1 of 6)

- **For inflation, we begin with real or "then year" data, normalize to a constant base year, and develop a model in base year dollars**
- **In order to budget we have to convert the model back to real of "then year" cost**
- **The modeling process does not need the transformation - instead, the information can be used in the model as a variable**

Inflation (2 of 6)

- **If we wish to compare the cost of a missile designed and built in the 1960s with a missile designed and built in the 2000s, we need to normalize the data to a common base year**
- **The effect of inflation across the decades makes the comparison meaningless otherwise**
- **For example, the average price of a house built in 1950 was less than $9,000 while in 2016 the average price is $355,000**
- **To have a meaningful comparison we have to consider inflation, as well as taking into account other changes, such as the fact that the average home today is much bigger than a house built in the 1950s, and has much different amenities**

Inflation (3 of 6)

- **This is all well and good for comparing historical data points**
- **But it doesn't mean we should model the data after it has been normalized for inflation**
- **Instead of normalizing the data before we model, we should add a variable that accounts for the year or years in which the project was executed and model the impact directly**

Inflation (4 of 6)

- **For example, applying and modeling the cost of reaction control subsystems for 62 NASA and Air Force missions with weight as the independent variable on normalized data results in the equation $0.17*Weight^{0.74}$**
- **Modeling the non-normalized data with the mid-point of design added as a variable yields the equation $0.07*Weight^{0.85}*Yr\ of\ Tech^{-0.12}$, where *Yr. of Tech = Year of Technology* is defined as the *mid-point of project design – 1960***
- **The first equation produces a cost in a constant base year, whereas the second equation produces cost in real year dollars, based on the year input variable**

Inflation (5 of 6)

- **Note that the value of the year variable is negative - why is this, when we know there is a strong and steady uptick in prices every year? The time coefficient reflects the overall real productivity growth over time, which on average exceeds inflation, reducing net costs overall over time, everything else being equal**
- **When we deflate the normalized data and compare it to the original, raw data, we find a Pearson's R^2 equal to 30%. When we compare the model on un-normalized data with the raw actual data we find a Pearson's R^2 equal to 39%, a big improvement over the normalized model**
- **The standard error of the normalized data is 358% vs. 278% for the non-normalized model**

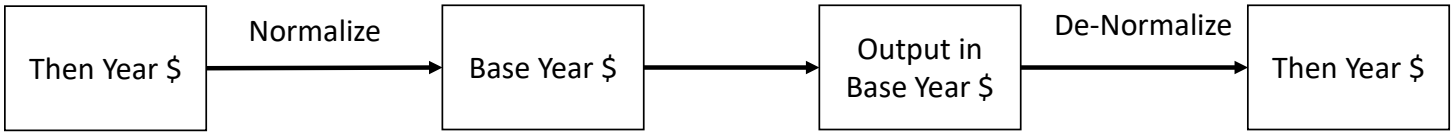
Inflation (6 of 6)

- **The process of normalization when applied to modeling should be called "noise-ification" since it is better to model the raw data directly**
- **Much of this is due to the nonlinearity of the data – if the coefficient of the power equation were equal to 1 then applying a linear filter to the data before and after modeling will have little to no impact**
- **But the application of a linear filter in the presence of nonlinearities, as seen with this example, when introduce noise and error into the equation**

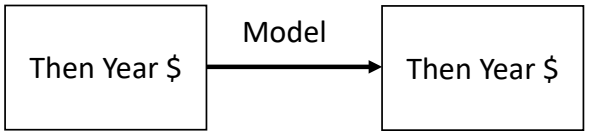
Direct Modeling Vs. Noise-ification

- **Direct modeling is also simpler and requires less work than noise-ification**

Noise-ification



Direct Modeling

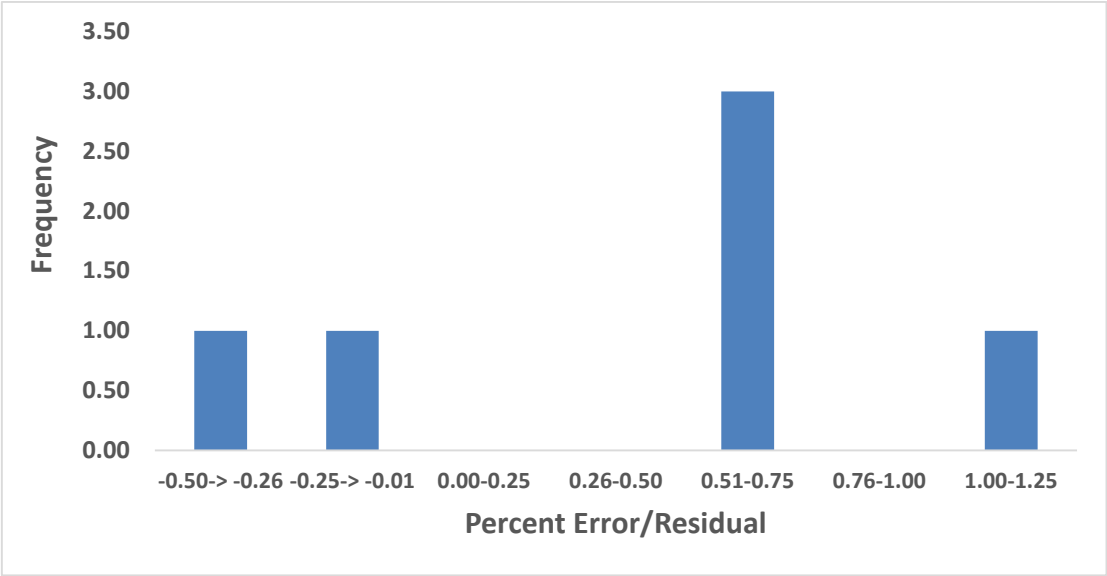


Bootstrap vs. Kernel Smoothing and Distribution Fitting

- **The bootstrap method, so named because it is akin to "pulling yourself up by your own bootstraps" repeatedly draws samples from a given data set to provide alternate outcomes**
- **Works well when you have sufficient data to calculate standard error and confidence intervals for nonparametric regressions**
- **However, when there is a small amount of data, there are large gaps in the data that are not realistic when trying to develop prediction intervals**
- **The use of bootstrapping with small data sets is a form of overfitting to the data**

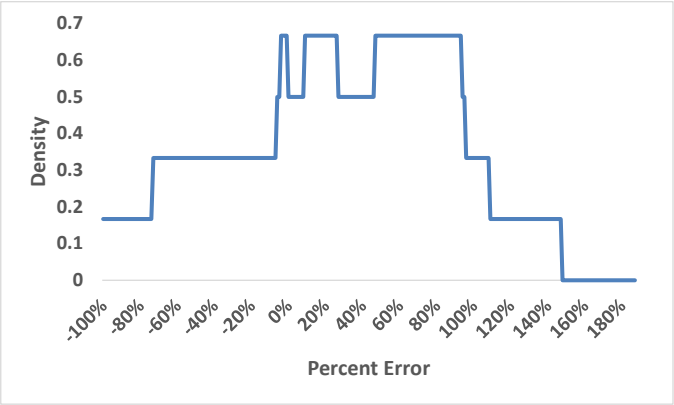
Bootstrap Example

- **Even though there is no reason why we could not see a percentage error between 0% and 50% but the bootstrap would ignore this possibility**



Kernal Smoothing

- **Kernel smoothing is a way to spread this histogram to peanut-butter spread some of the error, based on the available data at hand**
- **This produces a continuous distribution that fills in the gaps left by the original discrete histogram**
- **Simple uniform kernel with bandwidth = 0.50**



- **An alternative is to use the mean and standard deviation to fit a continuous distribution and use that to model the residuals**

Summary (1 of 2)

- **Prediction is a perilous enterprise – the process of model development is filled with pitfalls**
- **Lure of overfitting is powerful, since it is a natural human tendency to want to explain actuals completely – we are hard-wired to look for patterns even where none exist**
- **Easy to confuse noise and signal, especially in small data sets**
- **We have discussed ways to avoid overfitting, including limiting the number of variables, splitting the data into training and testing sets, and cross-validation**

Summary (2 of 2)

- **Normalization of data prior to modeling injects additional noise that can be avoided by directly modeling the phenomenon using the data**
- **Using bootstrapping to calculate standard errors and confidence intervals for small data sets – in such cases kernel smoothing and fitting continuous distributions to the sample moments is a better approach and helps avoid overfitting the residuals**

References (1 of 2)

1. Babyak, M.A., "What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models," Psychosomatic Medicine 66 (Feb. 19, 2004).

2. Draper, N.R., and H. Smith, *Applied Regression Analysis*, 3rd Ed., 1998, John Wiley and Sons, New York.

3. Dyson, F., "A Meeting with Enrico Fermi," Nature 427 (22 January 2004), page 297.

4. Feldman, D., and Springer, S., "Algebraic Formulas for Prediction Bounds on CER-Based Estimates," presented at the 2006 Society of Cost Estimating and Analysis Annual Conference, Tyson's Corner, VA, June 2006.

5. Foussier, P., *From Product Description to Cost: A Practical Approach, Volume 2: Building a Specific Model*, 2006, Springer-Verlag, London.

6. Harrell. F.E., *Regression Modeling Strategies*, 2010, Springer-Verlag, New York.

7. Mitchell, T., *Machine Learning*, 1997, McGraw-Hill, Boston, Massachusetts.

8. Petty, C., C. Smart, and J. Lawlor, "Seven Degrees of Separation: The Importance of High-Quality Contractor Data in Cost Estimating," presented at the International Cost Estimating and Analysis Association Annual Conference, June 2015, San Diego, CA.

9. Prince, A., "The Dangers of Parametrics," presented at the International Cost Estimating and Analysis Association Annual Conference, June 2016, Atlanta, GA.

10. Silver, N., *The Signal and the Noise: Why So Many Predictions Fail – But Some Don't*, 2012, Penguin Books, New York.

References (2 of 2)

11. Smart, "Bayesian Parametrics: How to Develop a CER with Limited Data and Even Without Data," presented at the International Cost Estimating and Analysis Association Annual Conference, June 2014, Denver, CO.

12. Taleb, N.N., *Fooled by Randomness*, 2nd ed., 2004, TEXERE, New York.

13. Vygen, T., *Spurious Correlations*, 2015, Hachette Books, New York, and http://www.tylervigen.com/spurious-correlations.

14. Wasserman, L., *All of Statistics: A Concise Course in Statistical Inference*, 2005, Springer, New York.