"Cross Validation for Model Selection in Regression"

Glenna M. Gordon
Adivsor: Dr. Mark Greenwood
Department of Mathematical Sciences
Montana State University – Bozeman

March 25, 2008

A writing project submitted in partial fulfillment of the requirements for the degree

Master's of Science in Statistics

1.2 An Example: Prostate Data

Data was collected on 97 men with prostate cancer who were scheduled to receive a radical prostatectomy. The researchers are interested in predicting the volume of the tumor based on various explanatory variables. The questions of interest are which model is closest to the "true" model and which model amongst those considered should be used for prediction and estimation of the parameters. We could devise a candidate set that contains all possible linear models of the form

$$\mathbf{y} = \beta_0 + \mathbf{X}\beta + \epsilon.$$

When the full model is fit, we see that the residuals seem to follow a right skewed distribution and there are violations of the constant variance assumption. In order to keep from violating these assumptions, I will also consider model building on the log-scale,

$$log(\mathbf{y}) = \beta_0 + \mathbf{X}\beta + \epsilon.$$

This suggests an alternative set of nonlinear models to consider are models of the form

$$\mathbf{y} = \beta_0 e^{\mathbf{X}\beta} + \epsilon$$

which have the same mean structure, different variance assumptions and are nonlinear in β. Once the set of candidate models has been determined, we could proceed with model selection methods in order to find the "best" model in our candidate set.

1.3 AIC & Kullback-Leibler Distance

The Kullback-Leibler (K-L) Distance between models f and g is a directed distance between the two models. It can be thought of as the "information lost when g is used to

1 Introduction

1.1 Model Selection

The problem of model selection arises when we have a set of data and we would like to determine the "true model" which generated the data or at least choose the model that is most supported amongst a set of reasonable candidate models. Typically, we have a set of n data points of the form $(\mathbf{x_i}, \mathbf{y_i})$ where $\mathbf{x_i}$ is a vector of k explanatory variables and $\mathbf{y_i}$ is a vector of responses. Often times, the data set is split into two pieces with the first piece to be used for model selection and the second piece to be used to assess the model's ability to predict future observations.

With regards to the model selection process, one should have specified a set of candidate models prior to data collection. It is often assumed that the candidate set contains the"true" model. Theoretically, the model which should be chosen as the "best" model is the one which minimizes a given criterion. Also, the model which minimizes a given criterion should be the "best" model. It is important to note, however, that if the candidate set does not contain "good" models, then even the "best" model in the candidate set may still not be a good model. Another potential problem is that if there are too many models in the candidate set, the chances of finding one distinctly preferred model will decrease and the ability of any criterion to find the "true" model is reduced. Therefore, it is important for the researcher to think about which variables are important in a practical sense and which variables may be important based on prior knowledge of the problem (Burnham & Anderson 1998).

overly complex models, however, it is asymptotically unbiased in large samples. There are several competing criteria, such as AIC_c (Hurvich & Tsai 1989), AIC_u (McQuarrie & Tsai 1998) and BIC (Schwartz 1978), which have different penalties for overfitting and have been shown to be more effective than AIC in small sample model selection.

1.4 MSEP as a target for model selection.

The mean squared error of prediction (MSEP) is an alternate target for model selection as it is a statistic that helps to evaluate the performance of a predictor (Droge 1996). In a regression setting, where we have n observations, $(y_1, y_2, ..., y_n)$, and a true model $\mathbf{y} = \mathbf{X}\beta + \epsilon$ with $\epsilon_i \sim N(0, \sigma^2)$, the true MSEP is defined as the variance of random errors plus the error from estimating the true regression model, or Mean Squared Error (MSE).

$$\begin{aligned} MSEP &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n} E\left(y_i - \hat{y}_i\right)^2 \\ &= \tfrac{1}{n} E||y - \hat{y}||^2 \\ &= \sigma^2 + \tfrac{1}{n}\textstyle\sum_{i=1}^{n} E\left(f(x_i) - \hat{y}_i\right)^2 \\ &= \sigma^2 + MSE. \end{aligned}$$

Small MSEP is possible when the variance of the random errors is small, when the candidate model is close to the true model, or in both situations. In theory, if we can find the model with minimum MSEP it will be closer to the true model than the rest of the candidate models. In practice, however, the true MSEP is unknown and must be estimated.

approximate f" or in a practical interpretation, "the distance from g to f" (Burnham & Anderson 1998). It is defined in the continuous case as $I(f,g) = \int f(x) log\left(\frac{f(x)}{g(x|\theta)}\right) dx$ where f and g represent probability distributions. It is equivalent to say

$$\begin{aligned} I(f,g) &= E_f\left(log\left(\tfrac{f(x)}{g(x|\theta)}\right)\right) \\ &= \int f(x) log(f(x))dx - \int f(x) log(g(x|\theta))dx \\ &= E_f[log(f(x))] - E_f[log(g(x|\theta))]. \end{aligned}$$

In model selection, the goal is to select a model as close to the true model as possible. In other words, we'd like to minimize $I(f,g)$ over all possible functions g. The K-L distance assumes that the true model, f, is fixed and only the approximating models, g_i, are allowed to vary depending on the known values of θ. As a result, $E_f[log(f(x))]$ is a constant that only depends on the unknown truth, so we can write $I(f,g) = C - E_f[log(g(x|\theta))]$.

Suppose we have two candidate models, g_1 and g_2. In this setting, if

$$-E_f[log(g_1(x|\theta))] < -E_f[log(g_2(x|\theta))] \Rightarrow E_f[log(g_1(x|\theta))] > E_f[log(g_2(x|\theta))],$$

then we could say that g_1 is closer to the true model than g_2.

In order to minimize the Kullback-Leibler (K-L) distance, the true model as well as the parameters in the approximating model must all be known. Akaike (1973, 1974) developed an estimator of the expected overall K-L Distance that he called "An Information Criterion" (AIC). It can be shown that $AIC = -2log(g(x|\hat{\theta})) + 2K$ where K is the dimension of $\hat{\theta}$. We can think of $-2log(g(x|\hat{\theta}))$ as a "fit" component with $2K$ serving as a penalizing term when the models are overfit since the dimension of $\hat{\theta}$ is larger when the model is overfit. AIC has a tendency to be biased in small samples with the bias leading to selection of

$$GCV = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 / \left(1 - \overline{h}\right)^2.$$

GCV has been shown to be less variable than CV (Craven and Wahba 1979) and tends to have higher correct selection rates than CV which will be shown in the simulation studies.

2.2 Full Cross Validation Theory

Full Cross Validation (FCV) is designed to help avoid the difficulties that can be encountered with cross validation (Droge 1996) by not requiring the removal of any data point(s). It is defined as $FCV = \frac{1}{n}\sum(y_i - \tilde{y}_i)^2$ where $\tilde{y}_i$ is the least squares prediction of y_i when y_i is substituted by $\hat{y}_i$, instead of leaving it out, in defining the prediction at the i^{th} design point (Droge 1996). FCV can allow for estimation of MSEP in situations where CV fails, however, it has a tendency to select overly complex models (Greenwood 2006). In the simulations, I will show that FCV becomes increasingly negatively biased when overfitting. This leads to selection of overfit models more often since we are minimizing CV and FCV.

In either case, it is important to note that complex models are undesirable. Often times it is difficult or expensive to collect data on many variables especially if it is not absolutely necessary for predictive accuracy. There may be situations when the model with the smallest CV or FCV value is quite complicated, but there is another, simpler, model whose CV or FCV value is close to the smallest CV or FCV value, say $CV < (1+\delta)CV$ for small $\delta > 0$ (Bunke, Droge and Polzehl 1999).

If we were to represent FCV in terms of the residuals from the linear projection model, $\left(y_i - \tilde{y}_{(i)}\right)$ can be shown to equal $(1 + h_{ii})\,(y_i - \hat{y}_i)$, which leads to

$$FCV = \frac{1}{n}\sum_{i=1}^{n}(1 + h_{ii})^2\,(y_i - \hat{y}_i)^2$$

2 Cross Validation vs. Full Cross Validation

2.1 Cross Validation Theory

Cross Validation (CV) gives an estimator of MSEP and is defined with respect to $n-1$ data points instead of all n data points for all n observations. It is defined as $CV = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_{(i)}\right)^2$ where $\hat{y}_{(i)}$ is the prediction of y_i leaving out the i^{th} data point, $(\mathbf{x_i}, y_i)$. In other words, we want to divide the sample "into a construction subsample (of size $n-1$) and a validation subsample (of size 1) in all (n) possible ways" (Stone 1974). It is important to note that CV based esimators have been shown to be asymptotically equivalent to AIC in linear models (Li 1987, Shao 1993). One benefit of CV based estimators is that there is no requirement of estimation of the error variance (Droge 1996) and it has fewer assumptions than AIC. Also, leave one out CV based estimators are fairly easy to calculate. One drawback of CV based methods is that they can fail in some nonlinear regression situations where parameter estimation algorithms may diverge for removal of the i^{th} observation. This occured for the nonlinear model simulations and will be discussed in further detail later in the paper.

CV can be represented in terms of the residuals in any linear projection model. As a result, $\left(y_i - \hat{y}_{(i)}\right)$ becomes $\frac{1}{(1-h_{ii})}\left(y_i - \hat{y}_i\right)$ which leads to

$$CV = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{(1-h_{ii})^2}\left(y_i - \hat{y}_i\right)^2$$

where h_{ii} are the leverages, or diagonal elements of the "hat" matrix, $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Replacing each leverage with its average, $\overline{h} = \sum h_{ii}/n$, leads to Generalized Cross Validation (GCV, Craven and Wahba 1979) where

simulations, with results presented only from two of them since the results were similar for the cases with high leverage observations. In the discussion that follows, a model is considered underfit if the number of covariates in the selected model is fewer than the number of covariates in the "true" model. An overfit model is one in which the number of covariates in the selected model is greater than the number of covariates in the "true" model.

The means, variances, bias and percent correctly selected for each criterion are displayed graphically by model order in Figures 1 and 2. In the plots of the means and of the biases we see that all criteria considered are positively biased when $p < p_0$. However, when $p > p_0$, the cross validation based measures are positively biased and the full cross validation measures become negatively biased. In the plot of variances vs model order we see that the variances for both CV and FCV based criteria decrease as $p \to p_0$. As p becomes larger than p_0, the variance of each measure increases although the full cross validation based measures are less variable than the cross validation based measures.

For the parameters used in these simulations, FCV and GFCV are very unsuccessful model selection criteria, performing even worse than AIC. Hurvich and Tsai (1989) show that the bias in the AIC leads it to have poor ability to discriminate between overfit and correctly specified models in small samples but that the AIC_c corrects this small sample bias and dramatically improves its model selection performance. By examing the model selection results, we can make two important observations. First, FCV-based measures tend to favor complex models. Second, the rate at which FCV selects overly complex models grows with the order of the overfit model that is considered. Although these results

where h_{ii} are the leverages, or diagonal elements of the "hat" matrix, $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Replacing each leverage with its average, $\overline{h} = \sum h_{ii}/n$, leads to Generalized Full Cross Validation (GFCV, Droge 1996) where

$$GFCV = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\left(1+\overline{h}\right)^2.$$

Similar to GCV, GFCV has been shown to be less variable than FCV (Droge 1996) and tends to have higher correct selection rates than FCV which will also be shown in the simulation studies.

2.3 Linear model simulations

A simulation study is used in order to evaluate the performance and behavior of CV and FCV based estimators compared with AIC. Independent Uniform(0,10) explanatory variables were generated for the results provided from the simulation. The response variables are built as functions of a subset of the generated, random explanatory variables plus independent, normal random errors with standard deviation 2.

Two different combinations of sample size and number of explanatory variables are considered; case 1 has five potential covariates, a true model that contains four, equally important covariates ($y_i = 1 + x_{1i} + x_{2i} + x_{3i} + x_{4i} + \epsilon_i$), and a sample size of $n = 15$. Case 2 contains ten potential covariates, a true model that contains four, equally important covariates, and a sample size of $n = 25$. Additionally, I explore the role of leverage on different estimators by simulating the explanatory variables from either Uniform(0,10) or Exponential($\lambda = 0.10$) distributions. Using an exponential distribution provides some simulated models that contain high leverage points. This leads to four different sets of

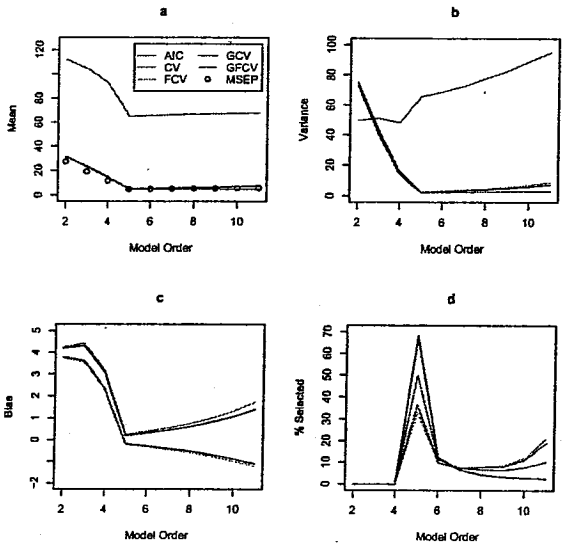

Figure 2: Results for Case 2, Nested Models (p^*=5) from 15,000 simulation runs. FCV and GFCV are the lowest dashed lines in all panels except in (d) where they increase dramatically for higher order models.

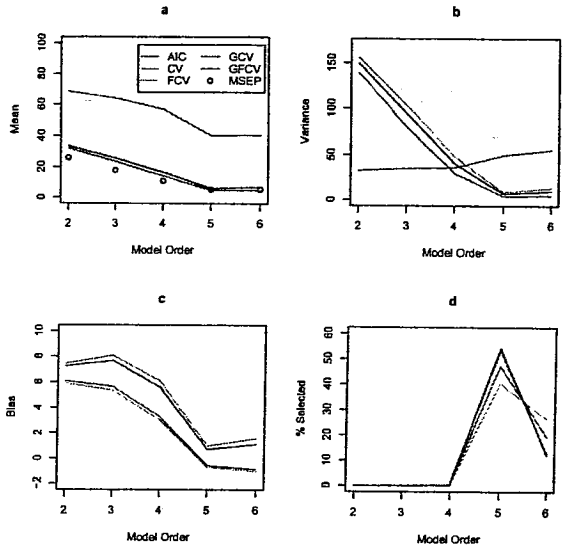

Figure 1: Results for Case 1, Nested Models (p^*=5) from 15,000 simulation runs. FCV and GFCV are the lowest dashed lines in all panels.

are specific to this set of linear regression models, similar selection of overly complicated models can be expected when using FCV in other situations.

In addition to the two cases discussed previously, I also considered a third case; all possible linear combinations of 7 covariates with a true model of 3 covariates and a sample size of $n = 97$. In this case, AIC_c had the highest correct selection rates with the FCV-based measures having the lowest correct selection rates. When considering how many times each model was selected as one of the top 6 preferred models, Table 6 shows that the FCV-based measures select the most complex model almost 3 times as often as AIC_c. Also, it is interesting to note that the true model was nested within all of the models that were selected by each of the criteria examined. This suggests that it is acceptable to examine only nested models when considering the performance of the different criteria.

Order	*CV*	*FCV*	*GCV*	*GFCV*	*AIC*
2	0.03	0.01	0.01	0.01	0.01
3	0.11	0.01	0.05	0.01	0.03
4	0.61	0.11	0.31	0.11	0.15
5	53.06	40.31	54.3	42.33	47.27
6	12.86	26.23	11.99	24.22	19.22

Table 3: *Case 1 where the true model contains 4 covariates (p^*=5 including the intercept term) and n=15. 15,000 iterations through model selection process, percentage of each order model selected.*

Order	*CV*	*FCV*	*GCV*	*GFCV*	*AIC*
2	0	0	0	0	0
3	0	0	0	0	0
4	0	0	0	0	0
5	67.23	33.27	68.08	36.66	50.09
6	12.6	10.11	12.18	9.95	11.35
7	6.73	7.41	6.61	7.35	7.52
8	4.51	7.93	4.39	7.85	6.7
9	3.49	8.43	3.29	8.13	6.53
10	2.93	12.04	2.89	11.12	7.65
11	2.51	20.81	2.56	18.93	10.17

Table 4: *Case 2 where the true model contains 4 covariates (p^*=5 including the intercept term) and n=25. 15,000 iterations through model selection process, percentage of each order model selected.*

$$\begin{aligned} \mathbf{y} &= e^{\beta_0+\mathbf{X}\beta+\epsilon} \\ &= \beta_0^* e^{\mathbf{X}\beta+\epsilon} \\ &= \beta_0^* e^{\mathbf{X}\beta} e^{\epsilon} \\ &= \beta_0^* e^{\mathbf{X}\beta} \epsilon^* \end{aligned}$$

where ϵ^* follows a Log Normal distribution. The model that may actually be desired by the researcher is $\mathbf{y} = \beta_0^* \mathbf{e}^{\mathbf{X}\beta} + \epsilon$ where Normal errors are considered, but the mean structure is based on the log-transformation. The implication that is often overlooked when fitting the $log(y)$ version of the model is that the error structure is being changed from an additive Normal error structure to a multiplicative Log-Normal error structure.

Order	$\overline{MSEP}$	$\overline{CV}$	$\overline{FCV}$	$\overline{GCV}$	$\overline{GFCV}$
2	26.15	33.6	32.03	33.4	32.22
3	18.12	26.18	23.47	25.79	23.77
4	11.17	17.24	14.19	16.76	14.46
5	5.33	6.31	4.65	6.04	4.78
6	5.6	7.17	4.58	6.72	4.74

Table 1: *Case 1 where the true model contains 4 covariates (p^*=5 including the intercept term) and n=15. 15,000 iterations through model selection process, mean of the criteria.*

Order	$\overline{MSEP}$	$\overline{CV}$	$\overline{FCV}$	$\overline{GCV}$	$\overline{GFCV}$
2	27.38	31.63	31.12	31.57	31.17
3	19.14	23.55	22.7	23.45	22.78
4	11.64	14.78	13.86	14.69	13.94
5	4.8	5.06	4.58	5.01	4.62
6	4.96	5.35	4.63	5.27	4.68
7	5.12	5.67	4.67	5.57	4.73
8	5.28	6.04	4.67	5.89	4.75
9	5.44	6.45	4.66	6.26	4.75
10	5.6	6.93	4.61	6.68	4.71
11	5.76	7.5	4.54	7.16	4.65

Table 2: *Case 2 where the true model contains 4 covariates (p^*=5 including the intercept term) and n=25. 15,000 iterations through model selection process, mean of the criteria.*

2.4 CV & FCV in Nonlinear models

In the linear model simulation studies, we examined the linear model $\mathbf{y} = \beta_0 + \mathbf{X}\beta + \epsilon$ where $\epsilon \sim N(0, \sigma^2)$. In many cases, the linear model may be inadequate and a nonlinear model may need to be considered. One common approach is to model the log of the response instead of the response itself. The model then becomes $log(\mathbf{y}) = \beta_0 + \mathbf{X}\beta + \epsilon$ also with $\epsilon \sim N(0, \sigma^2)$, which is still a linear model for $log(y)$. Exponentiating both sides, we obtain the following:

Model $f(\mathbf{X}) = \mathbf{X}\beta + \epsilon$	AIC	AIC_c	CV	FCV	GCV	GFCV
$f(x_2, x_3, x_4)$	83.3	88.5	85.4	80.5	85.2	81.1
$f(x_2, x_3, x_4, x_5)$	67.3	72.1	68.2	65.2	69.5	65.9
$f(x_2, x_3, x_4, x_6)$	66.3	72	69.6	64.4	68.7	65
$f(x_2, x_3, x_4, x_7)$	67.4	74.2	69.8	65.5	70.5	65.2
$f(x_2, x_3, x_4, x_8)$	70	75.2	71.1	66.2	71.5	68
$f(x_2, x_3, x_4, x_5, x_6)$	32.6	30.6	32	33.3	31.7	32.9
$f(x_2, x_3, x_4, x_5, x_7)$	29.6	27.6	29.3	30.4	29.2	30.2
$f(x_2, x_3, x_4, x_6, x_7)$	30.4	29.2	28.5	30.6	30.1	30.4
$f(x_2, x_3, x_4, x_5, x_8)$	33.6	31	33.2	34.2	32.3	34.4
$f(x_2, x_3, x_4, x_6, x_8)$	32.7	31.3	32.3	33.9	32	33.4
$f(x_2, x_3, x_4, x_7, x_8)$	31.1	29.9	31.2	31.3	30.7	31.1
$f(x_2, x_3, x_4, x_5, x_6, x_7)$	12	8.7	11.7	13.9	11.1	13.2
$f(x_2, x_3, x_4, x_5, x_6, x_8)$	13.2	9.5	11.2	13.9	11.5	14.3
$f(x_2, x_3, x_4, x_5, x_7, x_8)$	13.5	9.2	11.9	15.7	11.4	15.4
$f(x_2, x_3, x_4, x_6, x_7, x_8)$	12.4	9.2	11	15.1	11.2	13.9
$f(x_2, x_3, x_4, x_5, x_6, x_7, x_8)$	4.6	1.8	3.6	5.9	3.4	5.6

Table 6: *Case 3 where the true model contains 3 covariates (p^*=4 including the intercept term) and n=97. 1,000 iterations through model selection process, percentage each model selected as one of top 6.*

2.5 Nonlinear model simulations using $\mathbf{y} = \beta_0 e^{\mathbf{X}\beta} + \epsilon$

For the nonlinear simulations, I examined all possible combinations of seven covariates where the model was $\mathbf{y} = \beta_0 e^{\mathbf{X}\beta} + \epsilon$ and the true model had 3 covariates. The potential problems discussed in the previous section arose during the nonlinear simulation study. As a result, only 500 simulations were run. Similar to the linear model simulations, the explanatory variables were independent, Uniform(-1,1) random variables with $\beta_1 = (1, 1, 1)$ and a variance of 4 was used for the error terms. This small variance was used primarily to provide numerical stability for the simulation runs. Additionally, I used a sample size of $n = 97$ to mimic the setting of the prostate data set as closely as possible. From Table 7, we see that in the linear simulation study, FCV based measures performed the worst, selecting

Model $f(\mathbf{X}) = \mathbf{X}\beta + \epsilon$	AIC	AIC_c	CV	FCV	GCV	GFCV
$f(x_2, x_3, x_4)$	46.9	53.7	49.8	45.4	50.2	45.1
$f(x_2, x_3, x_4, x_5)$	10.2	9.3	9.5	9.8	9.6	10.4
$f(x_2, x_3, x_4, x_6)$	9.2	9.1	9.2	9.5	9.1	9.4
$f(x_2, x_3, x_4, x_7)$	7.9	7.6	7.3	7.1	7.7	7.3
$f(x_2, x_3, x_4, x_8)$	10.6	9.8	10.7	11.2	10.6	11.1
$f(x_2, x_3, x_4, x_5, x_6)$	2.6	1.8	2.2	2.6	2.1	2.7
$f(x_2, x_3, x_4, x_5, x_7)$	1.6	1.1	1.6	1.8	1.5	1.8
$f(x_2, x_3, x_4, x_6, x_7)$	1.6	1.3	1.5	1.8	1.5	1.6
$f(x_2, x_3, x_4, x_5, x_8)$	2.2	1.6	1.8	2.7	1.8	2.3
$f(x_2, x_3, x_4, x_6, x_8)$	1.9	1.3	1.8	1.9	1.7	2
$f(x_2, x_3, x_4, x_7, x_8)$	3.2	2.4	2.8	3.8	2.8	3.9
$f(x_2, x_3, x_4, x_5, x_6, x_7)$	0.6	0.2	0.5	0.8	0.3	0.7
$f(x_2, x_3, x_4, x_5, x_6, x_8)$	0.3	0.3	0.3	0.4	0.2	0.3
$f(x_2, x_3, x_4, x_5, x_7, x_8)$	0.5	0.2	0.4	0.5	0.4	0.6
$f(x_2, x_3, x_4, x_6, x_7, x_8)$	0.4	0.2	0.4	0.4	0.3	0.4
$f(x_2, x_3, x_4, x_5, x_6, x_7, x_8)$	0.3	0.1	0.2	0.3	0.2	0.4

Table 5: *Case 3 where the true model contains 3 covariates (p^*=4 including the intercept term) and n=97. 1,000 iterations through model selection process, percentage of each order model selected.*

When fitting the model, $y = f(\mathbf{x})$, we are assuming that $f(\mathbf{x})$ is a nonlinear function of $\mathbf{x}$. However, when $f(\mathbf{x})$ is a linear function of $\mathbf{x}$, we can consider it as a special case of the nonlinear model and the results of this section will still apply.

One problem that may arise in the nonlinear setting is that the leverages have the potential to be greater than or equal to 1. Such leverages are called superleverages and occur because the hat matrix used to calculate the leverages is only an approximate projection operator (St. Laurent & Cook 1992). Superleverages pose a problem for calculating CV using the leverages and typical residuals. One implication is that if the leverage is equal to 1, then $\frac{1}{1-h_{ii}}$ is undefined and if $h_{ii} > 1$, then $(y_i - \hat{y}_i)/(1 - h_{ii})$ may not be approximately equal to $y_i - \hat{y}_{-i}$. So in order to compute CV, all n models need to be fit. This can become computationally intensive and may cause problems, especially during simulation studies.

vored a model with two covariates, patient age and capsular penetration. AIC, GCV, and GFCV all favored the same model with three covariates; patient age, seminal vesicle invasion and capsular penetration whereas CV and FCV favored the same model with five covariates; prostate weight, seminal vesicle invasion, capsular penetration, gleason score, and percentage of gleason scores that were a 4 or 5.

AIC	AIC_c	CV	FCV	GCV	GFCV
2 4 5	2 5	1 4 5 6 7	1 4 5 6 7	2 4 5	2 4 5
2 5	2 4 5	1 4 5 6	1 4 5 6	2 5	2 5
1 4 5 6	1 5	4 5 6 7	2 4 5 6 7	1 4 5 6	1 4 5 6
1 2 4 5	1 4 5 6	4 5 6	1 2 4 5 6 7	1 4 5	1 2 4 5
1 4 5	1 4 5	2 4 5	4 5 6 7	1 5 6	1 4 5 6 7
1 5 6	1 5 6	2 4 5 7	2 4 5	1 2 5	1 4 5
1 2 5	1 2 5	4 5	4 5 6	1 2 4 5	1 5 6
1 5	1 2 4 5	1 4 5	1 4 5	1 5	1 2 5
1 4 5 6 7	4 5	1 5 6	1 5 6	2 4 5 6	2 4 5 6 7
2 4 5 6	2 4 5 6	2 4 5 6	2 4 5 6	1 4 5 6 7	1 5

Table 8: *Top 10 linear models selected by each criterion for the prostate data set. Each number responds to a particular covariate.*

AIC	AIC_c	CV	FCV	GCV	GFCV
1 2 3 4 5	1 2 4 5	3 4 5	3 4 5	1 2 4 5	1 2 3 4
1 2 4 5	1 2 3 4 5	4 5 7	4 5 7	1 2 3 4 5	1 2 4 5
2 4 5	2 4 5	4 5	1 4 5 7	2 4 5	2 3 4 5
2 3 4 5	2 3 4 5	3 5	3 4 5 6	2 3 4 5	2 4 5
1 2 3 5	1 2 3 5	1 4 5 7	3 5	1 2 3 5	1 2 3 5
1 2 3 4 5 6	1 2 4 5 7	3 4 5 6	3 4 5 7	1 2 3 4 5 6	1 2 3 4 5 6
1 2 3 4 5 7	1 2 4 5 6	4 5 6	4 5	1 2 4 5 7	1 2 3 4 5 7
1 2 4 5 7	1 2 3 4 5 6	3 5 6	4 5 6 7	1 2 4 5 6	1 2 4 5 7
1 2 4 5 6	1 2 3 4 5 7	4 5 6 7	4 5 6	1 2 3 4 5 7	1 2 4 5 6
2 4 5 7	1 2 5	5 7	3 5 6	1 2 5	2 4 5 7

Table 9: *Top 10 nonlinear models selected by each criterion for the prostate data set.*

In running the prostate data set through the nonlinear model selection process, CV and FCV both chose a model with three covariates, while the rest of the criterion selected

the most complex model about 3 times as often as AIC_c. For the nonlinear simulations, FCV based measures performed even worse than expected, selecting the most complex model more than 6 times as often as AIC_c.

Model $f(\mathbf{X}) = \beta_0 e^{\mathbf{X}\beta+\epsilon}$	AIC	AIC_c	CV	FCV	GCV	GFCV
$f(x_2, x_3, x_4)$	79.4	85.8	80.4	63.4	82	77.2
$f(x_2, x_3, x_4, x_5)$	69.6	74.4	64	55.6	70.8	67.8
$f(x_2, x_3, x_4, x_6)$	64.8	67.8	59.8	48.2	65.6	63.2
$f(x_2, x_3, x_4, x_7)$	66	68.6	61.2	52.6	67.2	64.8
$f(x_2, x_3, x_4, x_8)$	64.2	69	60	51	65.8	62.6
$f(x_2, x_3, x_4, x_5, x_6)$	30	29.4	31	34.2	29.8	29.8
$f(x_2, x_3, x_4, x_5, x_7)$	36.6	35.8	39.8	42.8	36.6	37.4
$f(x_2, x_3, x_4, x_6, x_7)$	31.6	30.8	30.4	34	31.4	32.2
$f(x_2, x_3, x_4, x_5, x_8)$	33	31.2	38	37.8	32.6	32.6
$f(x_2, x_3, x_4, x_6, x_8)$	31	30.6	32.8	32.2	30.6	31
$f(x_2, x_3, x_4, x_7, x_8)$	31.4	31.6	33.8	35.2	31.8	31.4
$f(x_2, x_3, x_4, x_5, x_6, x_7)$	14.8	11.8	17.8	25	14	16.6
$f(x_2, x_3, x_4, x_5, x_6, x_8)$	15.8	9.6	14.6	24.4	13	17.2
$f(x_2, x_3, x_4, x_5, x_7, x_8)$	14	10.8	13.6	23.6	12.6	15.4
$f(x_2, x_3, x_4, x_6, x_7, x_8)$	13	10	13.8	21.4	11.8	13.8
$f(x_2, x_3, x_4, x_5, x_6, x_7, x_8)$	4.8	2.8	9	18.6	4.4	7

Table 7: *Case 4 where the true model contains 3 covariates (p^*=4 including the intercept term) and n=97. 500 iterations through model selection process, percentage each model was selected as one of top 6 nonlinear models.*

2.6 Prostate Data Results

The explanatory variables in the prostate data set are prostate weight, age, benign hyperplasia amount, seminal vesicle invasion, capsular penetration, Gleason score, percentage of Gleason scores that are 4 or 5, and prostate specific antigen. For the model selection process, these are numbered 1, 2,...,8 respectively. The response variable is the volume of the cancerous tumor.

When I ran the prostate data set through the linear model selection process, AIC_c fa-

for generalized linear models.

In both the linear and nonlinear simulation studies, CV and FCV performed worse than their respective generalized versions. Also, for both the linear and nonlinear simulations, GCV and GFCV were better than AIC, but not AIC_c. Based on the performance of the CV and FCV based criteria in the simulation studies, I do not trust them completely for the prostate data results. However, in examining the preferred models across all criteria, there seems to be several important variables; prostate weight, patient's age, seminal vesicle invasion, and capsular penetration. In spite of the fact that interactions were not considered in either the simulation studies or the analysis of the prostate data set, the poor performance of CV and FCV based criteria in these simple settings should be a warning against using them in more complex situations, including those which account for interactions.

models with four or more covariates. For the prostate data set, AIC tended to favor more xomplex models than . Also, based on the correct selection rates of CV and FCV based methods in both the linear and nonlinear model simulations, it is likely that overfit models were selected by both criteria. One rule of thumb that is often used is to see which models have AIC values within 2 units of the model with the minimum AIC value. Those models are then considered equivalent models. For the prostate data set, over 40 models were within 2 units of the minimum AIC value. For simplicity, I chose to examine only the top 10 linear and the top 10 nonlinear models. Upon closer examination of the top 10 linear (shown in Table 8) and the top 10 nonlinear (shown in Table 9) models selected, it became apparent that the important variables are prostate weight, patient's age, seminal vesicle invasion, and capsular penetration.

3 Conclusions & Future Research

One problem that was seen with the FCV based criteria is that they have a tendency to have large negative bias. One possible direction for future research would be to explore this bias and try to determine if there is a correction factor that may decrease the bias in FCV-based criteria. Another problem that is seen is the potential for superleverages. When a situation occurs where superleverages exist, the use of leverages in computing CV and FCV based estimates of MSEP is impossible or very inaccurate. It would be interesting to explore simulations which force superleverages to exist. Finally, exploring the impact of different types of error structures, such as Poisson or Binomial, on model selection would be a natural next step in my research, potentially considering CV and FCV related measures

References

[1] Akaike, H. (1973), "Maximum likelihood identification of Gaussian autoregressive moving average models," *Biometrika*, **60**, 255-265.

[2] Akaike, H. (1974), "A new look at statistical model identification," *IEEE transactions on Automatic Control*, **19**, 716-723.

[3] Bunke, O., Droge, B. and Polzehl, J. (1998), "Splus tools for model selection in non-linear regression," *Computational Statistics*, **13**, 257-281.

[4] Bunke, O., Droge, B. and Polzehl, J. (1999), "Model selection, transformations and variance estimation in nonlinear regression," *Statistics*, **33**, 197-240.

[5] Burnham, K. and Anderson, D. (1998), *Model Selection and Multimodel Inference*. New York, Springer.

[6] Craven, P. and Wahba, G. (1979), "Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation," *Numerical Mathematics*, **31**, 377-403.

[7] Droge, B. (1996), "Some comments on cross-validation," in W. Hardle and M. Schimek (Eds.), *Statistical Theory and Computational Aspects of Smoothing*. Physica, Heidelberg, 178-199.

[8] Hurvich, C. and Tsai, C. (1989), "Regression and time series model selection criteria," *Biometrika*, **76**, 297-307.

[9] Kullback, S. (1968), *Information theory and statistics*. New York, Dover.

[10] McQuarrie, A. and Tsai, C. (1998), *Regression and Time Series Model Selection*. Singapore, World Scientific Publishing Co.

[11] Schwarz, G. (1978), "Estimating the dimension of a model," *The Annals of Statistics*, **6**, 461-464.

[12] Stone, M. (1974), "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society-B*, **36**, 111-147.