K-Fold Cross-Validation is Superior to Split Sample Validation for Risk Adjustment Models

Randall P. Ellis[1] and Pooja G. Mookim[2]

June 7, 2013

Abstract: This paper examines cross-validation techniques, with a particular focus on assessing the predictive validity of risk adjustment models as commonly estimated. We validate that K-Fold cross-validation is more efficient than a 50-50 split sample and illustrate that overfitting with rich risk adjustment models remains meaningful even in samples of a million observations. A new estimation algorithm is described that efficiently calculates K-Fold cross-validated R-squared and other measures of goodness of fit using only three (XXX verify) passes through the data, and hence can be applied easily on sample sizes in the millions without sorting or relying on repeated split-sample techniques. Analysis of K-fold cross-validation results using a large claims dataset is used to calculate the standard deviation and bias of fitted R-squares for different models and sample sizes, which have a larger bias in moderately large sample sizes than most researchers would realize. Programs that implement the algorithm in SAS and STATA are presented that can be easily used on any sample.

Acknowledgements: The authors thank Verisk Health for providing funding and data support and Kevin Lang and Timothy Layton for the helpful comments and suggestions.

[1] Department of Economics, Boston University, 270 Bay State Road; Boston MA 02215; ellisrp@bu.edu

[2] XXX

1 Introduction

Risk adjustment models are now widely used for health plan payment, patient management, severity adjustment, and performance measurement and it has become the norm to estimate these models on extremely large samples (N > 5 million) that enable large numbers of explanatory variables to be included (Ellis and Layton, 2013). The rich explanatory models that are feasible with large samples are not always appropriate on smaller samples, or for different dependent variables where overfitting can be a serious concern. Assessing overfitting is particularly important when alternative specifications are compared that differ not only in the predictors used, but also in the in the demographics, year, country, or choice of the dependent variable to be predicted (e.g., Winkelman and Mehmud, 2007). Split sample validation is commonly used, in which a fraction of the data (usually 50 percent) is used for estimation, and the remainder is used for validation. We make two contributions in this paper. The first is to show that overfitting is still a meaningful concern even with up to a million observations with today's sophisticated risk adjustment models. Second, we show that while split sample validation is useful, it is an inefficient approach when the goal is to simply validate existing risk adjustment models and structures and assess overfitting. Instead we demonstrate that K-fold cross validation, which sequentially uses all of the data for both estimation and validation, is meaningfully more efficient, computationally feasible (particularly for linear models), and easy to explain. Programs written for SAS and STATA are linked and provided in an appendix that make it easy to implement the algorithms on other data. (XXX put in link to web page here.)

A classic example of split sample validation is the 2007 Society of Actuaries report (Winkelman and Mehmud, 2007) which evaluated 12 distinct claims-based risk assessment models and several dozen alternative specifications using a single 50-50 split sample validation on a standardized sample of 617,683 individuals. Many other important methodological papers also evaluate alternative models of annual health spending using split sample methods (Mullahy, 1998; Manning and Mullahy, 2001; Pope et al, 2004; Basu et al (2004), Manning et al. (2005); and Fishman et al., 2006, Dixon et al XXXX). Most of these papers do not explore the fact that by using only a single split of their full sample, their results are sensitive to the particular split sample created. An important exception is Buntin and Zaslavsky (2004) who evaluate eight linear and nonlinear models using the Medicare Current Beneficiary Survey data on 10,134 individuals. Their validation uses 100 different replications of 50/50 splits of their sample for validation.

It is well known that estimating models of health care costs is problematic due to the heavily right-skewed nature of the distribution of non-zero annual costs (Jones, 2010). Less commonly emphasized is that the explanatory variables are also often highly skewed. Estimation of predictive models using OLS produces biased measures of statistical significance and can lead to 'overfitting'. While many early studies advocated strongly for nonlinear models to reduce the overfitting problem (Duan et al 1983) this preference was driven heavily by the limited sample sizes used for estimation. Several more recent studies (Fishman et al, 2006; Deb and Burgess, 2007; Ellis and McGuire, 2007; Ellis et al, 2013) suggest that the overfitting problems of OLS models largely disappear when very large samples sizes (over a million individuals) are used, an issue we also revisit here. Because of their ability to accommodate enormous numbers of covariates (in the hundreds), their computational speed for estimation, simplicity to use for prediction and ease of explanation, linear predictive models have reemerged as the preferred specification over nonlinear techniques for risk adjustment purposes. Although nonlinear models remain popular among academic researchers, and may be essential for hypothesis testing in small to moderate size samples, none of the commercially available risk adjustment models evaluated by the Society of Actuaries (Winkelman and Mehmud, 2007) use nonlinear models for prediction.

In this paper we focus on the validation of an existing model, not validation done in the process of choosing regressors to use in new models. If the researcher is using data to define explanatory variables, choose exclusions and interactions, or evaluate diverse nonlinear structures, then the K-fold cross validation techniques described here can be misleading. K-fold cross validation can help identify overfitting that results from estimation, but it cannot easily be used to understand overfitting due to model design and selection. For model development, relying on validation in new samples is the preferred method.

This paper makes two contributions. First, we show that the overfitting problem can be substantial even with sample sizes as large as one million, but that overfitting largely disappears in samples in the millions. The magnitude of the overfitting problem even in samples over 100,000 has perhaps been underappreciated in studies using such "small to moderate" size samples, and such small samples cannot themselves be relied upon to validate the extent of the overfitting problem. Second, we describe an efficient algorithm for implementing K-fold cross validation in linear models. This efficient algorithm is applied to large empirical samples of several million records, taking approximately three to five times the clock time of running a single OLS regression model. The algorithm uses the K-Fold cross validation technique developed in the statistics literature. Although we develop the algorithm using health expenditure data predicted using a linear risk adjustment framework, the method is general and could be applied to any data.

In this paper we present results solely for the R-squared measure of predictive ability and Copas tests of overfitting. The R-squared is attractive as a unit-free measure easily interpreted across models and samples, and the Copas test is a widely used measure of overfitting in linear models. It is straightforward to calculate alternative measures such as the root mean square error, mean absolute deviation, and predictive ratios by selected percentiles using the results from K-fold validation, and these measures are generated by the programs attached to this paper with further passes through the data. (XXXneed to add to the programs).

Risk adjustment can be used for many purposes, including health plan capitation payment (Ash et al, 1989; van de Ven and Ellis, 2000, Dixon et al XXXX), provider profiling (Thomas, et al, 2004a, 2004b), case management, plan rating and underwriting (Cumming and Cameron, 2002), and quality assessment and improvement (Iezzoni, 2013). For this paper, we examine three risk adjustment models predicting total spending: a simplified model that uses only 18 age and gender categories, a prospective model that use diagnostic information to predict subsequent year health spending, and a concurrent model that uses the same diagnostic information to predict health spending in the same year. For the diagnostic models, the diagnosis based explanatory variables are the 182 explanatory variables generated by the Hierarchical Condition Category (HCC) model (Pope et al, 2000) that underlies the classification system used for the Medicare Advantage, Medicare Part D and HHS Health Insurance Exchange risk adjustment models.[3]

2 Model Prediction and Cross-Validation

A common approach for choosing among competing alternative model specifications is based on their *validated* rather than within sample or *fitted* predictive power. One well-known early discussions cross-validation is Stone (1974, 1977), who developed the idea of splitting a dataset and then estimating a model on one part of the data and evaluating it using the other part. The two most common approaches are data splitting and K-Fold cross-validation.

[3] Each of these implemented risk adjustment models use 60-80 HCCs rather than the full set, while in this paper we use all 182.

2.1 Data Splitting or Split Sample Technique

The most common form of cross-validation is data-splitting or split sample validation. In this approach, the researcher selects (usually randomly) from the total set of observations available a "training" or "estimation" sample to estimate the model and subsequently uses the model to predict the dependent variable for the remaining holdout or “validation” sample. Predictive validity is assessed by using some measure of correlation between the values from the holdout sample and the values predicted by the model. Alternatively, dependent variable in the validation sample can be regressed on the predicted value using coefficients from the estimation sample and a COPAS test (XXX need cite) of whether the coefficient in that regression differs from one is a test of whether there is evidence of overfitting.

Traditionally, data splitting is done only once rather than several times. This makes the results dependent on which data points end up in the training set and which end up in the test set. Sometimes it can lead to unexpected results, such as when the validated R-square is larger than the estimation sample R-square. Following Buntin and Zaslavsky (2004), the standard practice today is to perform this exercise repeatedly and then take the mean of the estimates. This “repeated split sampling” is used in our analysis below.

With a large number of draws the mean R-square in the training sample will be above the mean in the validation sample, but since split samples only use part of the data for calibration, estimates are never as efficient as when the entire sample is used. If overfitting is an issue, then using only a subsample to measure overfitting reduces statistical precision in both the training *and* the validation samples, and potentially increases the divergence between R-squares from the two samples.

2.2 K-Fold Cross Validation

An alternative approach called “K-fold” cross-validation makes more efficient use of the available information. The algorithm for this approach is as follows:

1. Randomly split the sample into K equal parts
2. For the kth part, fit the model to the other K-1 parts of the data, and use this model to calculate the prediction for each observation in the kth part.
3. Repeat the above step for k=1, 2….K and combine the K sets of prediction to create a full sample of actual and predicted values
4. Use the actual and predicted values to generate any measures of goodness of fit that are desired.

If K equals the sample size (N), this is called N-fold or "leave-one-out" cross-validation. Even if K is two, this method differs from the "split sample" method where only a single subset (the validation set) is used to estimate the prediction error instead of k=2 different subsets.[4]

One difficulty with K-fold cross-validation is that it can be computationally slow with nonlinear models (including L' regression, tree structured methods for classification and nonlinear regression). Even for OLS K-fold cross validation can be slow when millions of observations and hundreds of explanatory variables are used. We describe a computationally efficient algorithm for conducting the K-fold cross validation below, and link computer code that implements the algorithm efficiently on the web.

2.3 COPAS Test

The COPAS test is a formal test of overfitting using split sample or K-fold cross validation. The following algorithm is used to perform this test using split samples.

1. Randomly split sample into two groups. The selection of groups can be 50-50 or 70-30 or (K-1, k). Call the first group A or the training sample and the other group B, or the validation sample
2. Estimate model on sample A and retain its coefficients $\hat{\beta}_A$
3. Forecast to sample B

$\hat{Y}_B = \hat{\beta}_A X_B$

4. Now regress the dependent variable from the validation sample i.e., Y_B on the predicted $\hat{Y}_B$ and test whether the slope is one. Hence estimate

$$Y_B = \delta_0 + \delta_1 \hat{Y}_B + \varepsilon \quad \text{and test } \delta_1 = 1$$

5. If reject the null hypothesis, then overfitting may be a problem. Simplifying the model such as by omitting variables or constraining them (“pruning”) is the normal prescription.

With split-sample validation, the convention is to repeatedly (100 or 1000 times) use different splits of the sample, conducting a COPAS test for each split, and then report the percentage of times the null hypothesis was rejected. For K-fold cross validation, the natural extension is to calculate the COPAS test

[4] Leave-one-out cross-validation is also easily confused with jackknifing. Both involve omitting each training case in turn and retraining the network on the remaining subset. But cross-validation is used to estimate generalization error, while the jackknife is used to estimate the within-sample bias of a statistic. In the jackknife, you compute some statistic of interest in each subset of the data. The average of these subset statistics is compared with the corresponding statistic computed from the entire sample in order to estimate the bias of the latter. You can also get a jackknife estimate of the standard error of a statistic. Jackknifing can be used to estimate the bias of the training error and hence to estimate the generalization error (Efron, 1982).

statistics once using the observed and out of sample predicted values for Y for the full sample. In what follows we calculate COPAS test statistics using both repeated split sample and K-fold validation methods.

3 A computationally efficient method for K-fold validation

Conducting multiple split samples and the straightforward application of K-fold cross validation generally require multiple passes through the dataset, which can be computationally time consuming when very large sample sizes and very large numbers of explanatory variables are involved. Part of the contribution of this paper is in verifying the usefulness of a computationally fast algorithm for conducting k-fold cross validation.

Our approach is easily explained using matrix notation. Let A_{-k} denote a matrix A generated while excluding the proper subset A_k of observations in set k. If Y is the Nx1 array of the dependent variable and X is the NxM matrix of explanatory variables, let Z = {X Y}. It is well known that the cross product matrix $Z^T Z$ contains all of the information needed to generate all conventional regression statistics, including betas, RSE and R-square. The algorithm we implement for a sample size of N is as follows.

1. Randomly sort the observations so that there is no significance to their order.
2. Estimate the OLS model using the full data set Z, retaining $Q = Z^T Z$.
3. For each of k subsamples of size N/K, created without replacement, generate $Q_k = Z_k{}^T Z_k$ and take matrix differences to generate $Q\text{-}_k = Q – Q_k = Z^T Z - Z_k{}^T Z_k$
4. Use Q_{-k} to calculate the array of OLS regression coefficients $\beta_{-k}(Q_{-k})$, and then generate predicted values $\hat{Y}_k$, which were not used in $\beta_{-k}(Q_{-k})$. Save these fitted values of $\hat{Y}_k$ in an $\{\hat{Y}_{K-Fold}\}$
5. After repeating steps 3 and 4 for all of the k samples, generate validated R-square, RSE, MAE and and COPAS test statistics measures for the full sample of size N using the original Y and $\{\hat{Y}_{K-Fold}\}$.

Reflecting the increased precision from larger samples, we repeated steps 1 through 5 for 1000 replications for small sample sizes of 1000, 2000, and 5000 observations; 100 replications for sample size of 10,000, 20,000, 50,000, 100,000, 200,000 and 500,000; 50 replications for the sample sizes of 1,000,000; and once for the entire sample (N = 4,688,092). We also explored the sensitivity of our results to various values of K=10, 100, and 1000.

4 Data

Data for this study came from the 2003 and 2004 MEDSTAT (now Truven Analytics) MarketScan Commercial Claims and Encounter Databases. These databases contain pooled and de-identified health information on to help its clients manage the cost and quality of healthcare they purchase on behalf of their employees. MarketScan is the data from these client databases. In this study we use the Commercial Claims and Encounters (CC&E) Database for 2003 and 2004, which contains the healthcare experience of approximately 10 million employees and their dependents in 2003 and 2004.[5] After excluding people who were not continuously eligible for coverage for all of 2003 and 2004, everyone Medicare eligible at any time, one person with implausibly high health spending in 2003, and people not in traditional indemnity, a preferred provider organization, a point of service plan, or a health maintenance organization, this left 4,688,092 individuals in our full sample.

Using this data, we evaluate three different model specifications.

- Age and sex model with 18 independent variables, (age/gender dummies)
- Prospective model with 18 independent age gender dummies and 182 hierarchical condition categories[6]
- Concurrent model with 18 independent age gender dummies and 182 hierarchical condition categories

5 Results

5.1 Descriptive statistics

Summary statistics are presented in Table 1. We see that both health spending in 2003 and health spending in 2004 have coefficients of variation (standard deviation divided by the sample mean) over 300, large skewness measures (in the 30's) and enormous kurtosis (over 2000). Note that all of these measures (the CV, skewness and kurtosis) are invariant to rescaling or normalization of the variable of interest. Also relevant are the moments of some of the explanatory variables. Age, gender, and dummy

[5] . This data was not used to calibrate or revise the DCG/HCC risk adjustment classification system used for validation, and hence this is an appropriate sample for model validation.

[6] The DxCG HCC classification system contains 184 HCCs, however two of them never occurred in our data and hence are omitted. These two were HCC 129 End stage Renal Disease (Medicare program participant), and HCC 173 Major Organ Transplant Status (e.g., heart, lung, etc.) which in the first case is impossible in our data by construction, and in the second case is sufficiently rare among non-Medicare eligibles to have not occurred in our sample.

variables reflecting their interaction all have relatively low CV (less than 500), and have low skewness and kurtosis. In contrast, a relatively rare HCC 1 such as HIV/AIDS, with a prevalence rate of .00075 has a CV that exceeds that of annual spending (CV = 3651), a skewness of 36, and a kurtosis of 1329. Hence, despite having an acceptable sample size of over 3400 cases with HIV/AIDS in the full sample, this variable will be subject to overfitting in modest size samples. Congestive heart failure, a binary variable with a nearly tenfold higher prevalence (mean = .00654) still has meaningful skewness and kurtosis.

5.2 Full sample results

Table 2 presents the results of estimating our three risk adjustment models using the full sample size of N = 4,688,092. Our base model is a prospective model, predicting 2004 total health care spending at the individual level using 18 age, gender, and diagnostic information from 2003. This base "prospective model" has 200 parameters: a constant term plus 182 hierarchical condition categories and 17 of 18 mutually exclusive age-gender categories. All of these explanatory variables are binary variables. We also estimate results for an "Age-sex model" using only the age-gender dummies and a "concurrent model", predicting 2003 spending in the same year as the diagnostic information (2003).

Table 2 reveals that the fitted and validated R-square measures for the prospective split sample model differ by only .005. The COPAS test on overfitting has a t ratio of -15.024 indicating that with 2.2 million records there is still some evidence of overfitting. In contrast the K-fold validation results differ at most by .001 when the full sample is used, and, hence, results are not overstated by overfitting. The age-sex model, with only 18 parameters, does not explain much of the variation in spending, but also shows no evidence of overfitting. The COPAS test statistic on the slope for the prospective HCC model is of borderline significance with a t ratio of 1.807 (p = .07 on two-tail test).

5.3 K-Fold versus split sample results

We next present results from K-Fold cross validation and compare them with split sample technique results. Table 3 presents these results for the prospective model. First consider the split sample results. For samples under 10,000, the R-square in the fitted models is grossly overstated, with highly negative validated R2. For a more respectable sample size of 10,000 the fitted R-square has a mean of .359, while the validated R2 mean remains negative. We see that the fitted and validated R-squares diverge markedly for smaller samples—validating the well-known results that significant overfitting remains a concern even with sample sizes of 100,000 in richly predictive models using highly skewed explanatory variables. As the sample size increases, this divergence gradually disappears, and the overfitting problem seems to be mitigated for large samples over 500,000 observations.

The superiority of the K-fold cross validation over split sample validation is revealed in Figure 1, which highlights that the K-fold R-square means are significantly closer to the true values than the split sample methods. These also reveal that the simple average of the fitted and validated R-square is a better estimate of the asymptotic predictive power of the model then just the validated R-square. Another feature to note in this figure is that for K=100, the fitted mean R-square from split sample techniques using a sample of N observations is statistically indistinguishable from the fitted mean R-square from K-Fold cross validation technique using a sample of N/2 observations. Hence the split sample validation results on 20,000 records gives nearly identical results to the K-fold validation results on 10,000 observations. This makes sense since splitting a 20,000 observation dataset into two parts and estimating the model using one part is almost identical to taking a 10,000 observation dataset and using 99% of it to estimate the model.

Figure 2 plots not only the means but also the 90 percent confidence intervals for the fitted and validated R-squares using K-fold validation on the prospective HCC model. The 90 percent confidence intervals for the split sample model are even wider. This figure reveals that the 90 percent confidence intervals for the validated and fitted values overlap considerably, so it is not unusual for the validated and fitted R-square values to be reversed for the split sample techniques, simply due to chance. In contrast, with K-fold validation the validated measure is guaranteed to be strictly less than the fitted value, since one can never do better using an out of sample model than using within sample methods. (See Efron and Tibshirani, 1998 for a demonstration.)

In Table 4 we repeat this exercise using only 18 age/sex dummies as our right hand side variables. The first four columns of table 4 show that the fitted and validated R-square estimates are very close to each other for the age-sex model even with as few as 10,000 observations. However, the 90% confidence intervals shown in figure 3 reveal that there is still a meaningful amount of variation in estimates of the R-square even with this simply parameterized model.

Figure 4 and the second set of columns in Table 4 present the concurrent model to show that K-Fold cross validation is useful for concurrent models as well. Overfitting is more significant in concurrent models with the mean R-squares differing by .005 even with a million individuals.

Table 5 evaluates the impact of different choices of how many folds should be used in the K-Fold cross validation exercise. Each cell was generated by taking the mean and standard deviation of the validated R2 from 100 replications for the given sample size for K = 10, 100 and 1000. Comparing across rows, we see that estimates of the validated R-square using K-fold validation is relatively stable across values of K with a slight improvement for going from K = 10 to K = 100, but no apparent improvement going from 100 to 1000. In part because of the computational savings we rely on K=100 for the rest of our results.

In Table 6 we show the average time it took for us to validate our model using split sample and K-fold techniques. It matters critically in this analysis whether the time taken generating the sample splits themselves are included in the estimates. Because split sample validation only estimates the model on half as much data, and forecasting the remaining half is very fast, split sample validation when efficiently programmed can take even less time than OLS. However, it is more interesting to note that the K-fold Cross validation only took at most 5 times longer than the OLS, despite running 100 regression models. Even for the full sample of 4.7 million records, K-fold validation took only 4:02 versus :43 for OLS. Straightforward bootstrap techniques with 100 repetitions will have taken on the order of 100 times as much time as OLS to generate similar results. All of the times shown here were generated on a basic Dell Pentium IV desktop that had only 2.8 GHZ of processor speed and 2.0 GB of RAM. Many research settings would typically have access to much faster machines. As a comparison, researchers at DxCG Inc. estimated and validated a concurrent regression model using this K-Fold algorithm with 13.65 million records and 835 explanatory variables. Our algorithm took only 3.3 times as much clock time (115 minutes) as doing OLS (35 minutes), despite doing 100 regressions with 835 betas, each on over 13 million records. The overfitting problem was also trivial, with only a .002 overstatement in the R-square. Bootstrap methods on this large sample could have taken multiple days to generate comparable statistics.

6 Conclusions

In this paper we have illustrated the value of using K-fold cross validation instead of split sample validation for validating linear models on large datasets. This technique is relevant in settings where the researcher is interested in comparing alternative sets of explanatory variables in a risk adjustment setting without exploring model specification, as in Winkelman and Mehmud (2007) and Manning et al (2008). If model selection tasks such as identifying which variables to include, searching among highly nonlinear models, or evaluating interactions and constraints are being considered, then split sample techniques will isolate the validation sample from contamination in ways that K-fold validation cannot.

This paper documents the magnitude of overstatement of the R-square using three specifications of common risk adjustment – age-gender, prospective and concurrent models. We have used the DCG risk adjustment framework for all of our estimates, but the techniques should be relevant for any setting in which overfitting is of concern. K-Fold cross validation is superior to split sample techniques, even when multiple splits are considered, since it achieves the same level of precision with half of the data. We have used the K-fold validation to calculate only one measure of goodness of fit – the R-square, but the individual level out-of-sample predictions can be used for any number of other measures – including mean absolute deviations, predictive ratios and grouped R-squares – with a simple modification.

Our demonstration that K-fold validation is relatively robust to relatively small values of K, such as ten, suggests that for nonlinear models, where computation time is a critical issue, K-fold cross validation using only K=10 may be attractive as an alternative to split sample techniques. Often the very large sample sizes available to researchers imply that the relevant choice of models is between using all of the data for a simple linear model versus a fraction of it for nonlinear estimation. We have not validated the relative attractiveness of these two competing approaches to model estimation.

Part of the contribution of this paper is that we develop a computationally efficient method of calculating K-fold validation which requires only on the order of five times as long as the amount of time running a simple OLS regression in large samples. Given that bootstrap techniques require a much larger multiple of time to generate comparable measures of out of sample predictive power, our hope is that this technique, long known in the statistical literature, will see increased use in empirical studies of large datasets.

Table 1 Summary Statistics

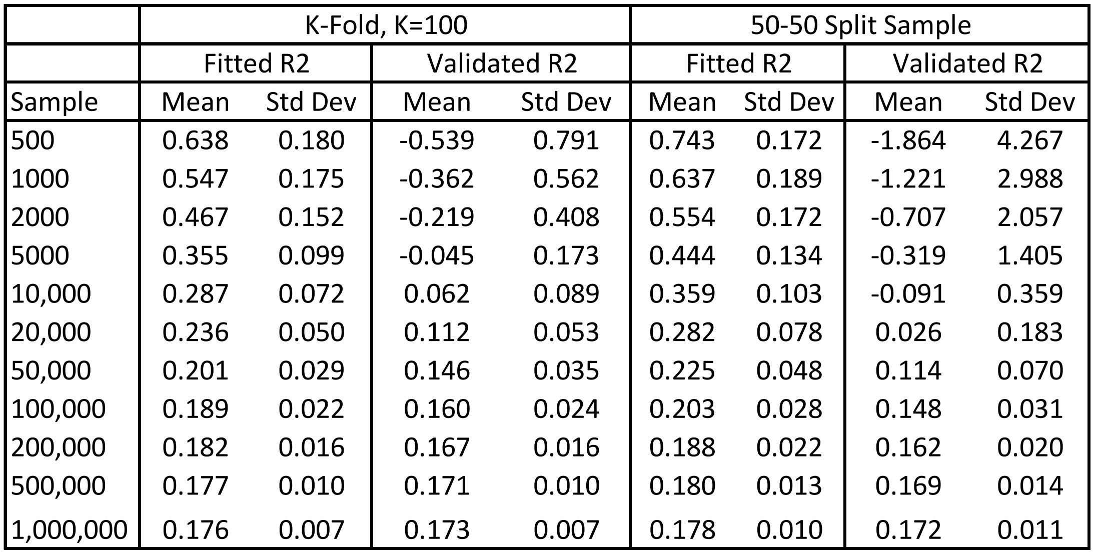
Marketscan data, 2003-2004, N = 4,688,092

	Mean	Std. Dev	CV*100	Skewness	Kurtosis	Maximum
Covered total charges, 2003	2905	10859	374	32	2,405	1,909,854
Covered total charges, 2004	3463	11675	337	27	2,174	2,222,606
Age (years)	34.95	18.41	53	-0.32	-1.21	63
Dummy if male, age = 0-5	0.048	0.21	447	4.25	16.03	1
HCC001 HIV/AIDS	0.00075	0.027	3651	36	1,329	1
HCC080 Congestive Heart Failure	0.00654	0.081	1233	12	148	1

Table 2 Full Sample Results

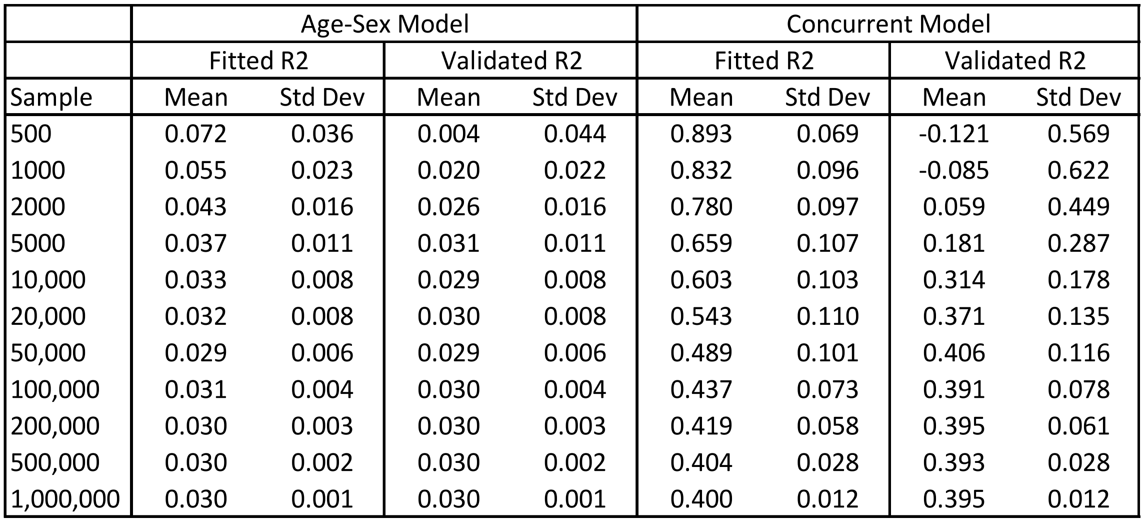
MEDSTAT Marketscan Data, 2003-2004, N = 4,688,092				
	Parameters	Fitted R-Square	Validated R-Square	COPAS T-Ratio
Prospective HCC	200	0.175	0.174	1.807
Prospective AgeSex	18	0.030	0.030	0.045
Concurrent HCC	200	0.398	0.397	2.028
Prospective HCC Split Sample Technique	200	0.177	0.172	-15.024

Table 3: R Squares generated for Prospective Model with 218 parameters, HCC+ AGE + SEX, 100-Fold, Cross Validation Vs. 50-50 Split



	K-Fold, K=100				50-50 Split Sample			
	Fitted R2		Validated R2		Fitted R2		Validated R2	
Sample	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
500	0.638	0.180	-0.539	0.791	0.743	0.172	-1.864	4.267
1000	0.547	0.175	-0.362	0.562	0.637	0.189	-1.221	2.988
2000	0.467	0.152	-0.219	0.408	0.554	0.172	-0.707	2.057
5000	0.355	0.099	-0.045	0.173	0.444	0.134	-0.319	1.405
10,000	0.287	0.072	0.062	0.089	0.359	0.103	-0.091	0.359
20,000	0.236	0.050	0.112	0.053	0.282	0.078	0.026	0.183
50,000	0.201	0.029	0.146	0.035	0.225	0.048	0.114	0.070
100,000	0.189	0.022	0.160	0.024	0.203	0.028	0.148	0.031
200,000	0.182	0.016	0.167	0.016	0.188	0.022	0.162	0.020
500,000	0.177	0.010	0.171	0.010	0.180	0.013	0.169	0.014
1,000,000	0.176	0.007	0.173	0.007	0.178	0.010	0.172	0.011

Table 4: R Squares generated for Age-Sex model and Concurrent Model using K-Fold Cross Validation



	Age-Sex Model				Concurrent Model			
	Fitted R2		Validated R2		Fitted R2		Validated R2	
Sample	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev	Mean	Std Dev
500	0.072	0.036	0.004	0.044	0.893	0.069	-0.121	0.569
1000	0.055	0.023	0.020	0.022	0.832	0.096	-0.085	0.622
2000	0.043	0.016	0.026	0.016	0.780	0.097	0.059	0.449
5000	0.037	0.011	0.031	0.011	0.659	0.107	0.181	0.287
10,000	0.033	0.008	0.029	0.008	0.603	0.103	0.314	0.178
20,000	0.032	0.008	0.030	0.008	0.543	0.110	0.371	0.135
50,000	0.029	0.006	0.029	0.006	0.489	0.101	0.406	0.116
100,000	0.031	0.004	0.030	0.004	0.437	0.073	0.391	0.078
200,000	0.030	0.003	0.030	0.003	0.419	0.058	0.395	0.061
500,000	0.030	0.002	0.030	0.002	0.404	0.028	0.393	0.028
1,000,000	0.030	0.001	0.030	0.001	0.400	0.012	0.395	0.012

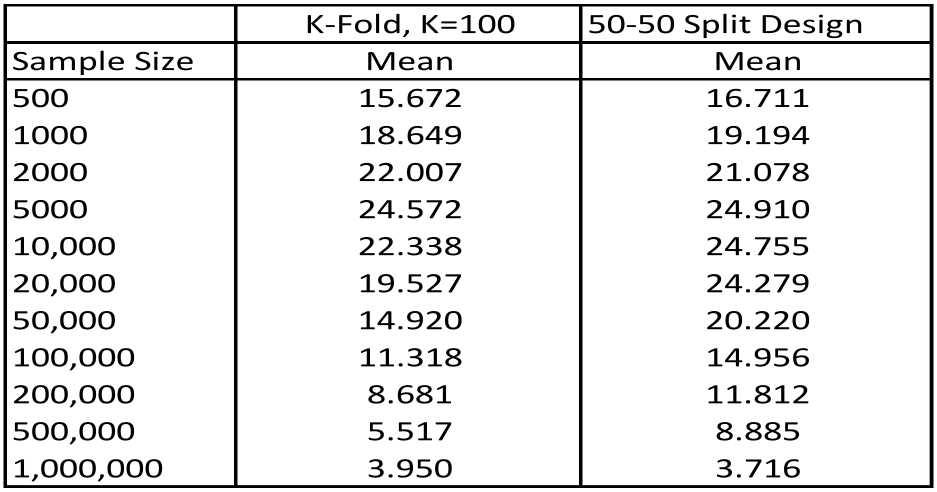
Table 5: Comparison of Validated R-Square Mean and Standard Deviation for various Choices of K and sample sizes, K-Fold Cross Validation on Prospective HCC + Age + Sex Model, and 218 parameters

	Validated R2 Mean			Validated R2 Std Dev		
Sample Size	K=10	K=100	K=1000	K=10	K=100	K=1000
2000	-0.154	-0.219	-0.127	0.220	0.408	0.203
5000	-0.047	-0.045	-0.032	0.139	0.173	0.134
10,000	0.050	0.062	0.067	0.095	0.089	0.084
20,000	0.099	0.112	0.108	0.057	0.053	0.056
50,000	0.142	0.146	0.146	0.035	0.035	0.034
100,000	0.157	0.160	0.158	0.023	0.024	0.023
200,000	0.165	0.167	0.165	0.015	0.016	0.015
500,000	0.170	0.171	0.171	0.010	0.010	0.010

Table 6: Comparison of Average Computer Time Utilized in validating in 100 samples of different sizes, Prospective Model, 218 parameters HCC + AGE +SEX

	OLS	50-50 Split Design	K-Fold, K=100
Sample Size	Time in Seconds	Time in Seconds	Time in Seconds
1000	0.246	0.047	20.640
2000	0.341	0.046	21.219
5000	0.276	0.078	21.984
10,000	0.288	0.109	22.796
20,000	0.411	0.187	23.063
50,000	0.737	0.391	24.640
100,000	1.588	0.750	27.266
200,000	2.463	1.609	34.407
500,000	7.382	3.375	45.297
1,000,000	17.593	8.547	69.234

Table 7: Average T-Ratio on Copas Test, Various Sample Sizes, 50-50 Split vs. K-Fold Cross Validation



	K-Fold, K=100	50-50 Split Design
Sample Size	Mean	Mean
500	15.672	16.711
1000	18.649	19.194
2000	22.007	21.078
5000	24.572	24.910
10,000	22.338	24.755
20,000	19.527	24.279
50,000	14.920	20.220
100,000	11.318	14.956
200,000	8.681	11.812
500,000	5.517	8.885
1,000,000	3.950	3.716

Figure 1: Fitted and Validated R^2 means by sample size, 200 parameters HCC + Age + Sex model, 50-50 Split Technique Vs. K-Fold Cross Validation

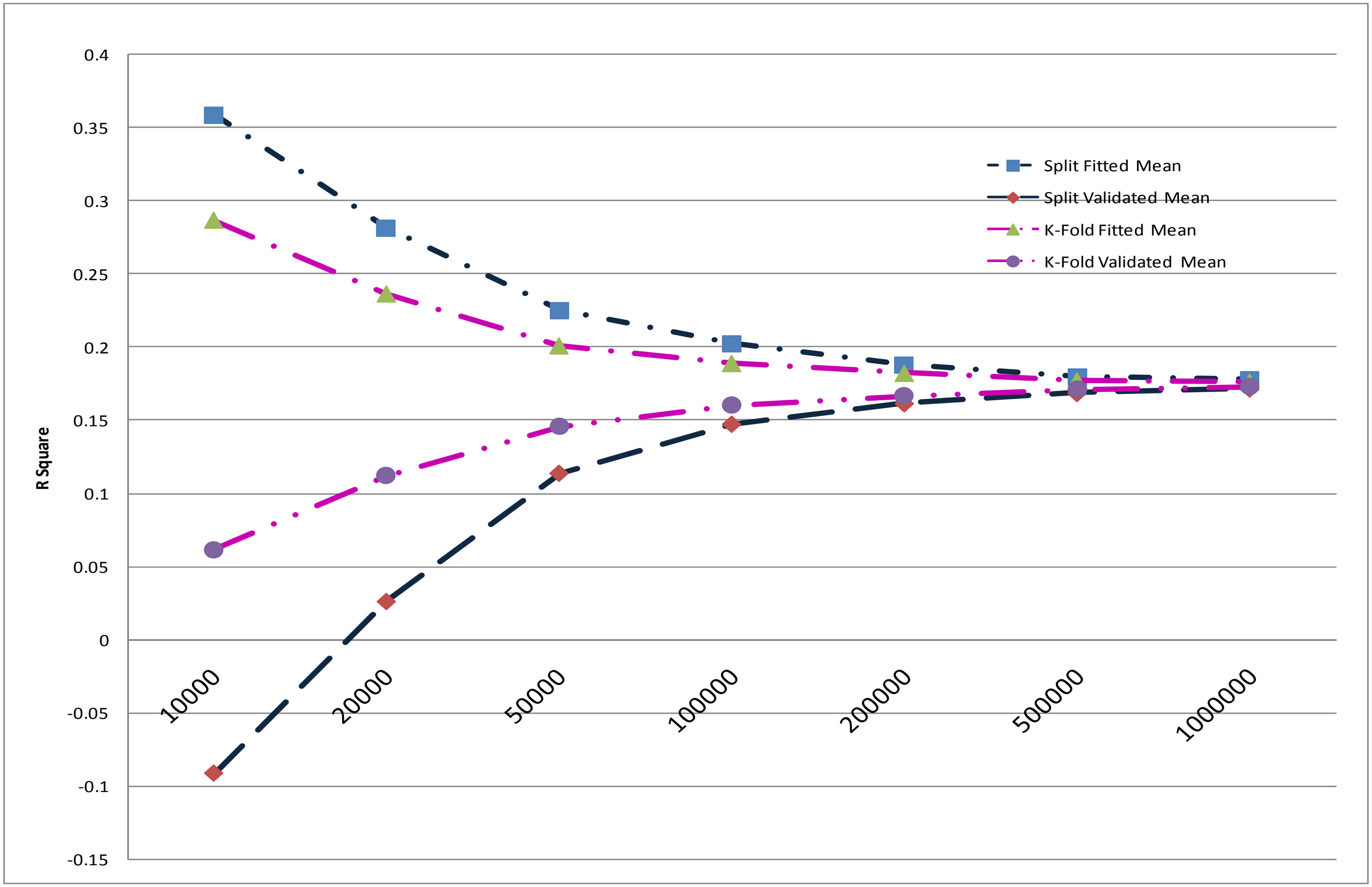


Figure 2: Fitted and Validated R2 by sample size, 200 parameter Prospective HCC + Age + Sex model, means and 90% confidence intervals, K-Fold Cross Validation Method

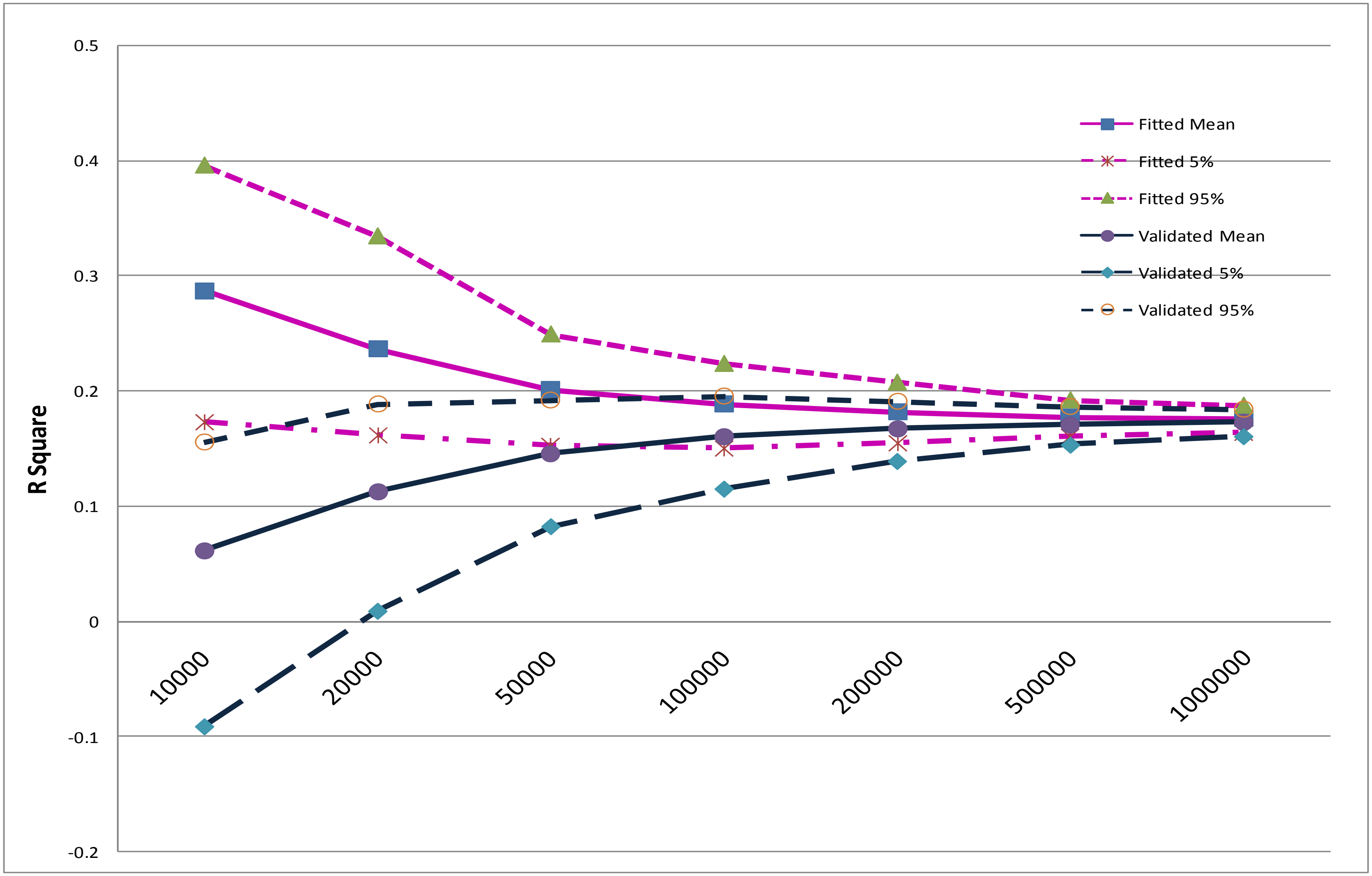


Figure 3: Fitted and Validated R^2 by sample size, 18 parameter Age + Sex model, means and 90% confidence intervals, K-Fold Cross Validation

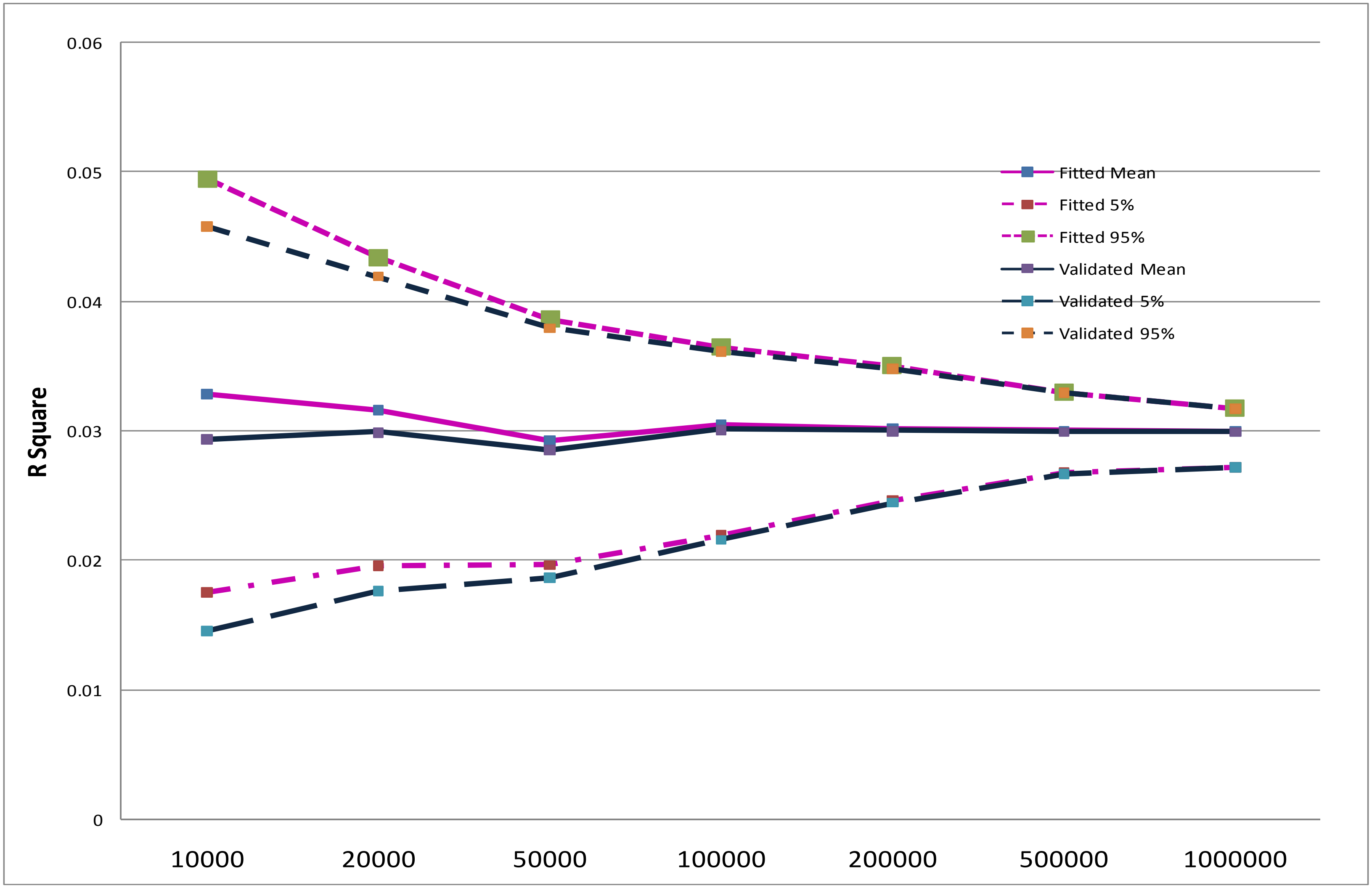
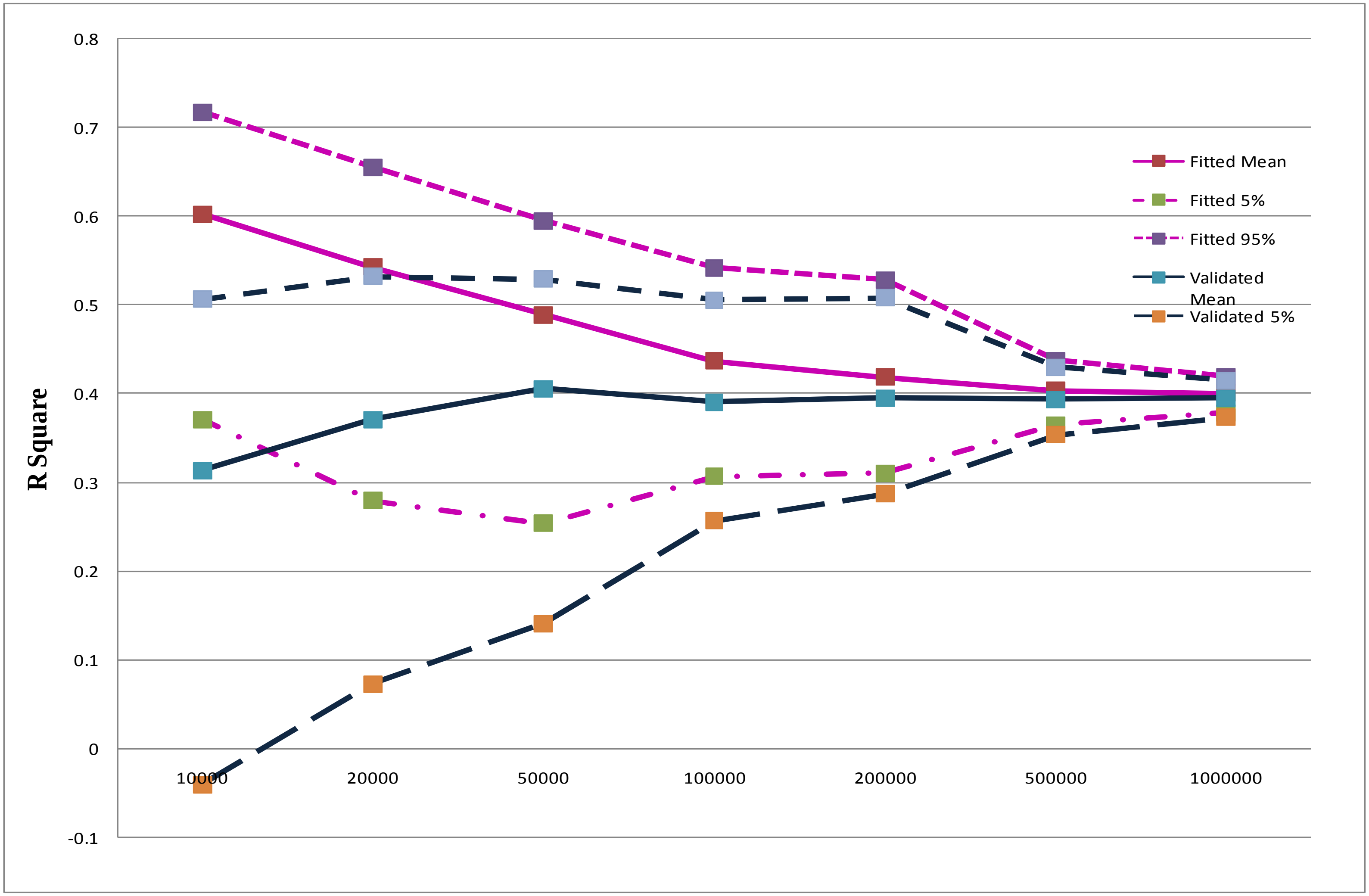


Figure 4: Fitted and Validated R^2 by sample size, 200 parameter Concurrent HCC + Age + Sex model, means and 90% confidence intervals, k-Fold Cross Validation



REFERENCES

1. Adamson DM, Chang S, Hansen LG “Health Research Data for the Real World: The Marketscan Databases” Thomson Medstat White Paper, January 2006.

2. Ash, A.S., Byrne-Logan, S. (1998), "How well do models work? Predicting health care costs", Proceedings of the Section on Statistics in Epidemiology of the American Statistical Association, Dallas, TX.

3. Ash, A.S., Porell, F., Gruenberg, L., et al. (1989) Adjusting Medicare capitation payments using prior hospitalization data, Health Care Fin Rev, 10(4):17-29.

4. Buntin, M.B., and Zaslavsky, A.M. (2004) Too much ado about two-part models and transformation? Comparing methods of modeling Medicare expenditures J Health Econ 23, 525-542.

5. Cooil, B. K., D. L. Rados and R. L. Winer (1987), "Cross-Validation for Prediction," *Journal of Marketing Research,* 24 (August), 27 1-279.

6. Cumming, R.B., Knutson, D., Cameron, B.A., Derrick, B. (2002) A Comparative Analysis of Claims-based Methods of Health Risk Assessment for Commercial Populations. Research study sponsored by the US Society of Actuaries.

7. Deb, P, and Burgess J.F. Jr. (2007). 'A quasi-experimental comparison of statistical models for health care expenditures.' Mimeo.

8. Duan, N., et al, 1983. A Comparison of Alternative Models for the Demand for Medical Care, J Bus & Econ Stat 1, pages 115-26.

9. Efron, B. (1982) The Jackknife, the Bootstrap and Other Resampling Plans, Philadelphia: SIAM.

10. Efron, B. and Tibshirani, R.J. (1993), An Introduction to the Bootstrap,London: Chapman & Hall.

11. Ellis, R.P. (2008) "Risk adjustment in health care markets: concepts and applications" in Lu, Mingshan, and Jonnson, Egon, Paying for Health Care: New Ideas for a Changing Society. Wiley-VCH publishers Weinheim, Germany.
12. Ellis, R.P., Fiebig, D.G., Johar, M, Jones, G., Savage, E. (2013) “Explaining Health Care Expenditure Variation: Large Sample Evidence Using Linked Survey and Health Administrative Data” Health Economics.
13. Ellis, R.P. Jiang S, and Kuo T.C. (2013) "Does service-level spending show evidence of selection across health plan types?" *Applied Economics*.45(13):1701-1712.

14. Ellis, R. P. and McGuire, T.G. (2007) "Predictability and predictiveness in health care spending" Journal of Health Economics. 26:25–48.

15. Ellis, R.P., Pope, G.C., Iezzoni, L.I., et al.: Diagnosis-Based Risk Adjustment for Medicare Capitation Payments. H*ealth Care Financing Review* 12(3):101-128. 1996.

16. Fishman, P. Sloan, K. Burgess J., Jr, Zhou C. , Wang, L. (2006) Evaluating Alternative Risk Assessment Models: Evidence from the US Veteran Population, Group Health Center for Health Studies working paper, May, 2006, available at: http://www.centerforhealthstudies.org/ctrstaff/fishman.html.

17. Iezzoni, L.I. (2003) Risk adjustment for measuring health outcomes. Fifth edition. Ann Arbor, MI: AcademyHealth HAP.

18. Manning, W.G. et al. (1987) Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment, Am Econ Rev 77, 251-77.

19. Manning, W.G., A. Basu, and J. Mullahy, (2005) Generalized Modeling Approaches to Risk Adjustment of Skewed Outcomes Data, J Health Econ 24, 465-488.

20. Manning, WG and Mullahy J. (2001) Estimating log models: To transform or not to transform? J Health Econ 20, 461-494.

21. Mullahy, “Much Ado about Two: Reconsidering Retransformation and the Two-Part Model in Health Econometrics”, Journal of Health Economics, 1998, v17 (3,Jun), 247-281.

22. Pope, G.C., R.P. Ellis, A.S. Ash, C.F. Liu, J.Z. Ayanian, D.W. Bates, H. Burstin, L.I.Iezzoni, M.J. Ingber (2000) Principal Inpatient Diagnostic Cost Group Models for Medicare Risk Adjustment Health Care Fin Rev 21, 93-118.

23. Pope, G.C., R.P. Ellis, A.S. Ash, J.Z. Ayanian, D.W. Bates, H. Burstin, L.I. Iezzoni, E.Marcantonio, B. Wu, (2000) Diagnostic cost group hierarchical condition category models for Medicare risk adjustment. Final Report to Health Care Financing Administration, December.

24. Rice, N. and Smith, P. (2001), Capitation and risk adjustment in health care financing: an international progress report, Milbank Q 79, 81-113.

25. Stone M. “Cross-Validatory Choice and Assessment of Statistical Predictions” Journal of Royal Statistical *Society. Series B (Methodological)*, Vol. 36, No. 2. (1974), pp. 111-147.

26. Thomas, J. W., Grazier, K.L. and Ward K. (2004a) Comparing Accuracy of Risk-Adjustment Methodologies used in Economic Profiling of Physicians. Inquiry 41, 218-231.

27. Thomas, J. W., Grazier, K.L. and Ward K. (2004b) Economic Profiling of Primary Care Physicians: Consistency among Risk-Adjusted Measures. Health Serv Res 39, 985-1003.

28. Van de Ven, W., Ellis, R. P., (2000) Risk adjustment in competitive health plan markets. In: Culyer, A.J., Newhouse, J.P.(Eds.), Handbook of Health Economics. North-Holland.

29. Winkelman R. and Mehmud S. (2007) “A Comparative Analysis of Claims Based Tools for Health Risk Assessment, Society of Actuaries, http://www.soa.org/files/pdf/risk-assessmentc.pdf.