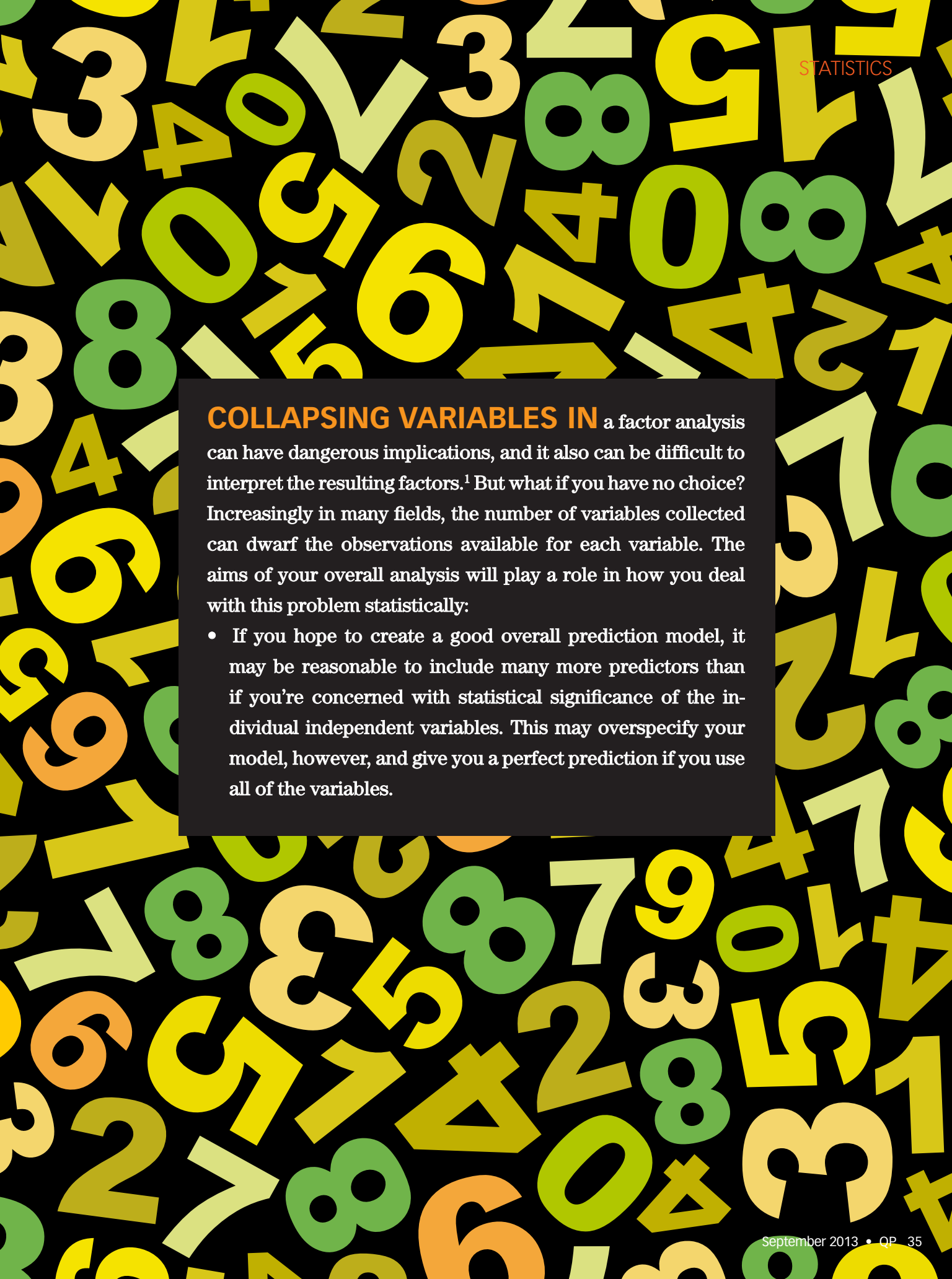
So Many Variables, So Few Observations

by Julia E. Seaman,
Christopher A. Seaman
and I. Elaine Allen

In 50 Words Or Less

- Too many predictors and not enough observations can present problems for a prediction model.
- There are five variable reduction techniques to address this scenario.
- An example illustrates the results of these techniques on the same data set with the same dependent and independent variables.

A look at **five variable reduction techniques to avoid troubles** in a prediction model



COLLAPSING VARIABLES IN a factor analysis can have dangerous implications, and it also can be difficult to interpret the resulting factors.[1] But what if you have no choice? Increasingly in many fields, the number of variables collected can dwarf the observations available for each variable. The aims of your overall analysis will play a role in how you deal with this problem statistically:

- If you hope to create a good overall prediction model, it may be reasonable to include many more predictors than if you're concerned with statistical significance of the individual independent variables. This may overspecify your model, however, and give you a perfect prediction if you use all of the variables.

- If you're testing a specific statistical model in which there may be reasons to include particular predictors, this type of theory-based model testing involves finding the predictors that provide a close connection between your theory and research question. In this case, understanding the relationship between the dependent variable and your independent variables may be most important before reducing the number of independent variables.
- If assessing relative importance of predictors is your aim, you may use a multistep process that includes principle components and factor analysis, among other techniques.[2]

While you may start by examining the bivariate relationship between the predictors and your dependent variable, identifying those predictors that are highly intercorrelated is equally important. It's this scenario in which variable reduction techniques can be most useful.

These large analyses—in which independent variables overwhelm the sample size in modeling—exist across many fields, from science to marketing to business. For example, in analysis of DNA sequence data, microarray data (gene expression) and protein expression data, 100,000 variables or more may be present with a much smaller sample size—in the hundreds, perhaps.[3] In analyzing marketing demand through online data tracking, thousands of variables are available for each unique visit to an organization's website.[4]

When the number of predictors is much greater than the number of observations, there are five statistical methods that can be used. These statistical techniques have been implemented in statistical software such as Statistical Product and Service Solutions (SPSS), Stata, Statistical Analysis System (SAS) and R.

To start, there are two simple approaches based on least-squares regression: stepwise forward regression and best subsets regression. For both methods, start with a model consisting only of the intercept. For both, add the predictor to the model with the smallest p-value (for that reason, all models with just one predictor) and compare p-values.

1. Stepwise forward selection

Add all possible predictors to the model in the last step and expand the model with the one with the smallest p-value or the one increasing the measure of fit (R^2 or standard error of the estimate).

Continue until some stopping criterion is met (that is, no increase in the R^2), and use the model with the highest R^2 or smallest standard error of the estimate (SEE) as your model.

2. Best subsets regression

Keep the best one-variable model using R^2. Compute all least-squares regression models with all possible two-variable models. Keep the best two-variable model. Compute all least-squares regression models with all possible three-variable models. Keep the best three-variable model.

Summary of overall results / TABLE 1

	Number of variables	Best fit (R^2)
Stepwise regression	9	0.988
Best subsets regression	9	0.921
Principal components analysis and regression	13 (in 3 factors)	0.897
Ridge regression	67 (weight = 0.05)	0.997
LASSO analysis	15	0.959

LASSO = least absolute shrinkage and selection operator

Stepwise regression results / TABLE 2

	Unstandardized coefficients		
	B	Std. error	p-value
Median home price: 2002	1.015	0.021	0.000
Percentage one-year increase	259368.274	17260.934	0.000
Students per computer	933.124	279.620	0.001
Library books per capita	1194.093	296.471	0.000
Colon/rectal-SIR	168.325	48.747	0.001
Property tax rate	–2250.794	529.125	0.000
Average tax bill	7.785	1.933	0.000
Average water bill	–25.425	8.284	0.003
MCAS SciTech—8th grade	–181.720	68.074	0.009
R^2 = 0.988			

LASSO = least absolute shrinkage and selection operator
MCAS SciTech = Massachusetts Comprehensive Assessment System science/technology
SIR = standardized incidence ratio
Std. = standard

Continue until all combinations of models with all variables are fit. You may want to limit this to some number $k < n < p$, in which n is the sample size and p is the number of variables available. Rank the fitted models by highest to lowest R^2, and choose the model with the highest value. For models with similar R^2 values, choose the model with the smallest number of variables.

The advantages of these first two approaches are that they're both easy to implement, automatic algorithms exist and they always produce a result. A disadvantage of both approaches is that stopping rules can be unstable because a small perturbation of the data can lead to different results. If independent variables are highly correlated, the resulting model will include only one of the variables. So you may miss the "best" model, making the results difficult to interpret. It also can be time consuming when building a best subsets model.

Overall, the stepwise regression and best subsets regression methods are often included as automated functions in many statistical packages. These packages often include options for changing the stopping criterion (R^2, standard error of the estimate, F-test or p-value) and the number of iterations to find a solution. The resulting model is the most optimized for the data, but it's difficult to interpret the relationship of predictors included and the outcome or the relationship between the predictors.

These methods are optimized for the data set, but there is a chance the model will pick unimportant features that are unique to the modeling data set and not important to an overall inferential model. Therefore, it may fail on new data sets.

One way to combat this issue is to have a training data subset to build the model, and a smaller testing data subset to assess how well the model predicts other data. With data being sparse compared to variables, however, this is not often possible.

Both stepwise and best subsets regression are unbiased methods and do not include any previous information of how variables are known to interact. This makes understanding the outcome difficult, but it is an easy way to build a prediction model.

3. Principle components with regression

Perform principal components analysis (PCA), or a factor analysis, on the design matrix X (covariance or correlation matrix) to create a set of smaller, uncorrelated and new variables:

- PCA will give you a new design matrix Z.
- Use the first $k < n < p$ total variables as your new variables based on the plot of the eigenvalues by component number, called a Scree plot, or the amount of variance explained.
- Perform an ordinary linear regression with the new variables.

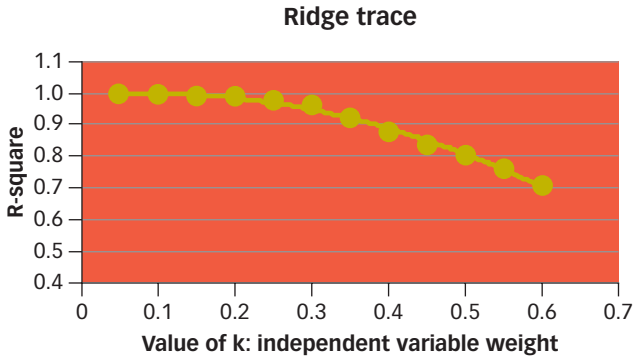
The advantage of the new design matrix Z is that it's orthogonal (by construction), generally reduces model size, and resulting components can usually be interpreted. The disadvantage is that you have not used the dependent variable when reducing the number of independent variables when using PCA. The resulting factors are orthogonal to each other, but you don't know how strong their relationship is to Y.

It is not always clear how many of the new factors you should use in the final model. This is also a two-stage method in which the factors are identified first and then used in the linear regression for prediction.

Similar to factor analysis, PCA is a transformation of the original, potentially correlated and independent variables into new variables, called factors, as a linear combination of the original variables. These new variables are used to perform the regression analysis to create a model.

Because PCA is carried out only on the independent variables without knowing the strength of their relationship to the dependent variable, examining the correlation with Y and only including independent variables with a strong relationship to Y in the PCA is

Ridge trace plot of Boston towns' data / FIGURE 1



Truncated output from subset regression showing change in R^2 and Mallow's C_p / TABLE 3

Number of variables	R-squared (R^2)	Mallow's C_p	Standard error of the estimate	Town population	Housing cost—one-year increase	Housing cost—five-year increase	Tax rate	Average tax bill	Car insurance cost	Water and sewage bill	Distance from Boston	Cost per pupil	Average teacher salary	Students per teacher	4th grade math MCAS	4th grade English MCAS	8th grade math MCAS	10th grade science MCAS	10th grade math MCAS	10th grade English MCAS	Total SAT score	Percentage of students taking SATs
1	78.2	136.8	57,015					X														
1	49.5	451	86,771														X					
2	87.3	39.1	43,736				X	X														
2	82.6	90.2	51,131					X					X									
3	89.3	18.8	40,260	X			X	X														
3	89.2	20.3	40,532				X	X					X									
4	90	13.9	39,249				X	X					X									
4	90	14.1	39,291	X			X	X			X											
5	90.4	10.9	38,533	X			X	X			X		X									
5	90.4	11.7	38,695	X			X	X	X		X											
6	90.8	8.7	37,953	X	X		X	X			X		X									
6	90.8	9.3	38,066	X			X	X	X		X		X									
7	91.2	7	37,433	X	X		X	X	X		X		X									
7	91.1	7.5	37,532	X	X		X	X	X		X	X										
8	91.4	5.8	36,996	X	X		X	X	X		X		X		X							
8	91.4	6	37,047	X	X		X	X	X		X	X	X									
9	91.6	6	36,840	X	X		X	X	X		X	X	X		X							
9	91.6	6.3	36,888	X	X		X	X	X		X		X		X							X
10	91.8	6	36,628	X	X		X	X	X		X	X	X		X							X
10	91.8	6.1	36,639	X	X		X	X	X		X	X	X		X		X					
11	91.9	7.1	36,631	X	X		X	X	X		X	X	X		X		X					X
11	91.9	7.1	36,634	X	X	X	X	X	X		X	X	X		X							X
12	92	7.9	36,585	X	X	X	X	X	X		X	X	X		X		X					X
12	92	8	36,610	X	X		X	X	X		X	X	X		X		X		X			X
13	92.1	8.8	36,557	X	X	X	X	X	X		X	X	X	X	X		X					X
13	92.1	9	36,586	X	X	X	X	X	X		X	X	X		X		X		X			X
14	92.2	9.9	36,565	X	X	X	X	X	X		X	X	X	X	X		X		X			X
14	92.1	10.2	36,631	X	X	X	X	X	X		X	X	X	X	X		X				X	X
15	92.2	11.3	36,636	X	X	X	X	X	X		X	X	X	X	X		X		X		X	X
15	92.2	11.9	36,757	X	X	X	X	X	X		X	X	X	X	X		X		X	X		X
16	92.2	13.2	36,826	X	X	X	X	X	X		X	X	X	X	X		X	X	X		X	X
16	92.2	13.3	36,834	X	X	X	X	X	X	X	X	X	X	X	X		X		X		X	X
17	92.2	15.2	37,019	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X		X	X
17	92.2	15.2	37,025	X	X	X	X	X	X		X	X	X	X	X		X	X	X	X	X	X
18	92.2	17.1	37,207	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X	X	X
18	92.2	17.1	37,225	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X
19	92.2	19	37,412	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

MCAS = Massachusetts Comprehensive Assessment System

SAT = Scholastic Aptitude Test

one possible extension of the method.

Additionally, it's good to use a subset of the data as a testing set to assess the model's accuracy and error. A great benefit of PCA is its ability to reduce many collinear variables into only a few significant variables for the model. Significant variables are determined by finding the inflection point (where the line flattens) on a Scree plot.

The new PCA variables may seem more complex because they are combinations of the original data, but often they can hint at underlying associations of the variables. In a model trying to predict heart attacks, for example, the variables of height and weight are combined positively into a single, strongly significant component. This is not surprising because obesity (a high weight-to-height ratio) is a major predictor of heart attack risk.

PCA is a useful technique for reducing the number of variables—especially when there is a known correlation among the variables—and can be a quick method to develop an optimized model.

4. Ridge regression

Ridge regression shrinks the regression coefficients by adding a penalty to the least-squares criterion. It does this by penalizing the residual sum of squares by adding a weight times the sum of the values of the squared regression coefficients where large coefficients are down-weighted. This reduces the overall variability of the prediction model and is a trade-off between goodness of fit (residual sum of squares) and a penalty:

1. The usual residual sum of squares = minimum ($\Sigma[Y - X\beta]^2$) or the square of observed minus the predicted values of the dependent variable.
2. The coefficient penalty = $\lambda\Sigma(\beta^2)$ is added to the residual sum of squares with a calculated weight λ that shrinks large values of the coefficients.
3. For $\lambda = 0$, this reduces to the usual least-squares fit. For $\lambda = \infty$, all of the coefficients are shrunk to equal zero. By varying λ, a plot called the ridge trace can be used to identify where the coefficients become stable (see example in Figure 1, p. 37).

The advantage of ridge regression is that it's capable of reducing the variability and improving the accuracy of linear regression models. These gains are largest in the presence of multicollinearity among the independent variables.

Variables added using the LASSO technique / TABLE 4

Step	C_p	R^2	Variable added
1	448.891	0.000	
2	407.044	0.074	+sat
3	148.272	0.510	+culture
4	52.070	0.675	+perpupil
5	31.854	0.712	+pctcoll4
6	32.986	0.714	+popdens
7	32.926	0.717	+polfire
8	31.745	0.722	+vote
9	27.533	0.753	+minority
10	25.526	0.779	+hsgradpc
11	19.618	0.803	+intclass
12	13.130	0.827	+cardeath
13	10.749	0.844	+pctpover
14	12.174	0.875	+unemp
15	13.369	0.897	+incr5
16	15.190	0.927	+popn
17	17.000	0.959	+poldense

LASSO = least absolute shrinkage and selection operator

What ridge regression doesn't do is variable selection, and it fails to take into account the fact that a parsimonious model with few parameters can't zero out coefficients; thus, you either end up including all the coefficients in the model or none of them.

There is no rule for identifying the appropriate λ and, while algorithms for ridge regression are included in major statistical software packages, the ridge regression method is not as automatic to apply as least-squares regression.

Ridge regression presents one option for reducing the variable weights, but not the number of variables. It produces a stable regression—in contrast to least squares alone—but leaves all the independent variables in the model, although many now have almost zero weight. Evaluation of the models is done by examining the ridge trace plot.

5. LASSO regression

This method is similar to ridge regression because it penalizes the regression coefficients, but it uses the absolute size of the coefficients rather than the square of

the coefficients to compute the penalty weight.[5]

By penalizing—or constraining the sum of the absolute values of the estimates—in the least absolute shrinkage and selection operator (LASSO) method, you end up with some of the parameter estimates at exactly zero, so you will reduce the number of independent variables in the model. The larger the penalty, the more the coefficient estimates are shrunk toward zero:

- The usual residual sum of squares = minimum ($\Sigma[Y - X\beta]^2$) or the square of observed minus predicted values of the dependent variable.
- Coefficient penalty = $\lambda\Sigma(||\beta||)$, in which $||\beta||$ is the absolute value of the coefficient, is added to the residual sum of squares with a calculated weight λ that shrinks large values of the coefficients.
- As the weight gets larger, the number of independent variables decreases, as shown in Figure 1.

In contrast to ridge regression, the LASSO method automatically does parameter shrinkage and variable selection. For large sample sizes and a large number of variables, the result approximates the least-squares solution. The method also works well with a large number of variables that are not multicollinear, as well as in the presence of multicollinearity.

The method and results, however, may change for different methods of scaling the data. The method is designed for working with standardized variables using the correlation matrix. LASSO algorithms are extremely fast compared with subset regression or PCA.

Designed to be computationally fast for large data sets that are accompanied by a large number of variables, LASSO provides an automatic way to scale and select your variables in many cases—including sparse matrixes. The fit of the models can be examined in a plot similar to the ridge trace.

Single data set example

A data set that contains multiple aspects about the towns surrounding Boston can help illustrate each technique. The outcome variable is the prediction of median home price in 2004 based on all the other variables.

The data set has 67 variables for 61 towns with no missing data and 89 towns with partial data. For some techniques, the full data set can be included for estimation, while other methods can use only complete data. All five techniques were implemented in SPSS, and the ridge regression and LASSO methods were validated in Stata because the syntax for the SPSS analyses aren't part of the main statistics component.

Table 1 (p. 36) summarizes the model-fitting results for all five techniques, with stepwise and subsets regression shown as separate models. In terms of R^2 values, stepwise and ridge regressions give the highest values (Table 2, p. 36, and Figure

Results of principal components analysis and regression / TABLE 5

	Unstandardized coefficients		
	B	Std. error	Sig.
(Constant)	400477.740	7542.081	0.000
Factor one—public safety	103431.549	7567.348	0.000
Factor two—health	42083.739	7567.348	0.000
Factor three—environment	24804.289	7567.348	0.001
$R^2 = 0.897$			

Factor variables
Public safety:
Violent crimes per capita
Structure fires per capita
Motor vehicle deaths per capita
Public spending per capita
Health:
Incidence per 100,000 of HIV
Incidence per capita of sexually transmitted diseases
Standardized incidence ratio of cancer
Heart disease deaths per capita
Overall death rate
Environment:
Percentage open space
Air pollution sources per square mile
Number of contaminated sites per square mile
Presence of radon

Sig. = significance
Std. = standard

In contrast to ridge regression, the **LASSO method automatically** does parameter shrinkage and variable selection.

1). Ridge regression includes all of the variables, albeit with different weights, so a higher R^2 is expected.

Stepwise regression includes all the possible variables showing an increase in R^2 and, similar to LASSO, has a stopping algorithm based on either increasing R^2 or decreasing SEE or Mallow's C_p, which is a goodness of fit statistics that incorporates the sample size and number of variables. A small value of C_p accompanying a large R^2 is desired.

Best subsets regression (Table 3, p. 38) and LASSO (Table 4, p. 39) give the next best fit of the model. Both use algorithms to find the smallest number of variables giving the best fit. With best subsets regression, all possible model combinations are fit and compared. The user picks the best model usually based on the largest R^2 combined with the smallest C_p.

With LASSO, the algorithm seeks to minimize any multicollinearity in the independent variables, and the output includes a table showing the effect of adding additional variables and a plot of the variables by importance. The optimal variables are shown by those that increase R^2 and decrease C_p.

Principal components and regression (Table 5) show the smallest R^2 and are the most difficult to implement because of decisions regarding how many factors to include, which variables are associated with which factors, and whether these factors can be easily interpreted. Having the worst fit of all the techniques, this also illustrates that this form of analysis does not include the dependent variable as a criterion for the calculation of factors.

Depending on the data set

None of the results from the techniques were identical to one another, although similar models were found between pairs of methods: stepwise regression and ridge regression, and the subsets regression and LASSO techniques.

The diversity in the fit of the models illustrates that some modeling techniques are better when prediction is the goal, while others are best when identifying uncorrelated or less-correlated predictors.

The best method with the best fit will change, however, depending on the data set, the relationships between independent variables, and the relationship between the independent and dependent variables. QP

REFERENCES AND NOTES

1. Julia E. Seaman and I. Elaine Allen, "Words of Caution," *Quality Progress,* July 2012, pp. 48-50.
2. An interesting perspective on choosing the best predictors is discussed by Steven M. Shugan in "The Anna Karenina Bias: Which Variables to Observe," *Marketing Science,* Vol. 26, No. 2, 2007, pp. 145-148. Shugan shows that those variables that are highly related to only part of the range of your dependent variable may seriously bias your results.
3. Sijian Wang and Ji Zhu, "Variable Selection for Model-Based High-Dimensional Clustering," *Biometrics,* Vol. 64, 2008, pp. 440-448.
4. Moez Hababou, Alec Y. Cheng and Ray Falk, "Variable Selection in the Credit Card Industry," proceedings from the New England SAS Users Group, 2006.
5. For more information about LASSO, see Robert Tibshirani's "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society,* Series B., Vol. 58, No. 1, 1996, pp. 267-288.

BIBLIOGRAPHY

Sogawa, Yasuhiro, Shohei Shimizu, Aapo Hyvärinen, Takashi Washio and Seiya Imoto, "Discovery of Exogenous Variables in Data With More Variables Than Observations," *Artificial Neural Networks,* Vol. 6352, 2010, pp. 67-76.

Sundberg, Rolf, and Philip J. Brown, "Multivariate Calibration With More Variables Than Observations," *Technometrics*, Vol. 31, No. 3, 1989, pp. 365-371.

JULIA E. SEAMAN is a doctoral student in pharmacogenomics at the University of California, San Francisco, and is a statistical consultant for Quahog Research Group in San Francisco. Seaman earned a bachelor's degree in chemistry and mathematics from Pomona College in Claremont, CA.

CHRISTOPHER A. SEAMAN is a data scientist at Atlassian Software in San Francisco and a statistical consultant for the Quahog Research Group. He is a doctoral student in mathematics and computer science and has a master's degree in mathematics from the Graduate Center of the City University of New York.

I. ELAINE ALLEN is professor of biostatistics and epidemiology at the University of California, San Francisco, and emeritus professor of statistics and entrepreneurship at Babson College in Wellesley, MA. She is cofounder of the Quahog Research Group. Allen earned a doctorate in statistics from Cornell University in Ithaca, NY, and is a member of ASQ.