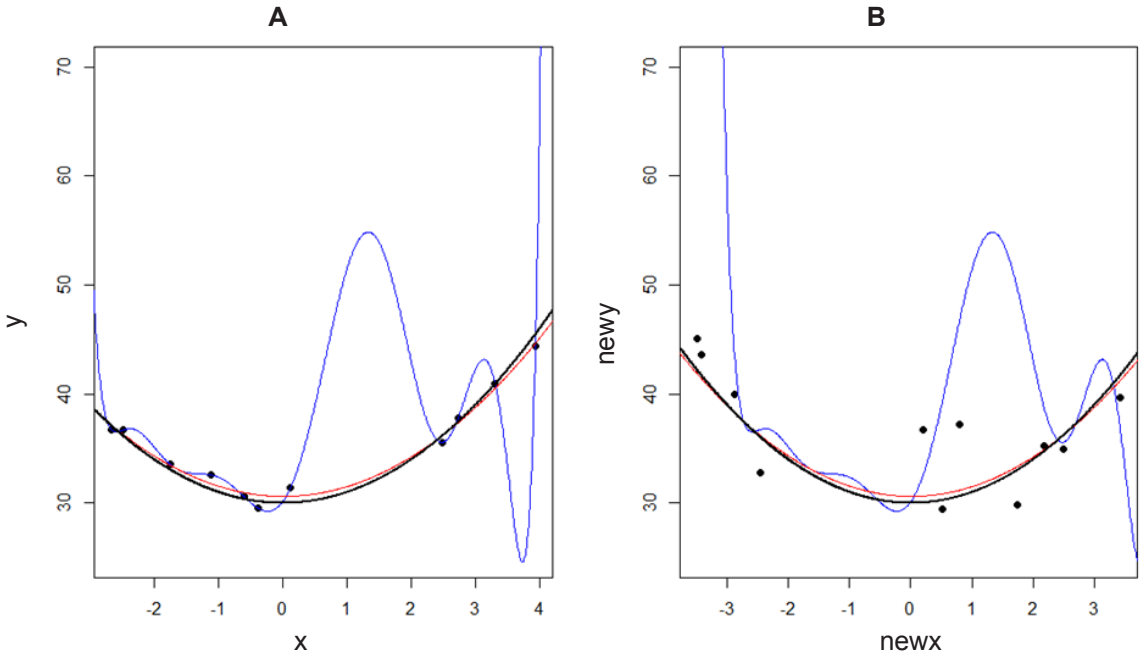
Model selection and model over-fitting

Shengping Yang PhD, Gilbert Berdine MD

I have collected some demographic, clinic, and blood test data from colon cancer patients treated from 2010 to 2012. I am interested in factors that potentially affect serum CRP levels shortly after diagnosis. I am thinking to use all the data (variables) in a regression model to evaluate such relationships. Is model over-fitting a concern?

Real world data very often consist of true signal and random noise. Although it's ideal to fit only the true signal using a perfect statistical model to explore the underlying relationship between the outcome $y=f(x)+\varepsilon$ and the independent variables (predictors) x, many times the true signal $f(x)$ and the random noise ε cannot be separated from each other. In other words, we can only observe y and x, and would not be able to observe $f(x)$, where $f()$ is the function that defines the underlying relationship between y and x.

Let's recall the process of data collection - in reality, it is rarely feasible to evaluate the underlying relationships using data of an entire population. The common approach is to take random samples from the population, and then using the sample data to infer such relationships. However, due to the random nature of sampling, as well as the random noise inherited in individual observations, it is common that statistical models fitted using different random samples (collected from the same population) will have different parameters. Moreover, in situations where ε is substantial, and the number of independent variables is large, a statistical model might mainly describe the random noise rather than the underlying relationship; this effect is called over-fitting. For example, if one



Corresponding author: Shengping Yang PhD
Contact Information: Shengping.Yang@ttuhsc.edu
DOI: 10.12746/swrccc2015.0312.160

fits *n* data points to n parameters, the fit will be exact, but it will be unlikely to work well with another sample of *n* points from the same population. The opposite problem – under-fitting – occurs when too few parameters are used. A model that is under-fitted will work well only if the parameter left out is homogeneously distributed throughout the population.

The classical example of model over-fitting is to fit data with a polynomial regression. Suppose that the true underlying relationship between *x* and *y* in a population is $y=30+x^2$ (this is rarely known in reality; we assume that it is known for demonstration purposes). First, we randomly sample 9 observations from the population (panel A), and fit the data with both a second order polynomial (red curve; the "correct" model), and a 10th order polynomial (blue curve: the "over-fitted" model). As we can see, although the 10th order polynomial has a perfect fit to the data, it substantially deviates from the true relationship (black curve). As a comparison, although the second order polynomial does not fit the data perfectly, it agrees well with the true relationship. Next, we take another independent random sample of 9 observations from the same population (panel B). It can be seen that the second order polynomial developed on the first sample has an acceptable fit also to the second independent sample. On contrast, the 10th order polynomial has a very poor fit. In other words, the 10th order polynomial model fits not only the underlying relationship, but also the variation associated with random error, thus it is a over-fitted model. Therefore, it is not surprising that such a model fits well to one sample, but poorly to another one.

Since the goal of any biomedical/clinical study is to disclose the true underlying relationship, it is critical to find the "correct" model in data analysis.

To find the "correct" model and avoid model over-fitting, many methods have been proposed. The majority of them adopt one of the following strategies: 1) penalize models with more parameters - since increased number of parameters in a model is associated with higher probability of modeling the random error, penalizing extra parameters reduces the risk of over-fitting; 2) use validation data set(s) to evaluate the performance of the fitted model - since the true underlying relationship is supposed to be consistent (the random noise is not) across samples randomly collected from the same population, models that consistently have good performance on different samples are more likely to be the one that models the true underlying relationship rather than the random noise.

1. Penalize models with more parameters

Penalizing additional parameters (predictors) can be achieved by either performing the traditional model selection (based on different criteria) or applying penalized regression models.

(A) Traditional model selection

i. Selection based on: *Adjusted R-squared, Mallows' Cp, AIC* and *BIC*

R-squared is the percentage of outcome variable variation explained by the model, and describes how close the data are to the fitted regression. In general, the higher the *R-squared* value, the better the model fits. However, *R-squared* always increases with additional predictors, thus models with more extra predictors always have higher *R-squared* values. The **adjusted** *R-squared* adjusts the *R-squared* value for the number of predictors, and it increases only if the additional predictor improves the model more than would be expected by chance. Thus, models with higher adjusted *R-squared* are generally considered better.

Mallows' Cp estimates the mean squared prediction error, and is a compromise among factors, including sample size, collinearity, and predictor effect sizes. The adequate models are those with *Cp* less than or equal to the number of parameters (including the constant) in the model.

AIC is "Akaike's Information Criterion", and *BIC* is "Schwartz' Bayesian Criterion." Both aim at achieving a compromise between model goodness of fit and

model complexity. The only difference between *AIC* and *BIC* is the penalty term, where *BIC* is more stringent than *AIC.* The preferred models are those with minimum *AIC/BIC.*

ii. Best subset / forward / backward / step-wise selection

Assuming that the total number of predictors is *p*, then the best subset selection fits 2^p total models, and chooses the best model based on criteria, such as, adjusted *R-squared, Mallows' Cp, AIC/BIC.* However, if the total number of predictors is large, for example, greater than 30, then the computation can be a big issue.

In *forward selection*, the most significant variable (based on certain pre-set confidence level) is added to the model one at a time, until no additional variable meets the criterion.

Backward selection starts with the full model that includes all the variables of interest, and then drop non-significant variables one at a time, until all the variables left are significant.

Step-wise selection allows both adding and dropping variables to allow dropped variables to be reconsidered.

As an alternative to the above traditional model selection methods, penalized regressions achieve coefficient estimation and model selection simultaneously.

(B) Penalized regressions (*LASSO* regression and *Elastic Net*)

The *LASSO* (Least Absolute Shrinkage and Selection Operator) achieves model selection by penalizing the absolute size of the regression coefficients. In other words, *LASSO* includes a penalty term that constrains the size of the estimated coefficients. As a result, solutions of the lasso regression will have many coefficients set exactly to zero, and the larger the penalty applied, the more estimates are shrunk towards zero. In general, the penalty parameter is chosen by cross validation to maximize out-of-sample fit.

Elastic Net regression was developed to overcome the limitations of *LASSO*, and in general outperforms *LASSO* when the predictors are highly correlated.

2. Cross validation

Any random sample will differ from its population. From a given population, two independent samples share the true underlying relationship, but not the sample-specific variation. It is important to assess how well the model developed on one sample performs on another independent sample, and fine tune model parameters. Cross validation is an easy-to-implement tool to make such assessments.

(A) *k*-fold cross validation

The *k*-fold cross validation is one of the most commonly used methods. Basically, the sample is randomly partitioned into *k* equal size subsets, then one of the k subsets is used as the validation set, and all the other *k*-1 subsets are used as the training set. This process is repeated *k* times (folds), so that each of the *k* subsets is used exactly once as the validation set. Results from the *k* validations can then be averaged to produce a single estimation. As a special case, if *k* equals *N*, which is the total number of observations, then the *k*-fold cross validation is called the leave-one-out cross validation.

(B) random sub-sampling validation

In *k*-fold cross validation, the proportion of the training/validation subsets depends on the number of iterations (folds). In situations where the total number of sample observations is small, it would make sense to use random sub-sampling validation, such as bootstrap. Note that the disadvantages of random sub-sampling are that some observations might never have been sampled, and the results might vary if such randomization is repeated.

Some other concerns:

Although over-fitting is a real issue in statistical modeling, model under-fitting can also have serious consequences. For example, a family history of cancer is a strong risk factor associated with breast cancer. Suppose that if in a breast cancer study, data on the family history were not collected, it is very likely that models developed using such data produce biased coefficient estimates due to complex relationships among predictors. Therefore, in any biomedical/clinical studies, investigators are expected to have a comprehensive understanding of the research topic, so that the study design does not have any fundamental flaw. In fact, statistical modeling would make sense only after all biomedical/clinical considerations are well addressed.

On the other hand, although model over-fitting should be avoided, sometimes it would be clinically sound to keep a predictor in the model even it does not have a "statistically significant" effect. In other words, if there is strong clinical or biomedical evidence that a factor is strongly associated with the outcome of interest, we should always include that factor in the model so that its effect can be adjusted.

Other times, costs of data collection might be a factor to be considered for determining whether or not to include a predictor into a prediction model. In situations where it is costly to measure a predictor, an easy-to-measure alternative should be considered, and even compromises may have to be made on the performance of such an alternative(s).

Overall, model selection is a critical step in data analysis. Considerations from both the clinical/biomedical and statistical aspects need to be well balanced to develop a meaningful model.

Author affiliation : Shengping Yang is in the Department of Pathology at Texas Tech University Health Sciences Center in Lubbock, TX. Gilbert Berdine is in the Department of Internal Medicine at TTUHSC in Lubbock, TX.
Published electronically: 10/15/2015

References

1. Akaike H. Information theory and an extension of the maximum likelihood principle. in Petrov BN, Csáki F. *2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR, September 2-8, 1971*, Budapest: Akadémiai Kiadó, p. 267-281, 1973.
2. Mallows CL. Some Comments on CP. *Technometrics* 1973; 15 (4): 661-675. doi:10.2307/1267380.
3. Schwarz Gideon E. Estimating the dimension of a model. *Annals of Statistics* 1978; 6 (2): 461–464, doi:10.1214/aos/1176344136.
4. Theil H. Applied economic forecasting. Amsterdam, The Netherlands: North-Holland, p 474, 1966.
5. Tibshirani R. Regression shrinkage and selection via the lasso. J Royal Statist Soc B 1996; 58(1): 267-288.
6. Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. *J Royal Statist Soc B* 2005; 67, 301–320.