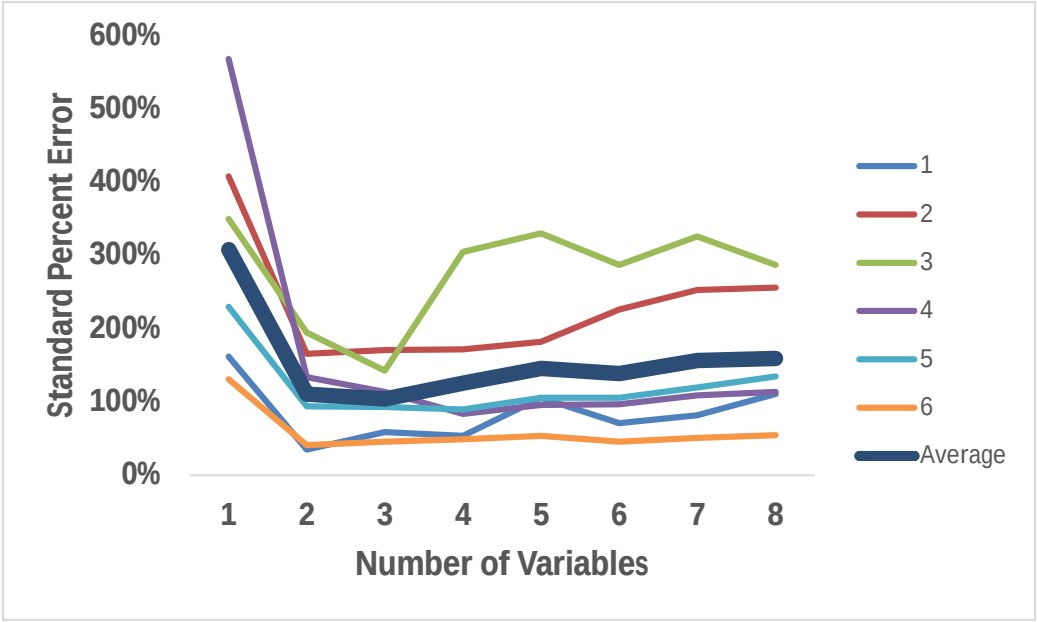
Cross-Validation Example

- **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**

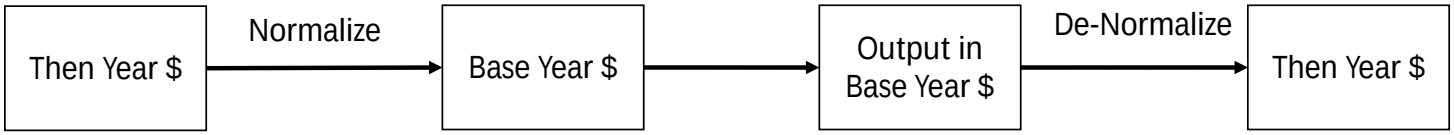
Inflation (6 of 6)

- **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**

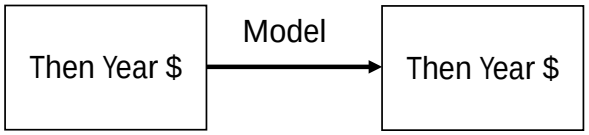
Direct Modeling Vs. Noise-ification

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

Noise-ification



Direct Modeling

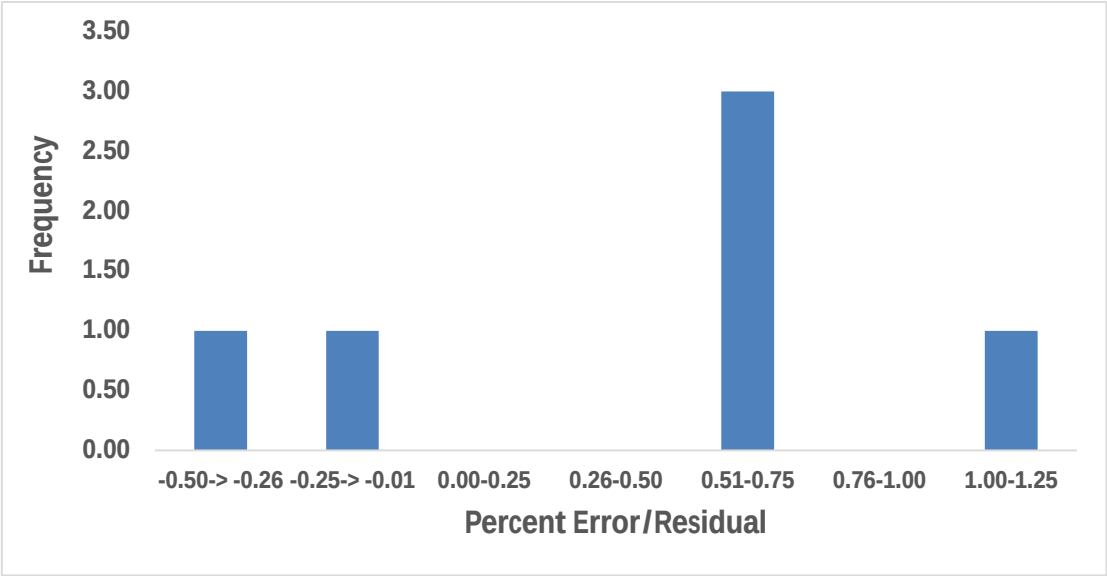


Boostrap vs. Kernel Smoothing and Distribution Fitting

- **The bootstrap method, so called 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**

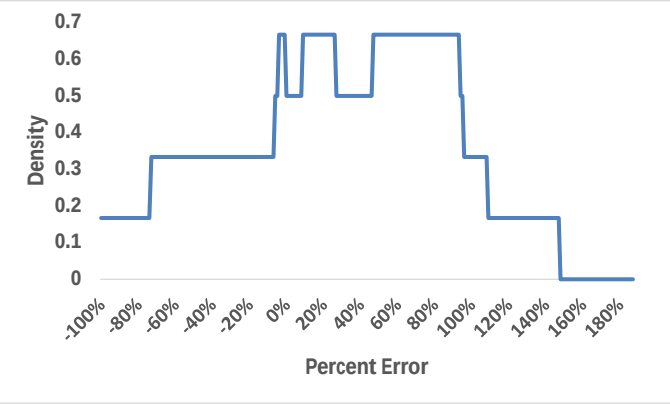
Boostrap Example

- **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)

- **We have discussed the issue that normalization of data prior to modeling injects additional noise that can be avoided by directly modeling the phenomenon using the data**
- **We have also discussed the potential issue of 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.**