Announcement

Simplifying a Multiple Regression Equation

Sometimes research questions involve selecting the best predictors from a set of candidates that, at the outset, seem equally likely to prove useful. The question is typically phrased, "Which ones of these predictors do I need in my model?" or "Which predictors really matter?" Although many methods have been proposed, **the standard purely statistical approaches for simplifying a multiple regression equation are unsatisfactory.** The reason is simple. With rare exception, **a hypothesis cannot be validated in the dataset that generated it**

Many multiple regression models contain variables whose t statistics have nonsignificant P values. These variables are judged to have not displayed statistically significant predictive capability in the presence of the other predictors. The question is then whether some variables can be removed from the model. To answer this question, many models are examined to find the one that's *best* in some sense.

The theoretical basis for concern over most simplification methods

The main concern is that many of the measures used to assess the importance of a variable were developed for examining a single variable only. They behave differently when assessing the *best*.

If you take a fair coin and flip it 100 times, common sense as well as probability theory says the chance of getting more heads than tails is 50%. However, suppose a large group of people were to each flip a coin 100 times. Again, both common sense and probability theory say that it is **unlikely** that the coin with the **most heads** has more tails than heads. For the best coin to show more tails than heads, they would *all* have to show more tails than heads. The chance of this becomes smaller the more coins that are flipped.

Most of the time (95%) there will be between 40 to 60 heads when a single fair coin is flipped 100 times. The chances of getting more than 60 heads are small, *for a single coin*. If we were suspicious of a particular coin and noticed that there were 70 heads in the next 100 flips, we'd have some statistical evidence to back up our suspicions.

Have a large group of people flip fair coins 100 times and the chances of *someone* getting more than 60 heads grows. On average, about 2.5% of the participants will get more than 60 heads. In this situation, we *might* become suspicious of the coin that recorded the most heads but, we'd have to test it again to be sure. If we had no reason other than the number of heads for being suspicious and were flip the coin another 100 times, it wouldn't be surprising to see it behave more typically this time.

To summarize:

- If our attention is drawn to a particular coin and it *subsequently* shows an excess of heads, we have some basis for our suspicion.
- If a large number of coins are flipped, we can't judge the coin with the largest number of heads as though it were the only coin flipped.

The same thing applies to building models. If there is a reason for singling out a predictor before the data are collected, then it is fair to say the variable has predictive value if it achieves statistical

significance. However,

- when many predictors are considered and
- there is nothing special about any of them before the data are collected and
- we judge them as though each were the only predictor being considered,

probability theory says that *something* will likely achieve statistical significance due to chance alone. We shouldn't be surprised if 1 of 20 such predictors achieves what would be statistical significance for a single predictor at the 0.05 level. This explains why measures that are unrelated to the response sometimes appear to be statistically significant predictors.

A similar problem arises when many variables predict the response equally well. Statistical theory says that, in any sample, some variables will appear to be better predictors than others. However, since all variables predict equally well, these particular variables are not really better. They appear that way due to chance, showing once again that **a hypothesis cannot be validated in the dataset that generated it**.

We need a procedure that can distinguish between variables that are truly better predictors and those that appear to be better due to the luck of the draw. Unfortunately, that procedure does not yet exist. Still, many procedures have been tried.

Stepwise Procedures

One approach to simplifying multiple regression equations is the stepwise procedures. These include **forward selection**, **backwards elimination**, and **stepwise regression**. They add or remove variables one-at-a-time until some stopping rule is satisfied. They were developed before there were personal computers, when time on mainframe computers was at a premium and when statisticians were considering the problem of what to do when there might be more predictors than observations.

Forward selection starts with an empty model. The variable that has the smallest P value when it is the only predictor in the regression equation is placed in the model. Each subsequent step adds the variable that has the smallest P value in the presence of the predictors already in the equation. Variables are added one-at-a-time as long as their P values are small enough, typically less than 0.05 or 0.10.

Backward elimination starts with all of the predictors in the model. The variable that is least significant--that is, the one with the largest P value--is removed and the model is refitted. Each subsequent step removes the least significant variable in the model until all remaining variables have individual P values smaller than some value, such as 0.05 or 0.10.

Stepwise regression is similar to forward selection except that variables are removed from the model if they become nonsignificant as other predictors are added.

Backwards elimination has an advantage over forward selection and stepwise regression because it is possible for a set of variables to have considerable predictive capability even though any subset of them does not. Forward selection and stepwise regression will fail to identify them. Because the variables don't predict well individually, they will never get to enter the model to have their joint behavior noticed. Backwards elimination starts with everything in the model, so their joint predictive capability will be seen.

Since variables are chosen because they look like good predictors, estimates of anything associated with prediction can be misleading. Regression coefficients are biased away from 0, that is, their magnitudes often appear to be larger than they really are. (This is like estimating the probability of a head from the fair coin with the most heads as the value that gained it the title of "most heads.") The t statistics tend to be larger in magnitude and the standard errors smaller than what would be observed if the study were replicated. Confidence intervals tend to be too narrow. Individual P values are too small. R^2, and even adjusted R^2, is too large. The overall F ratio is too large and its P value is too small. The standard error of the estimate is too small.

These objections have been well known and are mentioned in even the earliest books on regression analysis, but they became codified in 1995 in a post by Frank Harrell as part of a spirited debate in the Usenet group sci.stat.consult, so they are now known in Usenet lore and sometimes beyond as *Harrell's Nine Points*.

> **Nominal Significance**: Stepwise procedures are sometimes described as adding variables one-at-a-time as long as they are statistically significant or removing them if they are nonsignificant. This means comparing a variable's P values to some value, often 0.05. With forward selection, we are looking at the smallest P value to decide what to include. With backwards elimination, we are looking at the largest P value to decide what to remove. However, these are not special P values. They are the same ones used to assess a single predictor in a multiple regression equation. For the reasons already stated, these P values are artificially small or large. It is incorrect to call them *statistically significant* because they don't take account of the selection procedure. To acknowledge this, many statisticians call such P values *nominally significant*, that is, significant in name only.

When variables are highly correlated, the ones that appear in the model do so as a matter of chance and can change with the addition of one or two more observations. The idea that our assessment of a particular predictor might change with the addition of one or two observations doesn't bother me for the most part. That's part of the game. We choose our test, collect our data, and calculate the results, letting the chips fall where they may. In multiple regression, the worst that can happen is that some coefficients and P values might change a bit. P values might move from one side of 0.05 to the other, but confidence intervals for regression coefficients will be grossly the same. The troublesome feature of *stepwise* procedures is that the characteristics of the report model can change dramatically, with some variables entering and others leaving.

A final condemnation of stepwise procedures is often encountered when missing data are involved. Stepwise procedures must exclude observations that are missing *any* of the *potential* predictors. However, some of these observations will not be missing any of the predictors in the final model. Sometimes one or more of the predictors in the final model are no longer statistically significant when the model is fitted to the data set that includes these observations that had been set aside, even when values are missing at random. (This was a tenth point raised by Paul Velleman in response to Harrell's Usenet Post.)

All Possible Regressions

Other simplification procedures examine all possible models and choose the one with the most favorable value of some summary measure such as adjusted R^2 or Mallows' C(p) statistic. "All Possible Regressions" has a **huge** advantage over stepwise procedures, namely, it can let the analyst see competing models, models that are almost as good as the "best"and possibly more meaningful to a subject matter specialist. However, whenever a model is chosen because of an extreme value of

some summary statistic, it suffers from those same problems already mentioned. While I've seen many discussions of examining all possible models, I've never seen a report of anyone doing this in practice.

Some investigators suggest plotting various summary statistics from different models as a function of the number of predictors. When the standard error of the estimate is plotted against the number of predictors, the SEE will typically drop sharply until some floor is reached. The number of predictors needed to adequately describe the data is suggested by where the floor starts. The final model might be chosen from competing models with the same number of predictors, or maybe 1 or 2 more, by our knowledge of the variables under study. In similar fashion, some statisticians recommend using knowledge of the subject matter to select from nearly equivalent models the first time C(p) meets its target of being less than or equal to p+1.

Data Splitting

Another approach to the problem is **data splitting**. The dataset is divided in two, at random. One piece is used to derive a model while the other piece is used to verify it. The method is rarely used. In part, this is due to of the loss in power (ability to detect or verify effects) from working with only a subset of the data. Another reason is a general because that different investigators using the same data could split the data differently and generate different models.

It has always struck me as a peculiar notion that one could use a subset of the data challenge from what was observed in the reaminder or in data as a whole. if the full dataset has some peculiarity, the laws of probability dictate that each of the two halves should share it.

The Bootstrap

Today, some analysts are looking to the **bootstrap** for assistance. Bootstrap samples are obtained by selecting observations with replacement from the original sample. Usually, bootstrap samples are the same size as the original sample. They can be the same size as the original sample because the observations composing a bootstrap sample are chosen independently with replacement (that is, when an observation is chosen, it is thrown back into the pot before another is chosen). The typical bootstrap sample will contain duplicates of some original observations and no occurrences of others. The stepwise procedure is applied to each bootstrap sample to see how the model changes from sample to sample, which, it is hoped, will give some indication of the stability of the model. I am not optimistic about this approach for reasons stated here.

So...what do we do?

Some analysts soldier on regardless and look for consistency among the methods. They gain confidence in a model if most every method leads to the same candidate. Perhaps there is some justification for this belief, but I am inclined to think not. If due to chance a particular set of variables looks better than it really is, it's unlikely that the reason for this excellence will be uncovered, regardless of the lens used to examine the data.

Perhaps this outlook is too pessimistic. In a November, 2000, post to the S-News mailing list for users of S-Plus, Jeff Simonoff presented a cogent argument for using automatic methods. He states that he considers stepwise methods obsolete but does talk about "all subsets regression" in his teaching. He is adamant about validation, but would use a version of data splitting to do it. The central point of his argument is given here in case the link to his post should become inoperative:

> I can't agree, however, with the comments...that state that these problems with inference measures imply "never do it." The arguments that inference methods are based on prespecified hypotheses didn't impress me 25 years ago (when I was learning statistics), and they still don't. Nobody *ever* does statistics this way; if we did, we would never identify outliers, look for transformations, enrich the model in response to patterns in residual plots, and so on (all of which also can increase the apparent strength of a regression). Further, I would argue that with the explosion of methods commonly called "data mining," these pieces of advice are ludicrously anachronistic. All subset regression is nothing compared to those kinds of methods. We are no longer in the era of small data sets isolated from each other in time; we are now in one of large (or even massive) ones that are part of an ongoing continuing process. In that context, I would argue that automatic methods are crucial, and the key for statisticians should be to get people to validate their models and correct for selection effects, not tell them what nobody ever believed anyway.

On one level, I've no argument with this stance. I would *qualify it* by saying that activities such as identifying outlier and searching for transformations within a narrow set of options (original or logarithmic scales) are fundamentally different in nature from automatic model fitting procedures because they are done to improve the validity of our models. No one would argue with stopping an outlier from producing a model that failed to fit the bulk of the data, nor would anyone argue for fitting a linear model to highly nonlinear data. The important point is that automatic methods **can** be useful **as long as the model is tested in other data sets**. Unfortunately, too often studies are *not* repeated, if only because there's no glory in it, and the results of automatic model fitting procedures are treated as though they came from validation studies.

I have no doubt that stepwise and "all possible models" procedures can identify gross effects such as the dependence of body weight on caloric intake. However, in practice these procedures are often used to tease out much more complicated and subtle effects. It is these less obvious relationships that, in my experience, are less likely to be reproduced. Saying that these procedures are fine as long as the model is validated may offer false hope in these cases.

Final Comments

It's easy to cheat. When we fit a model to data and report our findings, it is essential to describe how we got the model so that others can judge it properly. **It is impossible to determine from the numerical results whether a set of predictors was specified before data collection or was obtained by using a selection procedure for finding the "best" model.** The parameter estimates and ANOVA tables don't change according to whether or not a variable selection procedure was used. The results are the same as what would have been obtained if that set of predictor variables had been specified in advance.

Perhaps the fundamental problem with automatic methods is that they often substitute for thinking about the problem. As Shayle Searle wrote in Section 1.1, Statistics and Computers, of his *Linear Models For Unbalanced Data*, published in 1987 by John Wiley & Sons, Inc., of New York:

> Statistical computing packages available today do our arithmetic for us in a way that was totally unthinkable thirty years ago. The capacity of today's computers for voluminous arithmetic, the great speed with which it is accomplished, and the low operating cost per unit of arithmetic--these characteristics are such as were totally unimaginable to most statisticians in the late 1950s. Solving equations for a 40-variable regression analysis

> could take six working weeks, using (electric) mechanical desk calculators. No wonder that regression analyses then seldom involved many variables. Today that arithmetic takes no more than ten seconds... But the all-important question would then be: Does such an analysis make sense?
>
> Thinking about such a question is essential to sane usage of statistical computing packages. Indeed, a more fundamental question prior to doing an intended analysis is "Is it sensible to do this analysis?". Consider how the environment in which we contemplate this question has changed as a result of the existence of today's packages. Prior to having high-speed computing, the six weeks that it took for solving the least squares equations for a 40-variable regression analysis had a very salutory effect on planning the analysis. One did not embark on such a task lightly; much forethought would first be given as to whether such voluminous arithmetic would likely be worthwhile or not. Questions about which variables to use would be argued at length: are all forty necessary, or could fewer suffice, and if so, which ones? Thought-provoking questions of this nature were not lightly dismissed. Once the six-week task were to be settled on and begun, there would be no going back; at least not without totally wasting effort up to that point. Inconceivable was any notion of "try these 40 variables, and then a different set of maybe 10, 15 or 20 variables". Yet this is an attitude that can be taken today, because computing facilities (machines and programs) enable the arithmetic to be done in minutes, not weeks, and at very small cost compared to six weeks of human labor. Further; and this is the flash-point for embarking on thoughtless analyses, these computing facilities can be initiated with barely a thought either for the subject-matter of the data being analyzed or for that all-important question "Is this a sensible analysis?"
>
> ...[V]ery minimal (maybe zero) statistical knowledge is needed for getting what can be voluminous and sophisticated arithmetic easily accomplished. But that same minimal knowledge may be woefully inadequate for understanding the computer output, for knowing what it means and how to use it.

If You Need Further Convincing

Everything I've written is true, but I've noticed that many people have trouble fully grasping it. It may seem reasonable that when a program is allowed to pick the best variables, everything will look better than it would if the predictors were picked at random, but the idea often remains an abstraction.

Simulations can make this more concrete. I'm not a Java programmer (I think it's for the best. Otherwise, I'd be doing nothing but programming!), so I don't have any applets to offer. However, I've written some SAS code to illustrate the problem.

The first example looks at whether the intake of various vitamins affects the time it takes to commute to work. One hundred fifty subjects keep a 7 day diary to record their dietary intake and the time it takes to commute to work. In the command language that follows, every pair of variables looks like a sample from a population in which the correlation coefficient is **rho**. Here, rho = 0, so the data are drawn from a population in which none of the variables are associated with each other.

If you paste the command language into SAS, you'll find that forward selection regression with a significance-level-to-enter of 0.05 will select something 54% of the time. That is, at least one vitamin will appear to be associated with commuting time in more than half of the instances when the

program is run, even though these observations are drawn from a population in which no two variables are associated!

[The constants in the variable definitions make the values look more realistic. For example, the commuting times will look like a sample from a normal distribution with a mean of 1 hour and a SD of 15 minutes (= 0.25 hour), the vitamin A values will look like a sample from a normal distribution with a mean of 800 IUs and an SD of 200 IUs, and so on. These adjustments are linear transformation, which have no effect on the correlations between the variables. Someone wanting simpler code and generic variables could change the definitions to
variable_name = rannor(0) + d;
to obtain random values from a normal distribution with a mean of 0.]

```
options ls=80 ps=56;

data analysis;
   rho = 0;
   c = (rho/(1-rho))**0.5;
     do i = 1 to 150;
       d = c * rannor(0);
       commute   =   1 + 0.25 * (rannor(0) + d);
       vit_A     = 800 +  200 * (rannor(0) + d);
       vit_B1    = 1.3 +  0.3 * (rannor(0) + d);
       vit_B2    = 1.7 +  0.4 * (rannor(0) + d);
       vit_B6    = 2.0 +  0.4 * (rannor(0) + d);
       vit_B12   = 2.0 + 0.35 * (rannor(0) + d);
       vit_C     =  55 +   14 * (rannor(0) + d);
       vit_D     =   8 +    2 * (rannor(0) + d);
       vit_E     =   9 +  2.2 * (rannor(0) + d);
       vit_K     =  60 +   12 * (rannor(0) + d);
       calcium   = 800 +  200 * (rannor(0) + d);
       folate    = 190 +   30 * (rannor(0) + d);
       iron      =  12 +    4 * (rannor(0) + d);
       niacin    =  15 +    3 * (rannor(0) + d);
       magnesium = 300 +   50 * (rannor(0) + d);
       potassium =  75 +   10 * (rannor(0) + d);
       zinc      =  13 +    3 * (rannor(0) + d);
       output;
     end;
   keep commute vit_A vit_B1 vit_B2 vit_B6 vit_B12 vit_C vit_D
          vit_E vit_K calcium folate iron magnesium niacin potassium zinc;

proc reg data=analysis;
    model commute = vit_A vit_B1 vit_B2 vit_B6 vit_B12 vit_C vit_D
            vit_E vit_K calcium folate iron magnesium niacin potassium zinc /
            selection=forward sle=0.05 ;
run;
```

SAS PROCs can be placed after the data step to check on the data, for example, to see that the correlation coefficients behave like a sample from a population in which they are all 0.

The second example is even more troublesome because it has some plausibility to it. Also, as you think about what you might do with the results of any one of these "experiments", you'll probably be reminded of a few published reports you've read.

Let the response be *birth weight* instead of commuting time, and let the vitamin variables measure the nutritional status of the baby's mother. We know nutrients are related to each other and it is likely that they will have an effect on birth weight. To reflect this in the data,

1. change all instances of *commute* to *bwt*,

2. let birth weight (in grams) be defined by
   bwt = 2200 + 225 * (rannor(0) + d);
   and
3. change *rho* to 0.50.

Then, data will be observations drawn from a population in which every pair of variables has a correlation of 0.50. We're no longer upset at seeing the predictors (vitamin levels) related to the response (birth weight) because it's now biologically plausible. However, since every pair of variables has the same correlation, the particular variables that enter the forward selection regression equation will be *a matter of chance alone*. This illustrates the danger of using an automated procedure to decide *which* predictors are important.

[back to The Little Handbook of Statistical Practice]
Gerard E. Dallal

Last modified: 05/23/2012 08:23:33.