Honest Confidence Intervals for the Error Variance in Stepwise Regression

Dean P. Foster and Robert A. Stine

Department of Statistics

The Wharton School of the University of Pennsylvania

Philadelphia, PA 19104-6340

March 15, 2006

Abstract

An honest confidence interval for the error variance in a stepwise regression is a one-sided interval that adjusts for the effects of variable selection. The endpoint of this interval may be many times larger than the usual endpoint. Such adjustments are most important when selecting variables from a large number of available predictors, particularly in situations with more available predictors than observations. An illustration using a regression model of stock market returns illustrates the calculations.

Key Phrases: Bonferroni inequality, model selection, selection bias.

1 Introduction

Stepwise regression is known for its ability to overfit data. Suppose that we wish to build a regression model based on n independent observations of a response variable Y and a large set of p potentially useful predictors $X_1, X_2, \ldots, X_p$. Virtually any statistics package will allow us to fit a sequence of approximating models of the form

$$\hat{Y}_k = \hat{\beta}_0 + \hat{\beta}_1 X_{j_1} + \cdots + \hat{\beta}_k X_{j_k} \ .$$

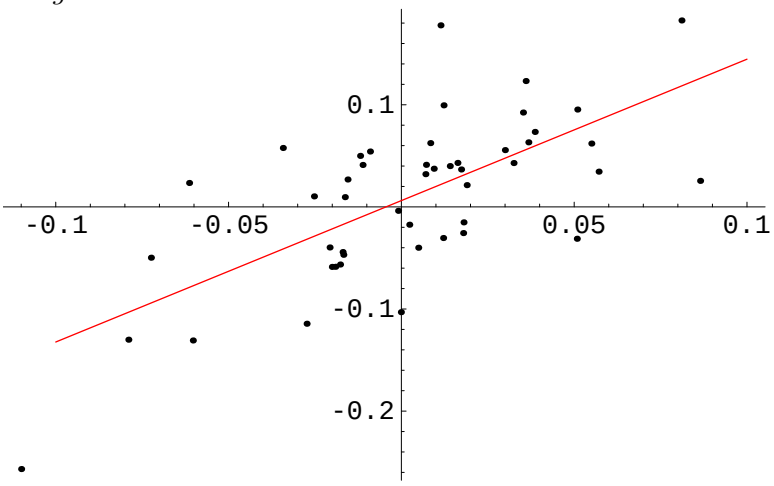
For each choice of k, the chosen model ideally minimizes the sum of squared residuals,

$$\text{Residual SS}_k = \sum_i (Y_i - \hat{Y}_{i,k})^2 \ ,$$

among all models with k predictors. Computational demands, however, limit the use of such best-subset regression models to small values of p, and generally simpler stepwise alternatives are used. These simpler algorithms (e.g., forward or backward stepwise regression) obtain comparable residual sums of squares unless the data possess certain forms of collinearity that are hard for greedy algorithms to recognize [1,10]. Whichever algorithm is used, the fitted model typically identifies numerous, apparently significant effects even when there is no association between the response and predictors. The optimism of the fitted model contrasts with poor out-of-sample prediction; models obtained by stepwise selection can fit quite well in-sample yet predict poorly when applied to new data.

This tendency of stepwise regression to overfit grows with the number of available predictors, particularly once $p > n$. Situations with more predictors than observations are common in financial modeling as illustrated below, but also occur in the physical sciences. For example, the expense of direct measurements has led to the use of near infrared spectroscopy (NIR) to assess the chemical composition of, for example, foods and waste products. In these analyses, the absorption of light at many (infrared) frequencies predicts the concentration of, say, protein in wheat or dioxin in smoke [9]. Miller [10] notes an application of NIR spectroscopy with 757 such predictors but only 42 observations. Another context rich in predictors is the analysis of weather patterns; a host of factors is available for predicting the chance of rainfall on a given day. Attempts to capture nonlinearities with interactions and polynomial terms further increase the number of predictors.

Figure 1: *The scatterplot of the monthly returns on McDonald's stock versus returns on the S&P 500 index during 2002–2005.*



For a financial illustration, we offer a regression model for the sequence of monthly returns on McDonald's stock from January, 2002 through December, 2005. Financial theory suggests that returns on McDonald's stock ought to be associated with contemporaneous returns on the stock market, and indeed this is the case. Figure 1 plots the monthly returns of McDonald's stock on the returns on the S&P500 index; the correlation is about 0.68. This is a well-known relationship, we are going to need to look elsewhere to "beat the market." The lure of discovering special features of the market is so powerful that some are tempted to try virtually any predictor that might lead to a money-making scheme. Table 1 summarizes the fitted coefficients of one such model; the 17 predictors were chosen from a set which includes the return on the S&P500 and 50 additional factors, labeled $X_1, \ldots, X_{50}$. We first performed forward stepwise regression using the criterion p-to-enter = 0.25, followed by backward stepwise with p-to-remove = 0.10. This two-step process (fit an initial model and then remove the insignificant terms) is similar to the approach studied in [7], but there $p < n$. The overall significance of the fit is impressive ($R^2 = 0.910$ and $F_{17,30} = 17.806$ with $p < 0.0001$). Perhaps more impressive (at least for those who have not read Freedman's paper), many of the individual p-values are quite significant, with six having p-values less than 0.0001. These are small p-values, even compared to the traditional Bonferroni threshold $0.05/51 = 0.001$.

We have found examples like this one to be useful in conveying the dangers of

Table 1: *Coefficients of a stepwise regression model for monthly returns of McDonald's stock. The overall $R^2 = 0.910$ ($F_{17,30} = 17.806$, $p < 0.0001$) and the residual variance $s_{17}^2 = 0.0298^2$.*

Term	Estimate	Std Error	t Ratio	p-value
Constant	0.0166	0.0061	2.72	0.0107
S&P 500	0.6863	0.1390	4.94	0.0000
X_4	-0.0147	0.0060	-2.45	0.0204
X_8	0.0129	0.0054	2.40	0.0230
X_{11}	-0.0185	0.0061	-3.04	0.0049
X_{19}	-0.0133	0.0050	-2.68	0.0119
X_{22}	0.0215	0.0058	3.69	0.0009
X_{28}	-0.0141	0.0059	-2.41	0.0223
X_{31}	-0.0155	0.0060	-2.57	0.0155
X_{33}	-0.0118	0.0050	-2.39	0.0233
X_{34}	0.0339	0.0058	5.83	0.0000
X_{35}	-0.0150	0.0045	-3.31	0.0024
X_{36}	0.0272	0.0055	4.99	0.0000
X_{37}	-0.0416	0.0053	-7.79	0.0000
X_{39}	0.0317	0.0051	6.26	0.0000
X_{44}	0.0293	0.0071	4.15	0.0003
X_{46}	-0.0352	0.0055	-6.46	0.0000
X_{48}	-0.0193	0.0059	-3.28	0.0026

overfitting to MBA students [6]. The example is compelling because, despite the great fit, each of the 50 additional predictors aside from returns on the S&P500 consists of independent Gaussian random noise. Each X_j is a sample of 48 observations from a standard normal distribution, simulated independently of one another and the response.

To cope with the problems brought on by variable selection, we focus on the familiar residual variance estimator

$$s_k^2 = \frac{\text{Residual SS}_k}{n-k-1} \; .$$

Because of the selection process, s_k^2 is biased and can grossly overstate the fit of the model. Berk [1] simulated this bias in problems with relatively few (between 4 and 15) predictors and $p < n/2$. Even in these problems, the bias of s_k^2 is as large as 25%, and the problem becomes much more severe as p/n increases. Simple adjustments to s_k^2 derived here conservatively allow for the effects of model selection and yield a one-sided confidence interval that compensates for overfitting.

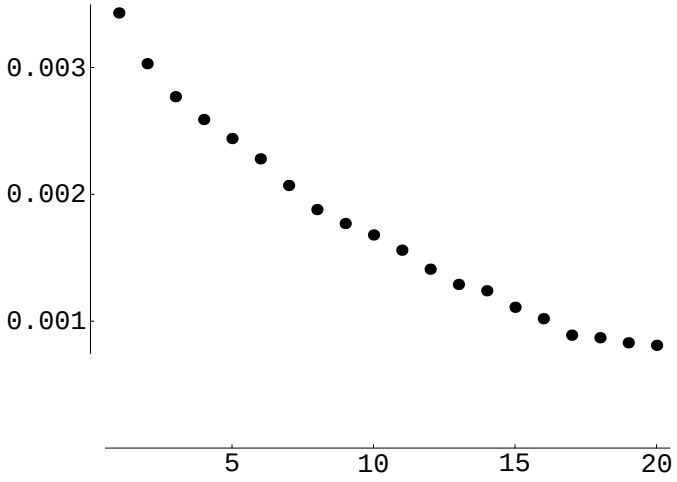
Concerns about overfitting in stepwise regression are not new and have been studied virtually since the introduction of this algorithm. For example, Draper and Smith [4, Chapter 6] followed their overview of stepwise methods with a discussion of overfitting that includes informal graphical methods for deciding when enough variables have been added. They suggested a "scree test" procedure (see their Figure 6.2) which identifies the appropriate choice for k by noting where the graph of s_k^2 on k flattens. They cautioned, though, that this procedure is suitable only when one has many more observations than included predictors. In our example with $p > n$, s_k^2 decreases steadily as k grows as shown in Figure 2.

More formal procedures for guarding against overfitting have considered the null distribution of R^2 under stepwise selection. Extending the results of Diehr and Hoflin [2], Rencher and Pun [11] simulated stepwise regression in a few models with p slightly larger than n (e.g., choosing $k = 4$ out of $p = 40$ with $n = 10$). Rencher and Pun also approximated the distribution of R^2 treating the $m = \binom{p}{k}$ possible R^2 statistics for a given k as a sample from a beta distribution. This approach leads to the approximate $1-\alpha$ upper critical value

$$\tilde{R}_\alpha^2 = B^{-1}\left(1 + \frac{\log \alpha}{m}\right) \; , \tag{1}$$

where B denotes the beta distribution with parameters $k/2$ and $(n-k-1)/2$. They

Figure 2: *The estimated error variance s_k^2 decreases steadily as the number of variables k in the stepwise model increases.*



improved this approximation by using their simulation results to adjust for dependence among the collection of m R^2 values, obtaining

$$R_\alpha^2 = B^{-1}\left(1 + \frac{\log \alpha}{(\log m)^{1.8m^{0.04}}}\right) . \tag{2}$$

The adjustment in effect reduces the number of possible models from m in (1) down to $(\log m)^{am^b}$, with a and b determined empirically. For the stock market example ($k = 17$, $p = 51$), this expression gives the critical value $R_{0.05}^2 = 0.906$, indicating that this model has not explained significant variation. Alternatively, others have proposed stopping rules for halting the selection process [3,5]. Neither these rules nor measures of the inflated size of R^2 appear in the output of standard stepwise software.

Our approach differs from these in several respects. First, we allow the model to have significant effects and adjust an interval for s_k^2 rather than present a critical value for the null case. Our method allows signal in the fitted model and places an upper confidence limit on the error variance given the fitted model rather than bound R^2 in the null case. Also, rather than the exception, our interest centers on models with as many or more predictors than observations. Finally, we offer a very simple approximate expression for the adjustments to s_k^2.

2 An Honest Confidence Interval for σ^2

The goal of this section is to produce a confidence interval for the error variance σ^2 that holds up under model selection. We first require a model that defines σ^2. We assume that the response vector Y is normal with arbitrary mean vector η and constant variance σ^2, $Y \sim N(\eta, \sigma^2 I_n)$ or

$$Y = \eta + \epsilon, \quad \epsilon_i \sim N(0, \sigma^2) \ ,$$

where the deviations ϵ_i are independent. Given that we allow $p \geq n$ (and indeed focus on this context), our interval is one-sided. With $p \geq n$, a perfect fit is possible so that the appropriate lower bound is zero. The challenge is to find an upper bound that implies a guaranteed level of fit. Since the mean η is unlikely to lie in the column span of $k < n$ chosen predictors, the resulting projection error inflates s_k^2. This effect works in the opposite direction of selection bias which leads to optimistically small estimates of error variation. Such lack of fit makes the selection-adjusted interval conservative in that the bounds are only so good as the set of predictors allows.

The usual confidence interval for σ^2 ignores the selection process. For the fitted model in Table 1, the residual standard error (or root mean squared error) is estimated as

$$s_k = \sqrt{\frac{\text{Residual SS}}{n-k-1}} = \sqrt{\frac{0.0266}{30}} = 0.0298 \ .$$

In the usual analysis, $(n-k-1)s_k^2/\sigma^2 \sim \chi^2_{n-k-1}$ so that

$$\mathrm{P}\left\{\sigma^2 \leq \frac{(n-k-1)s_k^2}{\chi^2_{n-k-1,0.05}}\right\} = 0.95 \ , \tag{3}$$

where $\chi^2_{d,\alpha}$ is the α quantile of a chi-squared distribution with d degrees of freedom. It follows that the 95% upper confidence limit for σ in the example is

$$\sigma \leq \sqrt{\frac{0.0266}{18.49}} = 0.038 \ .$$

In fact, this endpoint ought to be about three and a half times larger.

To adjust for variable selection, we begin with the assumption that the stepwise process has found the best fitting model from the set of $m = \binom{p}{k}$ possible models with k regressors. As a result,

$$\frac{(n-k-1)s_k^2}{\sigma^2} = \min(z_1^2, \ldots, z_m^2) \ ,$$

where each $z_i^2 = RSS_i/\sigma^2$ is the normalized residual sum of squares for the ith model, with $i = 1, \ldots, m$. In general, the z_i^2 are dependent, and each is distributed as a noncentral χ^2_{n-k-1} random variable. Noncentrality arises since some fits exclude important predictors (if in fact any are useful).

The Bonferroni inequality provides a critical value that allows for the minimization, dependence, and noncentrality. We need to replace the usual quantile $\chi^2_{n-k-1,\alpha}$ by a value C such that

$$\begin{aligned} 1-\alpha &\leq \mathrm{P}\{\frac{(n-k-1)s_k^2}{\sigma^2} \geq C\} \\ &= \mathrm{P}\{\min(z_1^2, \ldots, z_m^2) \geq C\} \\ &= \mathrm{P}\{z_1^2 \geq C \cap \cdots \cap z_m^2 \geq C\} \\ &= 1 - \mathrm{P}\{z_1^2 < C \cup \cdots \cup z_m^2 < C\} \end{aligned} \tag{4}$$

The inequality (4) obtains if we bound

$$\mathrm{P}\{z_1^2 < C \cup \cdots \cup z_m^2 < C\} \leq \sum_{i=1}^{m} \mathrm{P}\{z_i^2 < C\} \leq \alpha \,, \tag{5}$$

and simply choose C such that for each marginal probability

$$\mathrm{P}\{z_i^2 < C\} \leq \frac{\alpha}{m} \,. \tag{6}$$

The α/m quantile from the extreme left tail of the χ^2_{n-k-1} density meets these needs, and we set

$$C = \chi^2_{n-k-1,\alpha/m} \,. \tag{7}$$

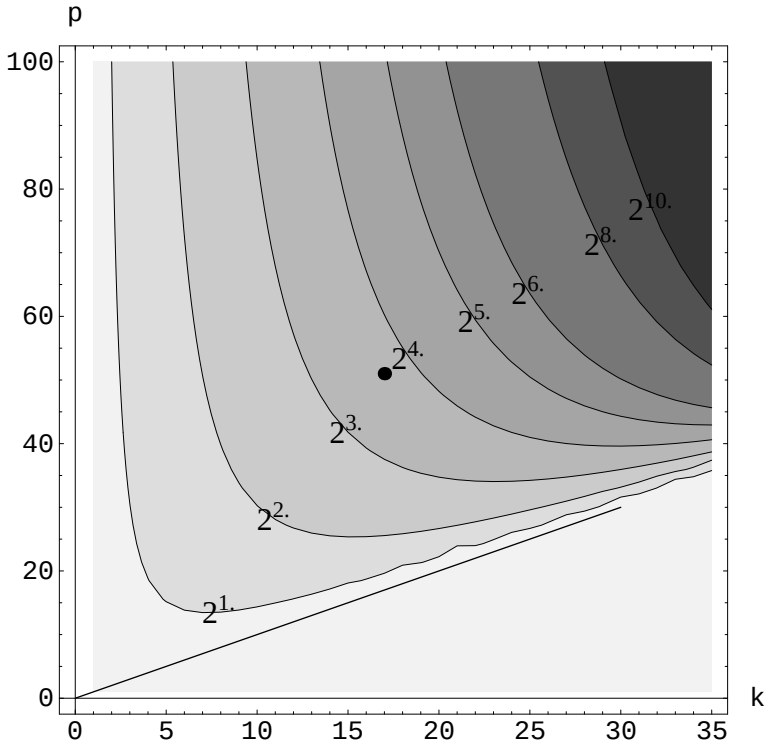
This choice also conservatively handles noncentrality. Since this choice for C satisfies (6) for a central χ^2 variate, this inequality also holds for noncentral χ^2's since noncentrality increases z_i^2. We term the resulting interval,

$$[0, \; \frac{(n-k-1)s_k^2}{\chi^2_{n-k-1,0.05/m}}]$$

an honest confidence interval for σ^2.

The adjusted quantiles that determine the honest interval change dramatically with k. The quantile $\chi^2_{n-k-1,\alpha}$ found in the usual interval is hardly affected by whether we choose $k = 5$, 10, or 15 predictors when $n = p = 50$, ranging from 29.8 down to 21.7. In contrast, the adjusted critical values with α replaced by α/m fall from 7.7 to 2.1. The contour plot in Figure 3 offers another view of this effect. The usual upper limit of

Figure 3: *Contour plot of the ratio of chi-squared critical values, $\chi^2_{n-k-1,\alpha}/\chi^2_{n-k-1,\alpha/m}$, for varying k and p with $\alpha = 0.05$ and $n = 48$. The point locates the stock market example. To convert the usual one-sided interval for σ^2 to an honest interval, multiply the endpoint by the value shown.*



the confidence interval for σ^2 is s_k^2 times $d/\chi^2_{d,\alpha}$. Adjusting for selection bias increases the latter factor to $d/\chi^2_{d,\alpha/m}$ by changing the tail probability. Figure 3 shows the ratio of critical values $\chi^2_{30,\alpha}/\chi^2_{30,\alpha/m}$ for various values of $1 \le k \le 35$ and $k \le p \le 100$ with $n = 48$ observations. The point in the figure locates the stock market example, and the diagonal line is a reminder that $k \le p$. For $k < 5$, there is relatively little selection bias since the ratio is about 2 or 3. As k grows, however, the selection effect increases rapidly.

For the illustrative model, the adjustment for selection bias makes the endpoint of the honest confidence interval for σ about three and a half times larger than that of the usual interval. Though we did not use best-subsets regression, we will assume that the best fitting model with 17 predictors is chosen from the collection of $m = \binom{51}{17}$ possible

models. The adjusted critical value (computed with *Mathematica*) is then

$$C = \chi^2_{30,\alpha/m} \approx 1.46 \ ,$$

compared to the usual critical value 18.49 at the 0.05 quantile. The upper limit for the resulting 95% confidence interval for σ is

$$\sigma \leq \sqrt{\frac{0.0266}{1.46}} = 0.135 \ ,$$

compared to the unadjusted endpoint 0.038.

Adjusting for selection as we have anticipates the lack of predictive power of the model. The addition of so many random predictors to the fit degrades the model's ability to predict future returns. For example, conditioning on $\hat{\beta}$ at the values in Table 1, the 16 random predictors can be expected to add $\sum_{j=2}^{17} \hat{\beta}^2 = 0.096^2$ to the mean squared error of prediction, much more variation than the usual limit accommodates. A simple model using a constant alone would be preferable since the observed standard deviation of McDonalds return during these four years is only 0.079.

3 Simulation Evidence

The proposed interval is conservative, possessing a larger upper endpoint than needed for the nominal coverage. How conservative? As one might suspect, the size of the excess coverage depends on the conditions of the model. In terms of the noncentrality, we believe that stepwise methods are most common (and appropriate) in problems characterized by substantial noise and relatively few meaningful predictors. In this setting, most of the z_i^2 are roughly central χ^2 random variables. As to the use of the Bonferroni inequality in (5), for m independent events with probabilities α_i, this inequality bounds

$$1 - \prod_{i=1}^{m}(1-\alpha_i) = \sum_i \alpha_i - \sum_{i<j} \alpha_i\alpha_j + \cdots \leq \sum_i \alpha_i \ .$$

In our case, $\alpha_i = \alpha/m$ and the error from using the Bonferroni inequality is on the order of $\alpha^2 = 0.0025$ and not of great concern. However, the z_i^2 assess overlapping subsets of predictors and are dependent. Bounds that ignore this dependence are conservative since one has not maximized over so many independent events. The

numerical adjustments of [11] that produce (2) account for some of these effects. With $p > n$, however, there is less overlap and less dependence.

We ran a small simulation to see how well the adjusted quantiles track the distribution of s_k^2 in stepwise regression. We used stepwise regression to select models for varying choices of k from a set of $p = n$ and $p = 4n$ random predictors. Comparison boxplots in Figure 4 summarize the observed distribution of the residual sums of squares of models fit with $n = 50$. In addition, the curve in each frame locates the quantile $\chi^2_{n-k-1,.05/m}$. The gap between the boxplots and $\chi^2_{n-k-1,.05/m}$ measure how conservative the intervals are. Since the boxplots lie above these points for all but the largest values of k, our procedure is conservative unless one is fitting many predictors. For example, with $k = 40$ and $n = p = 50$ about 1% of the residual sums of squares are below the $\chi^2_{9,.05/m}$ quantile.

Here are the details of the simulation. For each of the 500 trials in the simulation, the predictors and response are independent standard normal samples. We ran a simple variation on stepwise regression to obtain the desired number of predictors. For each choice of k, we ran forward stepwise to identify a model with an excess of predictors (including about $1.15k$ predictors), then used backward stepwise to fewer than k predictors (about $0.9k$), and finally ran forward stepwise again to select the final model. This procedure was followed sequentially as k was increased for each trial. This little "oscillation" produced smaller residual sums of squares for large k than were obtained by either forward stepwise or the usual mixed forward/backward algorithm. For example, the lower quartile of the residual sums of squares obtained by the oscillating procedure in the simulation with $k = 40$ and $n = p = 50$ is 60% of the quartile obtained by forward stepwise. Although both are near zero, these differences are quite large on the logarithmic scale of Figure 4.

4 Understanding the Honest Interval

The tiny tail probabilities in these calculations obscure how k, n, and p influence the selection-adjusted endpoint. Indeed, it can be quite hard to compute such an extreme quantile for the χ^2 distribution. The approximation derived in this section for $\chi^2_{d,\alpha/m}$ remedies both problems.

Figure 4: *Simulated residual sums of squares obtained by stepwise regression models, shown with the* $\chi^2_{n-k-1,.05/m}$ *quantile (line). (a) Boxplots summarize 500 independent trials with* $n = 50$ *and* $p = 50$.

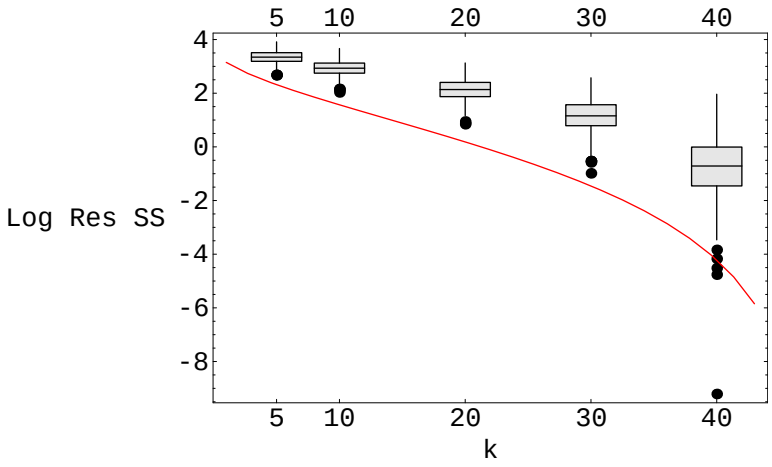
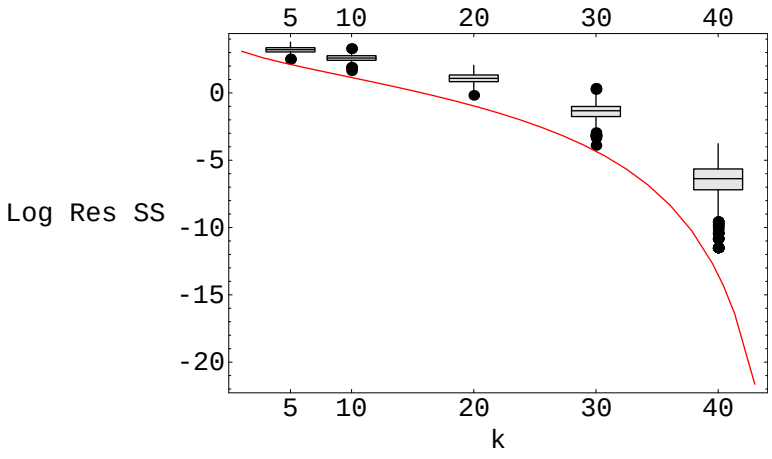


Figure 4: *(b) Summary of 500 trials with* $n = 50$ *and* $p = 200$.



Our approximation arises from a series expansion for the left tail probability of a χ^2 random variable. Assume that the degrees of freedom $d = n - k - 1$ is an even integer and write $h = d/2$. Then integration by parts shows that the critical value $C = \chi^2_{d,\alpha/m}$ satisfies

$$\begin{aligned}
\frac{\alpha}{m} &= \int_0^C \frac{t^{h-1}e^{-t/2}}{\Gamma(h)2^h}dt \\
&= \int_0^{C/2} \frac{t^{h-1}e^{-t}}{\Gamma(h)}dt \\
&= e^{-C/2}\left(\frac{(C/2)^h}{h!} + \frac{(C/2)^{h+1}}{(h+1)!} + \cdots\right) \\
&= e^{-C/2}\frac{(C/2)^h}{h!}\left(1 + \frac{C/2}{h+1} + \frac{(C/2)^2}{(h+1)(h+2)} + \cdots\right) \\
&= e^{-C/2}\frac{(C/2)^{d/2}}{(d/2)!}\left(\frac{1}{1-\frac{C}{d}}\right)^{\theta}, \qquad (8)
\end{aligned}$$

for some $0 < \theta < 1$, assuming the ratio $C/d < 1$ as is the case in the left tail. Now approximate the log of the left hand side of (8) using Stirling's formula $\log n! = (n + \frac{1}{2})\log n - n + (1/2)\log 2\pi + O(1/n)$ to obtain

$$\begin{aligned}
\log \alpha/m &= \log \alpha - \log \binom{p}{k} \\
&= \log \alpha - \frac{1}{2}\log\frac{p}{k(p-k)} - pH(k/p) + O(1)\,, \qquad (9)
\end{aligned}$$

where H denotes the entropy of a Boolean random variable,

$$H(x) = -x\log x - (1-x)\log(1-x), \quad 0 < x < 1\,,$$

and define $H(x) = 0$ for $x = 0, 1$. The entropy is roughly quadratic near its maximum at $x = 1/2$. Equating (9) to the log of the right hand side of (8) gives (after another application of Stirling's formula and simplifying)

$$\log d/C = 1 + \frac{2p}{d}H(k/p) - \frac{1}{d}\left(C + 2\log\alpha + \log\frac{dk(p-k)}{p} + 2\theta\log(1 - C/d) + O(1)\right). \qquad (10)$$

If we drop the parenthesized collection of terms multiplied by $1/d$ (which includes the nominal level α), we have an asymptotic upper bound for the multiplier of s_k^2 in the upper limit of the confidence interval (3),

$$d/C \le \exp\left(1 + \frac{2p}{d}H(k/p)\right). \qquad (11)$$

This bound is asymptotic in the sense that (11) obtains as $d \to \infty$, with k and p fixed or held fixed proportionally to n.

If we ignore the inequality in (11) and solve for C, we obtain the approximate quantile

$$\tilde{C} = \frac{d}{\exp\left(1 + \frac{2p}{d} H(k/p)\right)} \; . \tag{12}$$

This approximation is accurate unless k is small. For small values of k, one obtains a better approximation to the tail quantile by solving

$$\log d/C = 1 + \frac{2p}{d} H(k/p) - \frac{C}{d} \tag{13}$$

for C. Table 2 compares $\tilde{C}$ and this better approximation to the actual χ^2 critical value for various modeling situations. The first part of the table fixes the ratio $k/n = 1/10$, and the second part holds $k/n = 1/2$. In either case, $p = 2n$ with $n = 50, \ldots, 250$. Accuracy of $\tilde{C}$ is adequate, about 10% below the chi-square value, when choosing numerous predictors, but is more than 20% too small when $k = n/10$. We will discuss the orthogonal quantiles in Section 5.

5 Discussion

The procedure used to find a one-sided confidence interval also implies a bias-corrected estimator for σ^2. Since the nominal level α does not appear in the expression (12) for $\tilde{C}$, we are in effect estimating σ^2 by

$$\frac{d\ s_k^2}{\tilde{C}} = s_k^2 e^{1 + \frac{2p}{d} H(k/p)} \; . \tag{14}$$

The exponential term is an adjustment for selection bias. For the stock market example, the corrected estimate of σ^2 is

$$0.0298^2 e^{1 + \frac{2\times 51}{30} H(17/51)} = 0.0298^2 \times 3.164 = .053^2 \; .$$

One might consider using such a bias corrected estimator to pick the appropriate value for k. That is, choose the model which minimizes (14). Some simple calculations show this procedure is related to model selection using a penalized likelihood criterion. Without a balancing measure of the amount of explained variation, however, choosing the model which has the smallest selection-adjusted upper confidence limit leads to

Table 2: *Comparison of the approximate quantile $\tilde{C}$ from (8) and the better approximation from (13) to $\chi^2_{n-k-1,.05/m}$ for models with $n = p$ and $k = n/10$ (top) and $k = n/2$ (bottom). Orthogonal quantiles $O_{.05}$ from a simulation of 2500 samples.*

		Nominal	Selection Adjusted		Approximations		Percentage Error	
k	$n = p$	$\chi^2_{d,.05/m}$	$O_{.05}$	$\chi^2_{d,.05/m}$	"Better"	$\tilde{C}$	"Better"	$\tilde{C}$
5	50	29.79	18.8	10.15	9.62	7.73	-5.2	-24
10	100	68.25	42.4	20.60	19.7	15.8	-4.5	-23
15	150	108.3	67.2	30.89	29.7	23.8	-3.8	-23
20	200	149.1	92.3	41.10	39.8	31.8	-3.2	-23
25	250	190.4	118.4	51.28	49.8	39.9	-2.8	-22
25	50	13.85	1.9	0.563	.502	.492	-11	-13
50	100	33.93	4.6	1.185	1.09	1.06	-8.1	-10
75	150	55.19	7.5	1.794	1.68	1.64	-6.5	-8.6
100	200	77.05	10.5	2.396	2.26	2.21	-5.5	-7.6
125	250	99.28	13.5	2.996	2.85	2.79	-4.8	-6.9

parsimonious models. Continuing with our illustration, the 95% one-sided confidence limit for σ^2 using no predictors (i.e., simply fit a constant) is

$$\frac{(n-1)s_0^2}{\chi^2_{n-1,.05}} = \frac{47 \times .0792^2}{32.27} = 0.096^2 \ .$$

The selection adjusted endpoint for a model with $k = 1$ predictors (the one predictor is the return on the S&P 500) gives the interval

$$\frac{(n-2)s_1^2}{\chi^2_{n-2,.05/51}} = \frac{46 \times .0586^2}{21.89} = 0.085^2 \ .$$

Because the upper endpoint is smaller with this predictor, this approach 'prefers' the model with one predictor over a model with just an intercept. Continuing, the best fitting model with one predictor returns $s_2^2 = 0.0550^2$. The endpoint of the honest interval for this model is

$$\frac{(n-3)s_2^2}{\chi^2_{n-3,2\times.05/(51\times50)}} = \frac{45 \times .055^2}{16.76} = 0.090^2 \ .$$

The endpoints for $k = 3, 4, \ldots$ are larger still in spite of the downward trend seen in Figure 2. Thus, a model selection procedure based on the upper endpoint for σ^2 chooses $k = 1$ and correctly recognizes the importance of the market return as a predictor.

With fast computing widely available, one can simulate more accurate quantiles for a given design matrix rather than rely on the conservative estimates given here. A common situation in which the computing is particularly easy is the special case of $p = n$ orthogonal predictors, as encountered in a wavelet regression. In this setting, one is not choosing from among all $\binom{p}{k}$ subsets, but rather picks the k predictors with the largest t statistics. Rather than use $\tilde{C}$ or $\chi^2_{d,\alpha/m}$, a more accurate upper bound for s_k^2 bound can be found rapidly by simulation. Ignoring the effect of fitting a constant, assume that the n orthogonal predictors are the columns of an $n \times n$ identity matrix. In this canonical form, the minimum residual sum of squares obtained by a model with k predictors is

$$RSS_k = \sum_{j=1}^{n-k} Y^2_{(j)} \ ,$$

where $Y^2_{(1)} < Y^2_{(2)} < \cdots < Y^2_{(n)}$ are the ordered squares of $Y_j \sim N(0,1)$. The sampling distribution of RSS_k in this context is hard to express analytically, but very easy to compute. The column of orthogonal quantiles in Table 2 includes the 5% points

from 2500 samples for each choice of n. These quantiles are several times larger than the conservative $\chi^2_{d,\alpha/m}$ bounds which allow for any type of dependence among the covariates.

We note in closing that our procedure is not needed when $p << n-1$. In this setting, one can and often should estimate σ^2 using the full model and use the resulting estimate to assess the various models, as recommended in [8]. Evidently, though, common statistics packages do not make this choice and instead estimate σ^2 sequentially when computing a stepwise regression. Our adjustment for selection bias is again relevant for such algorithms.

References

1. Berk, K. N. (1978). Comparing subset regression procedures. *Technometrics*, **20**, 1–6.
2. Diehr, G. and D. R. Hoflin (1974). Approximating the distribution of the sample R^2 in best subset regressions. *Technometrics*, **16**, 317–320.
3. Draper, N. R., I. Guttman and H. Kanemasu (1971). The distribution of certain regression statistics. *Biometrika*, **58**, 295–298.
4. Draper, N. and H. Smith (1966). *Applied Regression Analysis*. Wiley, New York.
5. Forsythe, A. B., L. Engelman, R. Jennrich and P. R. A. May (1973). A stopping rule for variable selection in multiple regression. *Journal of the American Statistical Association*, **68**, 75–77.
6. Foster, D. P., R. A. Stine, and R. Waterman (1998). *Business Analysis using Regression*. Springer, New York.
7. Freedman, D. A. (1983). A note on screening regression equations. *American Statistician*, **37**, 152–155.
8. Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.
9. Martens, H. and T. Naes (1989). *Multivariate Calibration*. Wiley, Chichester.
10. Miller, A. J. (1990). *Subset Selection in Regression*. Chapman and Hall, London.

11. Rencher, A. C. and F. C. Pun (1980). Inflation of R^2 in best subset regression. *Technometrics*, **22**, 49–53.