Paper 106-2009

The Problem of Regression Assumptions and the Use of Predictive Modeling

Patricia B. Cerrito, University of Louisville, Louisville, KY

ABSTRACT

Predictive modeling includes regression, both logistic and linear, depending upon the type of outcome variable. However, as the datasets are generally too large for a p-value to have meaning, predictive modeling uses other measures of model fit. Generally, too, there are enough observations so that the data can be partitioned into two or more datasets. The first subset is used to define (or train) the model. The second subset can be used in an iterative process to improve the model. The third subset is used to test the model for accuracy.

The definition of "best" model needs to be considered as well. In a regression model, the "best" model is one that satisfies the criteria of uniform minimum variance unbiased estimator. In other words, it is only "best" in the class of unbiased estimators. As soon as the class of estimators is expanded, "best" no longer exists, and we must define the criteria that we will use to determine a "best" fit. There are several criteria to consider. For a binary outcome variable, we can use the misclassification rate. However, especially in medicine, misclassification can have different costs. A false positive error is not as costly as a false negative error if the outcome involves the diagnosis of a terminal disease. We will discuss the similarities and differences between the types of modeling.

Another consideration is the assumptions required for regression; predictive modeling is more nonparametric in nature. We will examine the assumption of normality and the use of the Central Limit Theorem.

INTRODUCTION

Regression has been the standard approach to modeling the relationship between one outcome variable and several input variables. Generally, the p-value is used as a measure of the adequacy of the model. There are other statistics, such as the r^2 and the c-statistic (for logistic regression) that are presented, but are not usually considered as important. However, regression has limitations with large samples; all p-values are statistically significant with an effect size of virtually zero. For this reason, we need to be careful when interpreting the model. Instead, we can take a different approach. Because there are so many data values available, we can divide them and create holdout samples. Then, when using predictive modeling, we can use many different models simultaneously, and compare them to find the one that is the best. We can use the traditional regression, but also decision trees and neural network analysis. We can also combine different models. We can focus on accuracy of prediction rather than just identifying risk factors.

In particular, we will discuss some of the issues that are involved when using both linear and logistic regression. Regression requires an assumption of normality. The definition of confidence intervals, too, requires normality. However, most healthcare data are exponential or gamma. According to the Central Limit Theorem, the sample mean can be assumed normal if the sample is sufficiently large. However, if the distribution is exponential, just how large is large enough? If we use nonparametric models, we do not have to be as concerned with the actual population distribution.

Additional assumptions for regression are that the mean of the error term is equal to zero, and that the error term has equal variance for different levels of the input or independent variables. While the assumption of zero mean is almost always satisfied, the assumption of equal variance is not. Often, as the independent variables increase in value, the variance often increases as well. Therefore, modifications are needed to the variables, usually in the form of transformations, substituting the log of an independent variable for the variable itself. Transformations require considerable experience to use properly. In addition, the independent variables are assumed to be independent of each other. While the model can tolerate some correlation between these variables, too much correlation will result in a poor model that cannot be used effectively on fresh data. A similar problem occurs if the independent variables have different range scales.

Probably the most worrisome is the assumption that the error terms are identically distributed. In order for this assumption to be valid, we must assume the uniformity of data entry. That means that all providers must enter poorly defined values in exactly the same way. Unfortunately, such an assumption cannot possibly be valid. Consider, for example, the condition of "uncontrolled diabetes," which is one coded patient condition. The term, "uncontrolled" is not defined. Therefore, the designation remains at the discretion of the provider to define the term. For this reason, different providers will define it differently.

We use a medical dataset to examine the Central Limit Theorem. The dataset is available from the National Inpatient Sample (NIS). While publicly available, there is a small fee required from the Healthcare Cost and Utilization Project (HCUP, http://www.ahrq.gov/data/hcup/) in order to use them ($20 for students per year; $200 for non-students). In addition, the user must complete a statement of use indicating that they understand what can and cannot be disclosed. The NIS contains information concerning hospital inpatient stays, currently from a total of 37 states. Specific information about the datasets is available at http://www.ahrq.gov/data/hcup/datahcup.htm. Information on the acquisition of the data is available at http://www.hcup-us.ahrq.gov/tech_assist/centdist.jsp. The advantage of using this dataset is that the outcome variables are all heavy-tailed with gamma or exponential distributions. Therefore, we can use a series of random samples of the dataset to investigate the properties of the Central Limit Theorem for such a population distribution.

DATA VISUALIZATION

We start with Figure 1, the bar graph of the hospital length of stay for all patients with diabetes. In the National Inpatient Sample, that includes just over 1 million patient stays. We use a bar graph to examine length of stay. Note that the distribution has a very heavy tail with the maximum stay of about 354 days. The average length of stay is equal to 5 days with a standard deviation of 5.8 days. Figure 2 reduces the values on the x-axis to a maximum of 50. In Figure 2, note the gaps that occur because of rounding in the length of stay.

Figure 1. Length of Hospital Stay for Patients with Diabetes

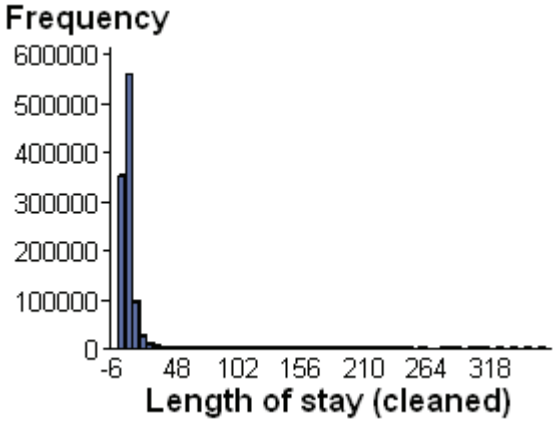


Figure 2. Length of Stay Limited to 50 Maximum

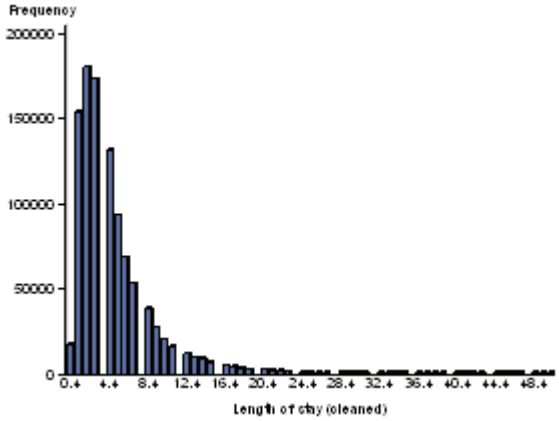


Figure 3 gives the best normal estimate of the population distribution. It significantly under-values the probability at the lower levels of length of stay, but also does not adequately estimate the outliers.

Figure 3. Normal Estimate of Population Distribution

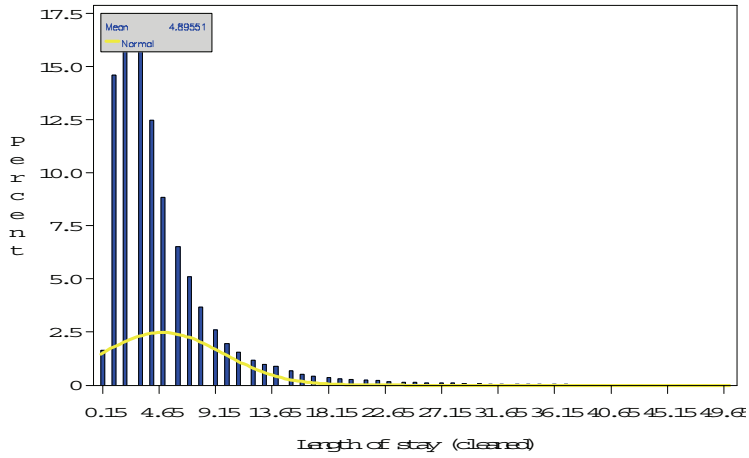
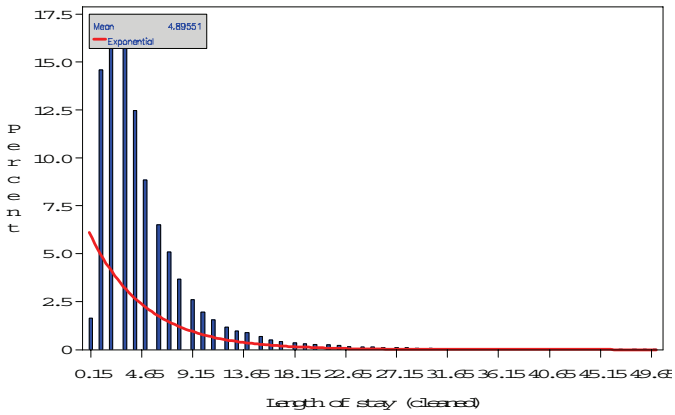


Figure 4 gives an exponential distribution estimate. It better follows the pattern of the bar graph, but it still under-values the height of the bars.

Figure 4. Exponential Estimate of Population Distribution



KERNEL DENSITY ESTIMATION

When a known distribution does not work to estimate the population, we can just use an estimate of that distribution. The histogram in Figures 1-4 can be smoothed into a probability density function. The formula for computing a kernel density estimate at the point x is equal to

$$f(x) = \frac{1}{na_n} \sum_{j=1}^{n} K\left(\frac{x - X_j}{a_n}\right)$$

where n is the size of the sample and K is a known density function. The value a_n is called the bandwidth. It controls the level of smoothing of the estimate curve. As the value of a_n approaches zero, the curve, f(x), becomes very jagged. As the value of a_n approaches infinity, the curve becomes closer to a straight line.

There are different methods available that can be used to attempt to optimize the level of smoothing. However, the value of a_n may still need adjustments, so SAS has a mechanism to allow you to do just that. Note that for most standard density functions, K, where x is far in magnitude from any point X_j, the value of $\sum K\left(\frac{x - X_j}{a_n}\right)$ will be very small. Where many data points cluster together, the value of the density function will be high because the sum of $x\text{-}X_j$ will be large, so that the probability defined by the kernel function will be large. However, where there are only scattered points, the value will be small. K can be the standard normal density, or the uniform density. Simulation studies have demonstrated that the value of K has very limited impact on the value of the density estimate. It is the value of the bandwidth, a_n, that has substantial impact on the value of the density estimate. The true value of this bandwidth must be estimated, and there are several methods available to optimize this estimate. The SAS code used to define this kernel density function is given below:

```
proc kde data=nis.diabetesless50los;
univar los/gridl=0 gridu=50 method=srot out=nis.kde50 bwm=3;
run;
```

We specify lower and upper grid values to bound the estimate. The method=srot attempts to optimize the level of smoothness in the estimate. The option, bwm=3, allows you to modify the optimal smoothness. The 'bwm' stands for bandwidth multiplier. With bwm=3, you take the value of an optimal bandwidth computed through the srot method (discussed below) and multiply it by 3 to increase the smoothness of the graph. The resulting estimate is saved in the nis.kde50 dataset so that it can be graphed. The result is given in Figure 5. Without the bwm=3 option, the estimate appears more jagged (Figure 6).

Figure 5. Kernel Density Estimate of Length of Stay for Patients With Diabetes

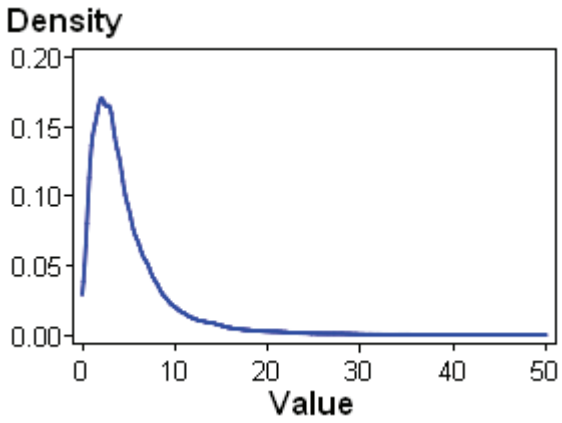
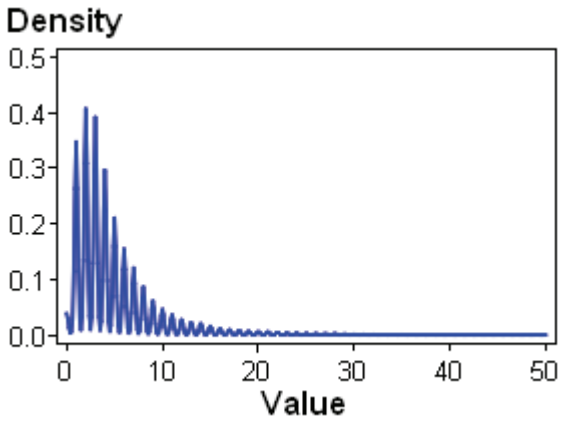


Figure 6. Kernel Density Estimate of Length of Stay for Patients With Diabetes Without Modifying the Level of Smoothness



PROC KDE uses only the standard normal density for K, but allows for several different methods to estimate the bandwidth, as discussed below. The default for the univariate smoothing is that of Sheather-Jones plug in (SJPI):

$$h = C_3\left\{\int f''(x)^2 dx, \int f'''(x)^2 dx\right\} C_4(K) h^{5/7}$$

where C_3 and C_4 are appropriate functionals. The unknown values that depend upon the density function f(x) are estimated with bandwidths chosen by reference to a parametric family such as the Gaussian as provided in Silverman:(Silverman, 1986)

$$\int f''(x)^2 dx = \sigma^{-5}\int \phi''(x)^2 dx \approx 0.212\sigma^{-5}$$

However, the procedure uses a different estimator, the simple normal reference (SNR), as the default for the bivariate estimator:

$$h = \hat{\sigma}\left[4/(3n)\right]^{1/5}$$

along with Silverman's rule of thumb (SROT):

$$h = 0.9\min[\hat{\sigma}, (Q_1 - Q_3)/1.34]n^{-1/5}$$

and the over-smoothed method (OS):

$$h = 3\hat{\sigma}\left[1/70\sqrt{\pi}n\right]^{1/5}$$

Figure 7. Kernel Density Estimate of Length of Stay for Patients With Diabetes and BWM=10.

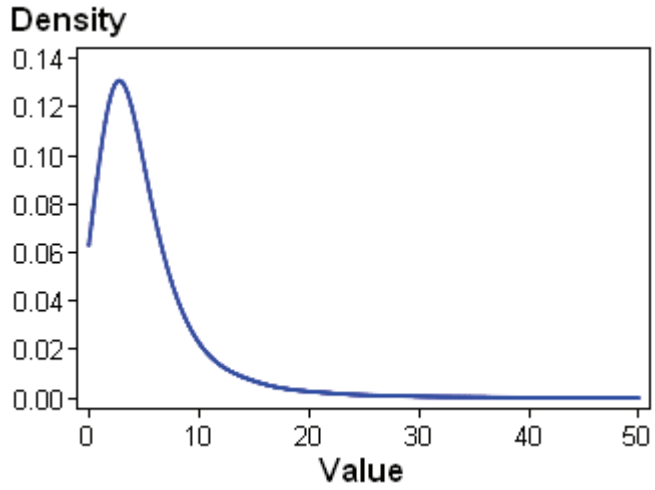
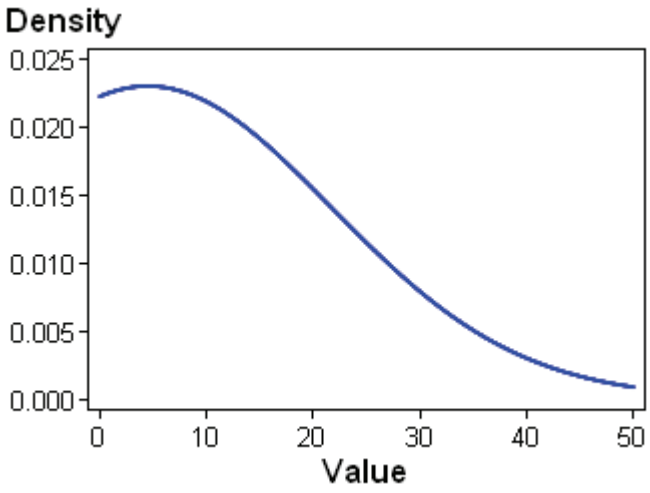


Figure 7 uses bwm=10 to increase the level of smoothing. It appears to have the most optimal level of smoothness. For a bwm<1, the curve becomes more jagged; for bwm>1, it becomes smoother. However, it can be too smooth. Figure 8 has a bwm of 100.

Figure 8. Kernel Density Estimate of Length of Stay for Patients With Diabetes and BWM=100.



CENTRAL LIMIT THEOREM

We must consider just how large n has to be for the Central Limit Theorem to be valid.(Battioui, 2007) To examine the issue, we take samples of different sizes to compute the distribution of the sample mean. The following code will compute 100 mean values from sample sizes starting with 5 and increasing to 10,000.

```
PROC SURVEYSELECT DATA=nis.nis_205 OUT=work.samples METHOD=SRS N=5 rep=100
noprint;
RUN;
proc means data=work.samples noprint;
  by replicate;
  var los;
  output out=out mean=mean;
run;
```

We change the value of N=5 in the first code statement to change the sample size. Once we have computed the means, we can graph them using kernel density estimation. We show the difference between the distribution of the population, and the distribution of the sample mean for the differing sample sizes. Figures 9-12 show the distribution of the sample mean compared to the distribution of the population for differing sample sizes. To compute the distribution of the sample mean, we collect 100 different samples using the above code. We compute the mean for the patient length of stay using the National Inpatient Sample.

Figure 9. Sample Mean With Sample=5

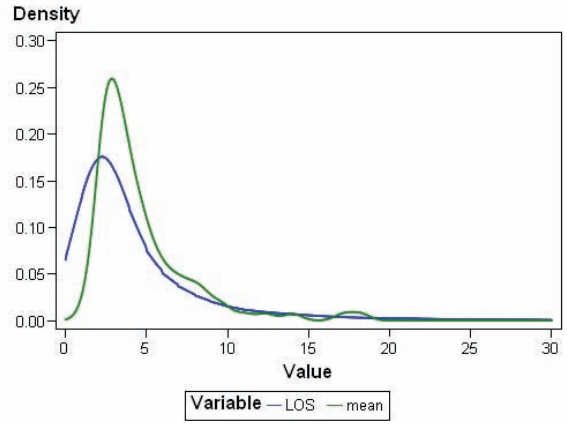


Figure 10. Sample Mean With Sample=30

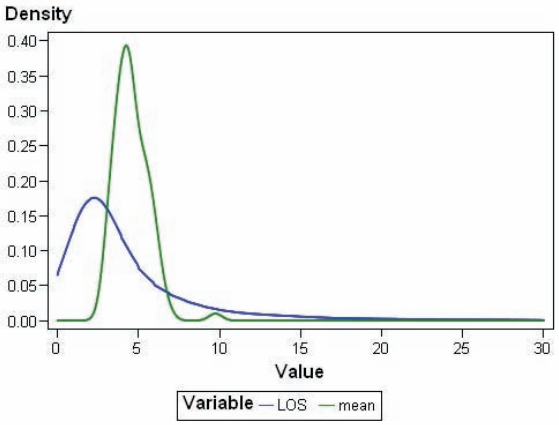


Figure 11. Sample Mean With Sample=100

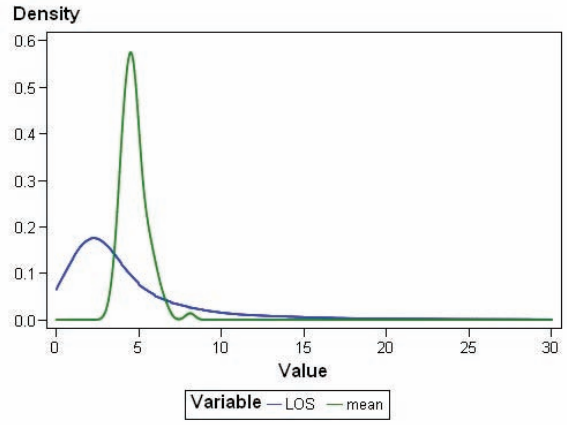
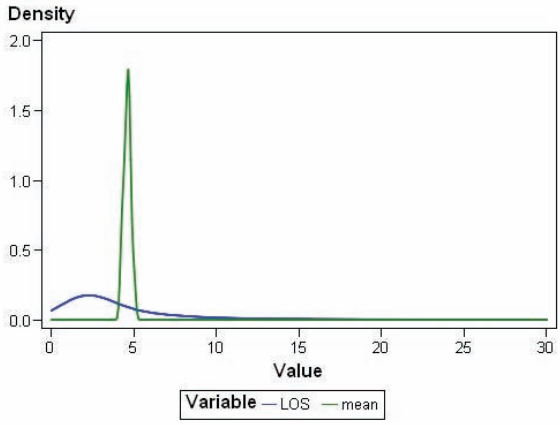


Figure 12. Sample Mean With Sample=1000



In Figure 9, the sample mean peaks slightly to the right of the peak of the population distribution; this peak is much more exaggerated in Figure 10. The reason for this shift in the peak is because the sample mean is susceptible to the influence of outliers, and the population is very skewed. Because it is so skewed, the distribution of the sample mean is not entirely normal. As the sample increases to 100 and then to 1000, this shift from the population peak to the sample peak becomes much more exaggerated. We use the same sample sizes for 1000 replicates (Figures 13-16).

Figure 13. Sample Mean for Sample Size=5 and 1000 Replicates

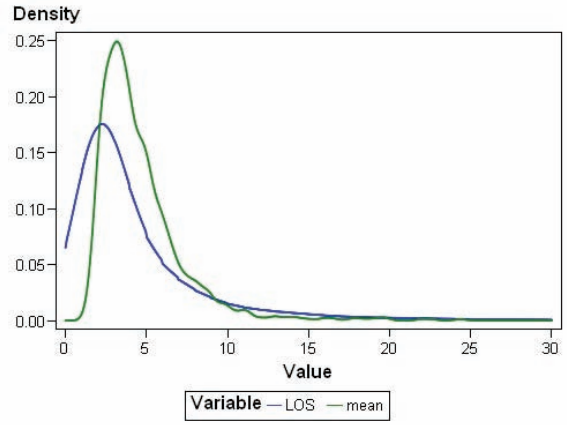


Figure 14. Sample Mean for Sample Size=30 and 1000 Replicates

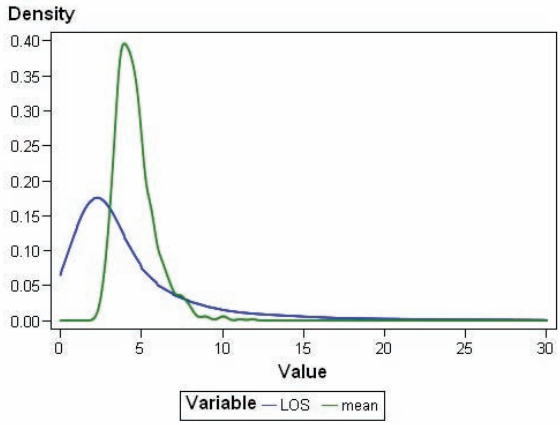


Figure 15. Sample Mean for Sample Size=100 and 1000 Replicates

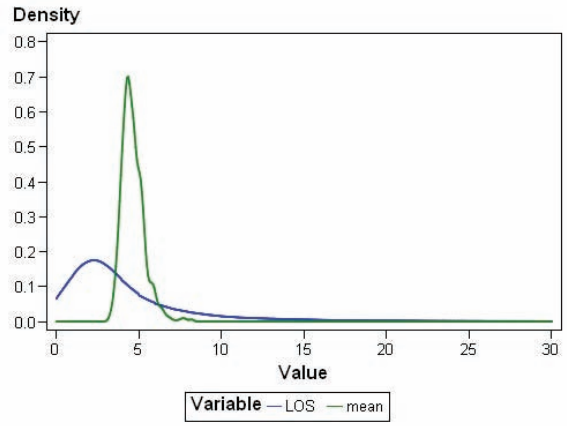
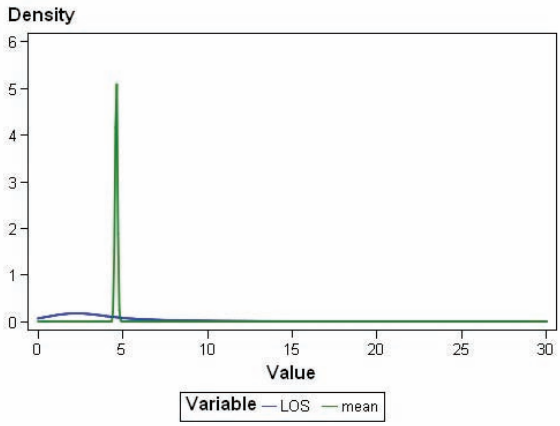


Figure 16. Sample Mean for Sample Size=1000 and 1000 Replicates



It is again noticeable that the sample mean has shifted away from the peak value of the population distribution because of the skewed distribution. However, the distribution of the mean is not normally distributed.

OUTLIERS IN REGRESSION

An assumption of normality is required for regression. If we assume normality when the distribution is exponential or gamma, the outliers will be under-counted. Consider the dataset here that is not normally distributed. We use a random sample of 1000 observations. The mean and standard deviation (assuming a normal distribution) are equal to 4.235 and 5.0479 respectively. Then, three standard deviations beyond the mean is equal to 19.3787 days. Two standard deviations beyond the mean is equal to 14.3308. In the random sample, the proportion of days beyond two standard deviations is equal to 35 when the normal probability indicates only 25 should be that large. The proportion beyond three standard deviations is equal to 20; the probability indicates that only 10 should be beyond that point. We also look at the outlier charges and the cost-to-charge ratio as determined by the hospital. The cost-to-charge ratio by hospital is provided in the National Inpatient Sample data. A cost-to-charge is the ratio of patient costs to the charges billed by the hospital. A ratio less than one indicates that the charges are much larger than costs; a ratio larger than on indicates the opposite. Figure 17 shows the variability in the cost-to-charge ratio across the different hospitals. Note the considerable variability from 0.1 to 0.8. Hospitals with a higher cost-to-charge ratio tend to bill charges that are more unreasonably in line with actual costs compared to hospitals with a rate of 1.0. Figure 18 shows a comparison between charges and charges x cost ratio.

Figure 17. Cost-to-Charge Ratio

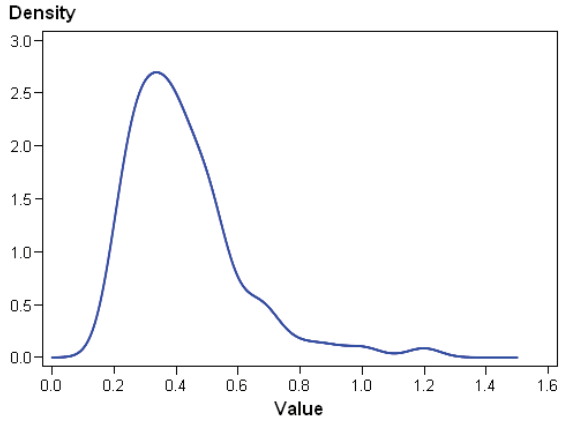
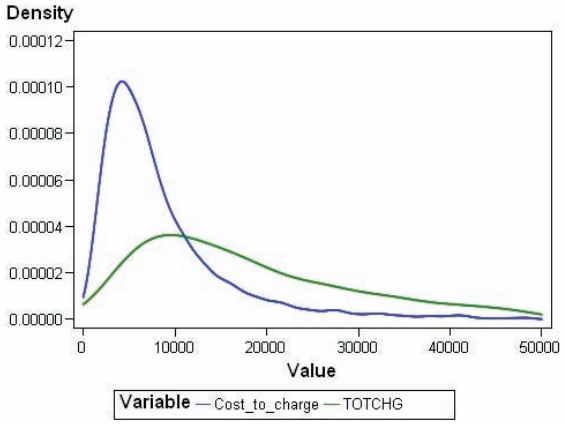


Figure 18. Comparison of Charges to Estimated Costs

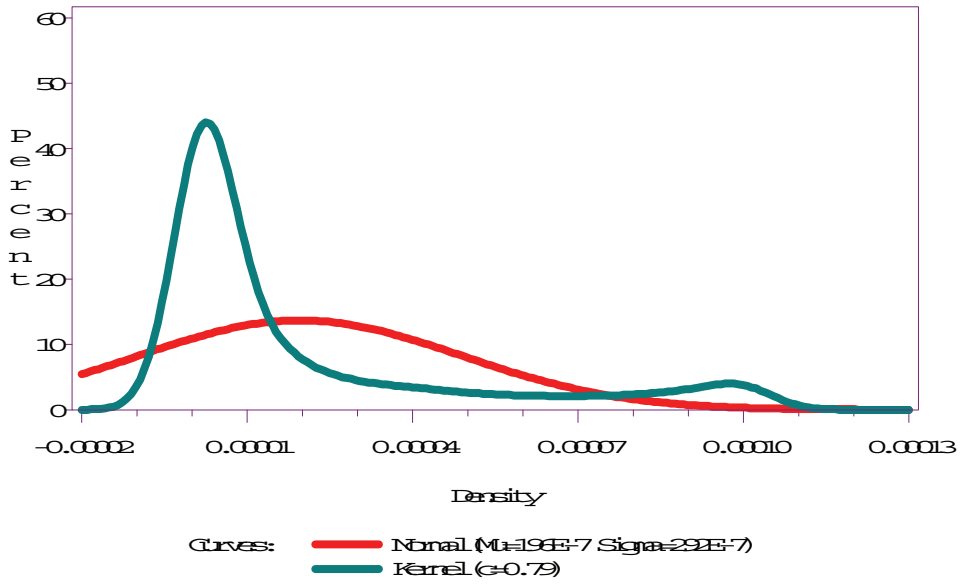


The cost-to-charge ratio, then, reduces the probability in the tail, but does not reduce its size. If we compare the kernel density graph of the cost-to-charge compared to the normal distribution assumption (Figure 19), it is clear that the assumption of normality will under-count the outliers. The following code is used to compare the kernel density to the normal distribution:

```
PROC CAPABILITY DATA = WORK.SORTTempTableSorted
             CIBASIC(TYPE=TWOSIDED ALPHA=0.05)
             MU0=0;
       VAR density;
       ;
        HISTOGRAM density / NORMAL ( W=10 L=1 COLOR=red MU=EST SIGMA=EST)
                            KERNEL ( W=10 L=1 COLOR=CX008080 C= MISE K=NORMAL)
                            NOBARS

        CAXIS=PURPLE
        CTEXT=BLACK CFRAME=WHITE
        CBARLINE=BLACK
        CFILL=GRAY;
         RUN;
```

Figure 19. Comparison of Kernel Estimate to Normality Assumption



A normal assumption increases the variance, but fails to count many of the extreme outliers. Therefore, hospital reimbursement formulas need to take the gamma population distribution into consideration in order to account for the population outliers.

LOGISTIC REGRESSION

We want to see if we can predict mortality in patients using a logistic regression model. There is considerable incentive to increase the number of positive indicators, called upcoding. The value,

$$\alpha_1 X_2 + \alpha_2 X_2 + \ldots + \alpha_{25} X_{25}$$

increases as the number of nonzero X's increases. The greater this value, the greater the likelihood that it will cross the threshold value that predicts mortality. However, consider for a moment that just about every patient condition has a small risk of mortality. Once the threshold value is crossed, every patient with similar conditions are predicted to die. Therefore, the more patients who can be defined over the threshold value, the higher the predicted mortality rate, decreasing the difference between predicted and actual mortality. There is considerable incentive to upcode patient diagnoses to increase the likelihood of crossing this threshold value. To simplify, we start with just one input variable to the logistic regression; the occurrence of pneumonia. Table 1 gives the chi-square table for the two variables.

Table 1. Chi-square Table for Mortality by Pneumonia

Table of pneumonia by DIED			
pneumonia	**DIED**		**Total**
Frequency **Row Pct** **Col Pct**	**0**	**1**	
0	7431129 98.21 94.97	135419 1.79 81.02	7566548
1	393728 92.54 5.03	31731 7.46 18.98	425459
Total	7824857	167150	7992007
Frequency Missing = 3041			

Approximately 7% of the patients with pneumonia died compared to just under 2% generally. However, if we consider the classification table (Table 2) for a logistic regression with pneumonia as the input and mortality as the outcome variable, the accuracy rate is above 90% for any choice of threshold value of less than 1.0, where 100% of the values are to predict non-mortality. Therefore, even though patients with pneumonia are almost 4 times as likely to die compared to patients without pneumonia, pneumonia by itself is a poor predictor of mortality because of the rare occurrence.

Table 2. Classification Table for Logistic Regression

Classification Table									
Prob Level	**Correct**		**Incorrect**		**Percentages**				
	Event	**Non-Event**	**Event**	**Non-Event**	**Correct**	**Sensi-tivity**	**Speci-ficity**	**False POS**	**False NEG**
0.920	782E4	0	167E3	0	97.9	100.0	0.0	2.1	.
0.940	743E4	31731	135E3	394E3	93.4	95.0	19.0	1.8	92.5
0.960	743E4	31731	135E3	394E3	93.4	95.0	19.0	1.8	92.5
0.980	743E4	31731	135E3	394E3	93.4	95.0	19.0	1.8	92.5
1.000	0	167E3	0	782E4	2.1	0.0	100.0	.	97.9

We now add a second patient diagnosis to the regression. Table 3 gives the chi-square table for pneumonia and septicemia.

Table 3. Chi-square Table for Pneumonia and Septicemia

Controlling for septicemia=0				**Controlling for septicemia=1**		
pneumonia	**Died**		**Total**	**DIED**		**Total**
Frequency Row Pct Col Pct	**0**	**1**		**0**	**1**	
0	7307726 98.60 95.20	103759 1.40 82.65	7411485	123403 79.58 83.06	31660 20.42 76.09	155063
1	368553 94.42 4.80	21783 5.58 17.35	390336	25175 71.68 16.94	9948 28.32 23.91	35123
Total	7676279	125542	7801821	148578	41608	190186

Of the patients with septicemia only (pneumonia=0), 20% died, increasing to 28% with both septicemia and pneumonia. For patients without septicemia but with pneumonia, 5% died. The classification table for the logistic regression is given in Table 4.

Table 4. Classification Table for Logistic Regression With Pneumonia and Septicemia

Classification Table									
Prob Level	**Correct**		**Incorrect**		**Percentages**				
	Event	**Non-Event**	**Event**	**Non-Event**	**Correct**	**Sensi-tivity**	**Speci-ficity**	**False POS**	**False NEG**
0.580	782E4	0	167E3	0	97.9	100.0	0.0	2.1	.
0.600	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.620	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.640	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.660	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7

Classification Table									
Prob Level	Correct		Incorrect		Percentages				
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.680	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.700	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.720	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.740	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.760	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.780	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.800	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.820	78E5	9948	157E3	25175	97.7	99.7	6.0	2.0	71.7
0.840	768E4	41608	126E3	149E3	96.6	98.1	24.9	1.6	78.1
0.860	768E4	41608	126E3	149E3	96.6	98.1	24.9	1.6	78.1
0.880	768E4	41608	126E3	149E3	96.6	98.1	24.9	1.6	78.1
0.900	768E4	41608	126E3	149E3	96.6	98.1	24.9	1.6	78.1
0.920	768E4	41608	126E3	149E3	96.6	98.1	24.9	1.6	78.1
0.940	768E4	41608	126E3	149E3	96.6	98.1	24.9	1.6	78.1
0.960	731E4	63391	104E3	517E3	92.2	93.4	37.9	1.4	89.1
0.980	731E4	63391	104E3	517E3	92.2	93.4	37.9	1.4	89.1
1.000	0	167E3	0	782E4	2.1	0.0	100.0	.	97.9

Again, for any threshold value below 98%, the logistic regression model will be over 90% accurate by identifying most of the observations as non-occurrences so that the false negative rate is over 70%. In other words, adding a second input variable did not change the problems with the regression, which are caused by attempting to predict a rare occurrence. We add Immune Disorder to the model (Table 5).

Table 5. Classification Table Adding Immune Disorder

Classification Table									
Prob Level	Correct		Incorrect		Percentages				
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.480	782E4	0	167E3	0	97.9	100.0	0.0	2.1	.
0.500	781E4	4907	162E3	11633	97.8	99.9	2.9	2.0	70.3
0.520	781E4	4907	162E3	11633	97.8	99.9	2.9	2.0	70.3
0.540	781E4	4907	162E3	11633	97.8	99.9	2.9	2.0	70.3
0.560	781E4	4907	162E3	11633	97.8	99.9	2.9	2.0	70.3
0.580	781E4	4907	162E3	11633	97.8	99.9	2.9	2.0	70.3
0.600	781E4	4907	162E3	11633	97.8	99.9	2.9	2.0	70.3
0.620	781E4	4907	162E3	11633	97.8	99.9	2.9	2.0	70.3
0.640	781E4	4907	162E3	11633	97.8	99.9	2.9	2.0	70.3

Classification Table									
Prob Level	Correct		Incorrect		Percentages				
	Event	Non-Event	Event	Non-Event	Correct	Sensi-tivity	Speci-ficity	False POS	False NEG
0.660	781E4	4907	162E3	11633	97.8	99.9	2.9	2.0	70.3
0.680	781E4	4907	162E3	11633	97.8	99.9	2.9	2.0	70.3
0.700	781E4	4907	162E3	11633	97.8	99.9	2.9	2.0	70.3
0.720	781E4	4907	162E3	11633	97.8	99.9	2.9	2.0	70.3
0.740	776E4	21322	146E3	65076	97.4	99.2	12.8	1.8	75.3
0.760	775E4	26363	141E3	78618	97.3	99.0	15.8	1.8	74.9
0.780	775E4	26363	141E3	78618	97.3	99.0	15.8	1.8	74.9
0.800	775E4	26363	141E3	78618	97.3	99.0	15.8	1.8	74.9
0.820	775E4	26363	141E3	78618	97.3	99.0	15.8	1.8	74.9
0.840	775E4	26363	141E3	78618	97.3	99.0	15.8	1.8	74.9
0.860	775E4	26363	141E3	78618	97.3	99.0	15.8	1.8	74.9
0.880	775E4	26363	141E3	78618	97.3	99.0	15.8	1.8	74.9
0.900	768E4	41608	126E3	149E3	96.6	98.1	24.9	1.6	78.1
0.920	757E4	51297	116E3	258E3	95.3	96.7	30.7	1.5	83.4
0.940	757E4	51297	116E3	258E3	95.3	96.7	30.7	1.5	83.4
0.960	757E4	51297	116E3	258E3	95.3	96.7	30.7	1.5	83.4
0.980	634E4	103E3	64219	149E4	80.6	81.0	61.6	1.0	93.5
1.000	0	167E3	0	782E4	2.1	0.0	100.0	.	97.9

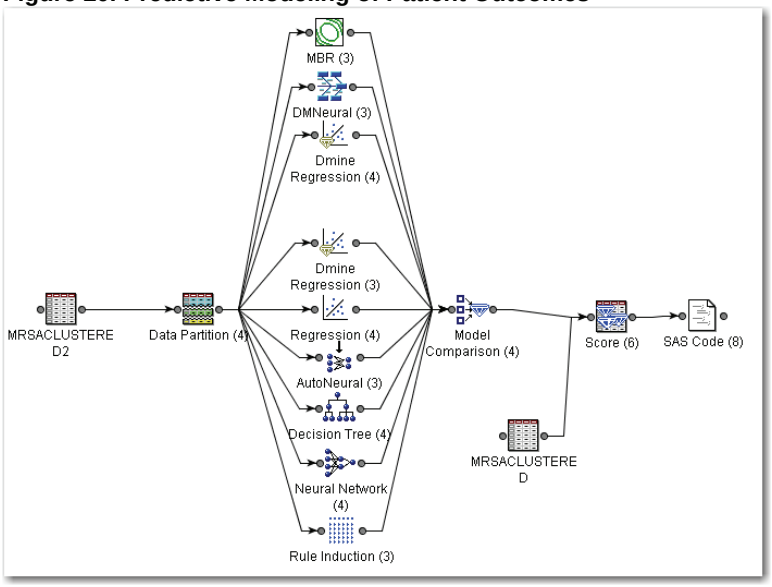
The problem still persists, and will continue to persist regardless of the number of input variables. We need to change the sample size so that the group sizes are close to equal.

PREDICTIVE MODELING IN SAS ENTERPRISE MINER

Figure 1 gives a diagram of a predictive model in SAS Enterprise Miner. Enterprise Miner includes the standard types of regression, artificial neural networks, and decision trees. The regression model will choose linear or logistic automatically, depending upon the type of outcome variable. Figure 20 shows that many different models can be used. Once defined, the models are compared and the optimal model chosen based upon pre-selected criteria. Then, additional data can be scored so that patients, in this example, at high risk for adverse events can be identified for more aggressive treatment.

The purpose of the partition node in Figure 20 is to divide the data into training, validation, and testing subsets, by default, a 40/30/30 split in the data. Usually, the datasets are large enough that such a partitioning is possible. The training set is used to define the model; the testing set is a holdout sample used as fresh data to test the accuracy of the model. The validation set is not needed for regression; it is needed for neural networks and any model that is defined iteratively. The model is examined on the validation set, and adjustments are made to the model if necessary. This process is repeated until no more changes are necessary.

Figure 20. Predictive Modeling of Patient Outcomes



For predicting a rare occurrence, one more node is added to the model in Figure 20, the sampling node (Figure 21). This node uses all of the observations with the rare occurrence, and then takes a random sample of the remaining data. While the sampling node can use any proportional split, we recommend a 50:50 split. Figure 22 shows how the defaults are modified in the sampling node of SAS Enterprise Miner to make predictions.

Figure 21. Addition of Sampling Node

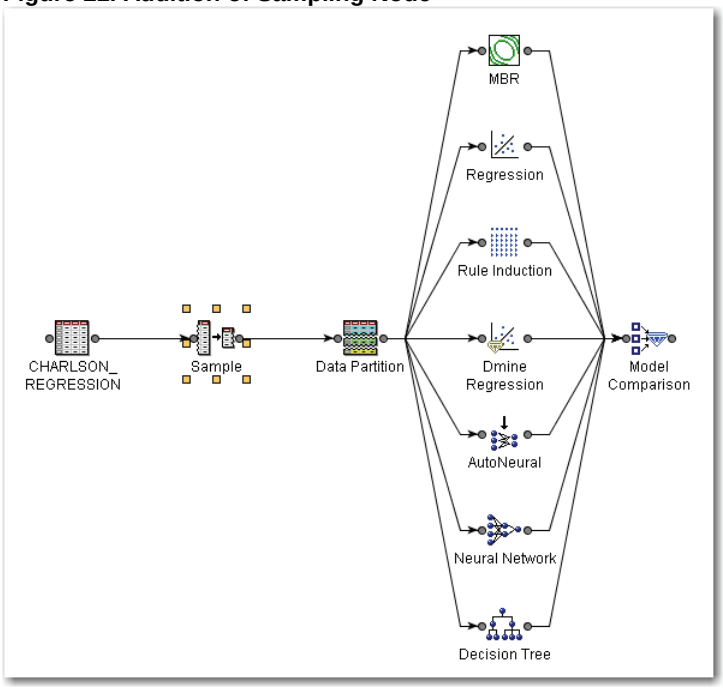
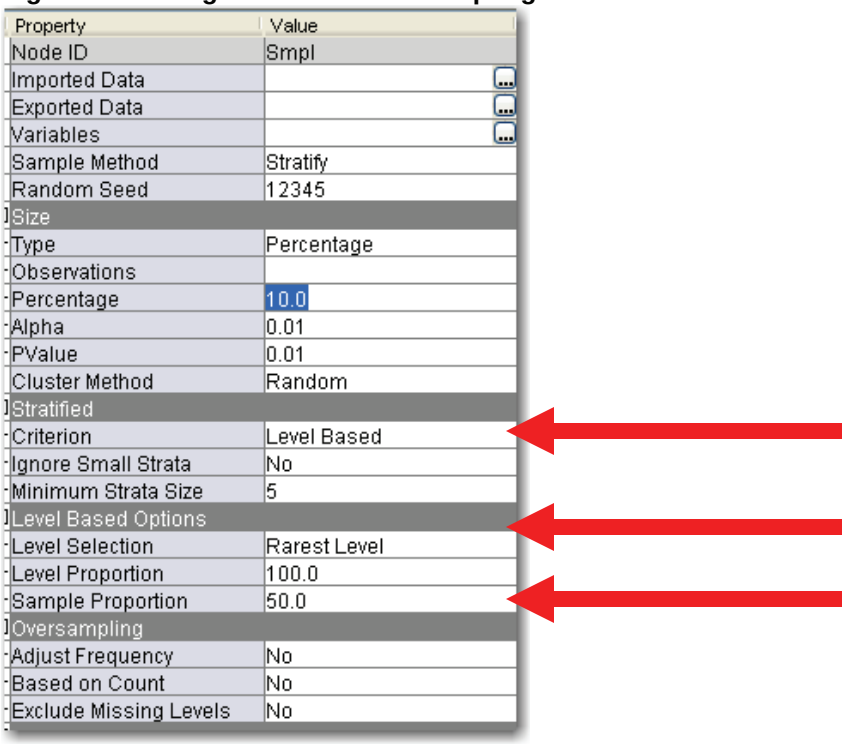


Figure 22. Change to Defaults in Sampling Node



The first arrow indicates that the sampling is stratified, and the criterion is level based. The rarest level (in this case, mortality) is sampled so that it will consist of half (50% sample proportion) of the sample.

Consider the problem of predicting mortality that was discussed in the previous section on logistic regression. We use just the same three patient diagnoses of pneumonia, septicemia, and immune disorder that we used previously. However, in this case, we use the sampling node to get a 50/50 split in the data.

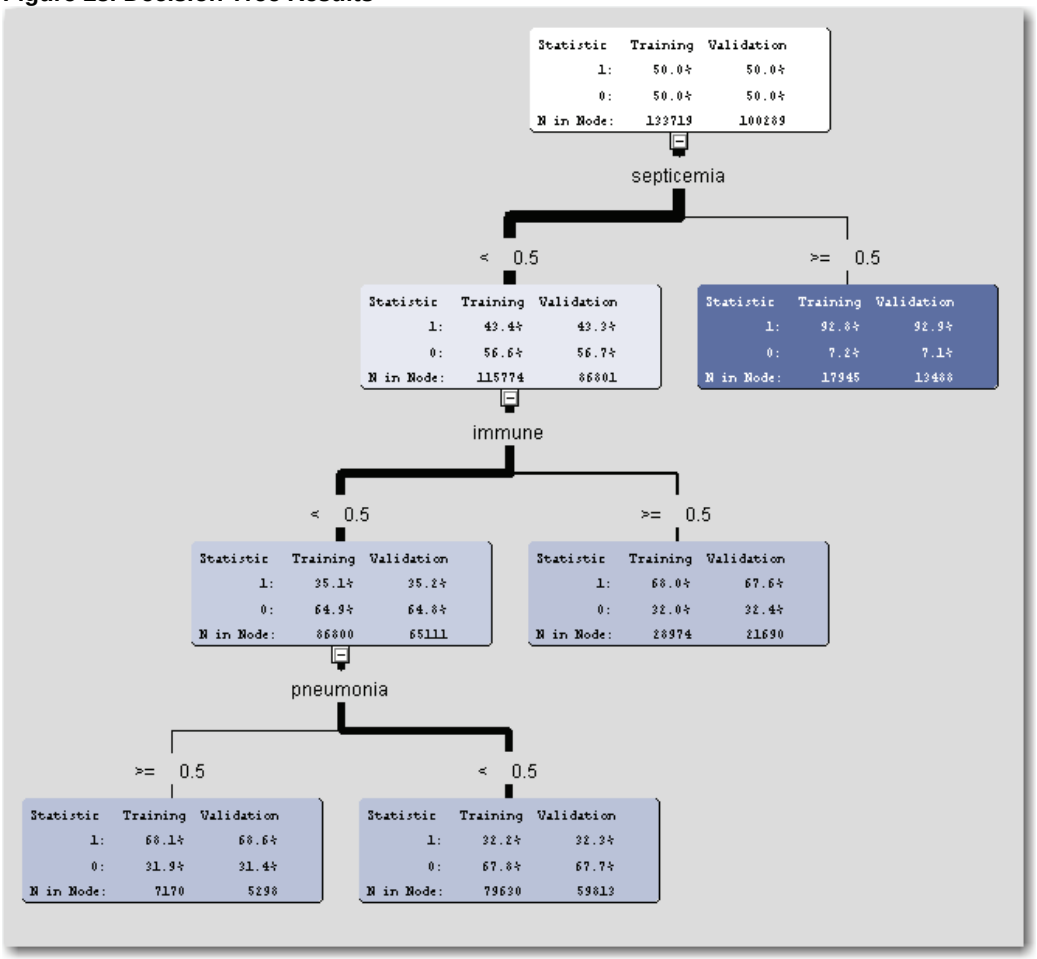
We use all of the models depicted in Figure 20. According to the model comparison, the rule induction provides the best fit, using the misclassification rate as the measure of "best". We first look at the regression model, comparing the results to those in the previous chapter when a 50/50 split was not performed. The overall misclassification rate is 28%, with the divisions as shown in Table 6.

Table 6. Misclassification in Regression Model

Target	Outcome	Target Percentage	Outcome Percentage	Count	Total Percentage
Training Data					
0	0	67.8	80.1	54008	40.4
1	0	32.2	38.3	25622	19.2
0	1	23.8	19.2	12852	9.6
1	1	76.3	61.7	41237	30.8
Validation Data					
0	0	67.7	80.8	40498	40.4
1	0	32.3	38.5	19315	19.2
0	1	23.8	19.2	9646	9.6
1	1	76.2	61.5	30830	30.7

Note that the misclassification becomes more balanced between false positives and false negatives with a 50/50 split in the data. The model gives heavier weight to false positives than it does to false negatives. We also want to examine the decision tree model. While it is not the most accurate model, it is one that clearly describes the rationale behind the predictions. This tree is given in Figure 23. The tree shows that the first split occurs on the variable, Septicemia. Patients with Septicemia are more likely to suffer mortality compared to patients without Septicemia. As shown in the previous section, the Immune Disorder has the next highest level of mortality followed by Pneumonia.

Figure 23. Decision Tree Results



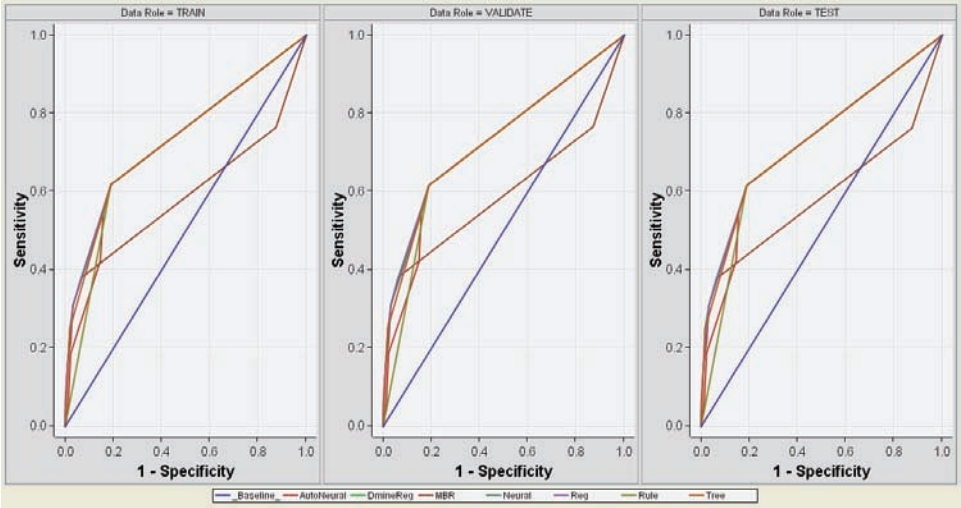
Since rule induction is identified as the best model, we examine that one next. The misclassification rate is only slightly smaller compared to the regression model. Table 7 gives the classification table.

Table 7. Misclassification in Rule Induction Model

Target	Outcome	Target Percentage	Outcome Percentage	Count	Total Percentage
Training Data					
0	0	67.8	80.8	54008	40.4
1	0	32.2	38.3	25622	19.2
0	1	23.8	19.2	12852	9.6
1	1	76.3	61.7	41237	30.8
Validation Data					
0	0	67.7	80.8	40498	40.4
1	0	32.3	38.5	19315	19.2
0	1	23.8	19.2	9646	9.6
1	1	76.2	61.5	30830	30.7

The results look virtually identical to those in Table 6. For this reason, the regression model, although not defined as the best, can be used to predict outcomes when only these three variables are used. The similarities in the models can also be visualized in the ROC (received-operating curve) that graphs the sensitivity versus one minus the specificity (Figure 24). The curves for rule induction and regression are virtually the same.

Figure 24. Comparison of ROC Curves



The above example only used three possible diagnosis codes. We want to expand upon the number of diagnosis codes, and also to use a number of procedure codes. In this example, we restrict our attention to patients with a primary diagnosis of COPD (chronic obstructive pulmonary disease resulting primarily from smoking). There are approximately 245,000 patients in the NIS dataset. Table 8 gives the list of diagnosis codes used; Table 9 gives a list of procedure codes used as well.

Table 8. Diagnosis Codes Used to Predict Mortality

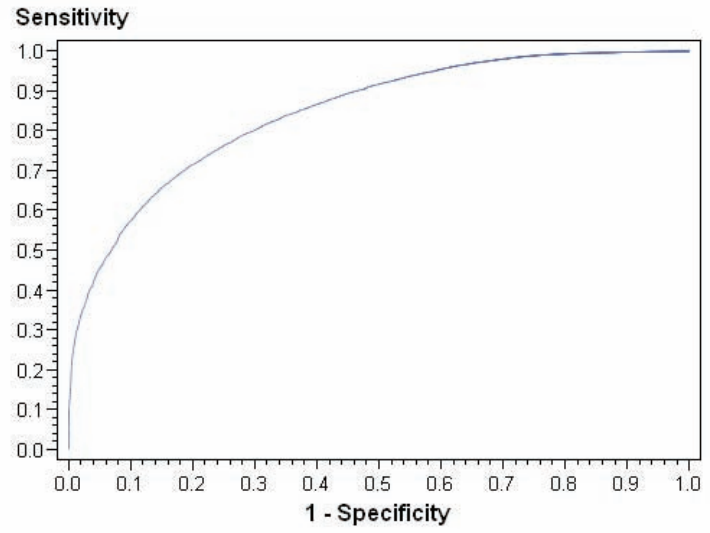
Condition	ICD9 Codes
Acute myocardial infarction	410, 412
Congestive heart failure	428
Peripheral vascular disease	441,4439,7854,V434
Cerebral vascular accident	430-438
Dementia	290
Pulmonary disease	490,491,492,493,494,495,496,500,501,502,503,504,505
Connective tissue disorder	7100,7101,7104,7140,7141,7142,7148,5171,725
Peptic ulcer	531,532,533,534
Liver disease	5712,5714,5715,5716
Diabetes	2500,2501,2502,2503,2507
Diabetes complications	2504,2505,2506
Paraplegia	342,3441
Renal disease	582,5830,5831,5832,5833,5835,5836,5837,5834,585,586,588
Cancer	14,15,16,17,18,170,171,172,174,175,176,179,190,191,193, 194,1950,1951,1952, 1953,1954,1955,1958,200,201,202,203, 204,205,206,207,208
Metastatic cancer	196,197,198,1990,1991
Severe liver disease	5722,5723,5724,5728
HIV	042,043,044

Table 9. Procedure Codes Used to Predict Mortality

pr	Procedure Translation	Frequency	Percent
9904	Transfusion of packed cells	17756	7.05
3893	Venous catheterization, not elsewhere classified	16142	6.41
9671	Continuous mechanical ventilation for less than 96 consecutive hours	10528	4.18
3324	Closed [endoscopic] biopsy of bronchus	8315	3.30
9672	Continuous mechanical ventilation for 96 consecutive hours or more	8243	3.27
3491	Thoracentesis	8118	3.22
3995	Hemodialysis	8083	3.21
9604	Insertion of endotracheal tube	7579	3.01
9921	Injection of antibiotic	6786	2.69
9394	Respiratory medication administered by nebulizer	6309	2.50
8872	Diagnostic ultrasound of heart	5419	2.15
4516	Esophagogastroduodenoscopy [EGD] with closed biopsy	4894	1.94
9390	Continuous positive airway pressure	4667	1.85
3327	Closed endoscopic biopsy of lung	3446	1.37
8741	Computerized axial tomography of thorax	3417	1.36
4513	Other endoscopy of small intestine	3277	1.30

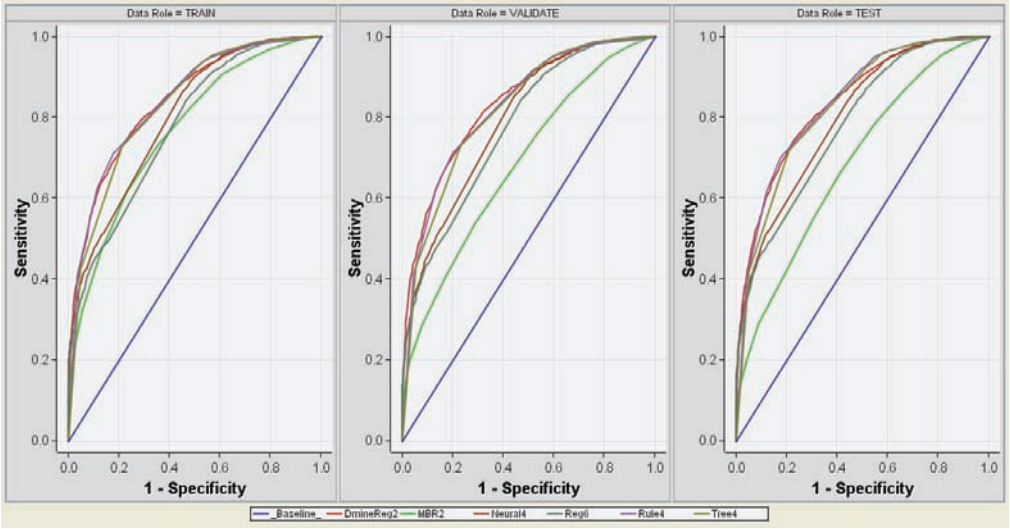
If we perform standard logistic regression without stratified sampling, the false positive rate remains small (approximately 3-4%), but with a high false negative rate (minimized at 38%). Given the large dataset, almost all of the input variables are statistically significant. The percent agreement is 84% and the ROC curve looks fairly good (Figure 25).

Figure 25. ROC Curve for Traditional Logistic Regression



If we perform predictive modeling, the accuracy rate drops to 75%, but the false negative rate is considerably improved. Figure 26 gives the ROC curve from predictive modeling.

Figure 26. ROC From Predictive Modeling



CHANGE IN SPLIT IN THE DATA

All of the analyses in the previous section assumed a 50/50 split between mortality and non-mortality. We want to look at the results if mortality composes only 25% of the data, and 10% of the data. Table 10 gives the regression classification breakdown for a 25% sample; Table 11 gives the breakdown for a 10% sample.

Table 10. Misclassification Rate for a 25% Sample

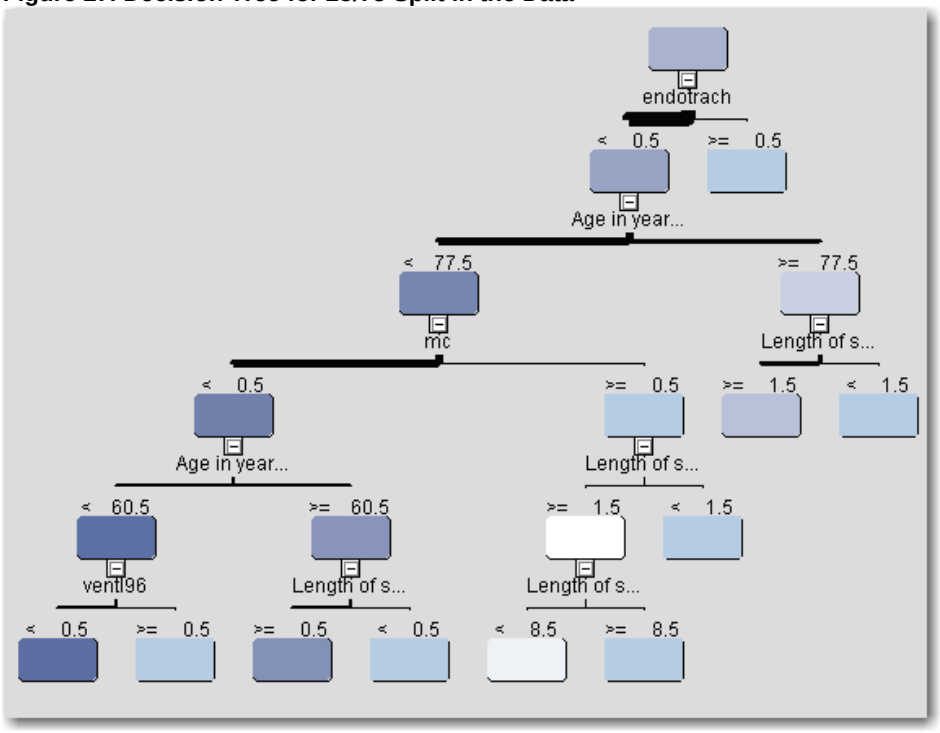
Target	Outcome	Target Percentage	Outcome Percentage	Count	Total Percentage
Training Data					
0	0	80.4	96.6	10070	72.5
1	0	19.6	70.9	2462	17.7
0	1	25.6	3.3	348	2.5
1	1	74.4	29.1	1010	7.3
Validation Data					
0	0	80.2	97.1	7584	72.8
1	0	19.8	71.7	1870	17.9
0	1	23.7	2.9	229	2.2
1	1	76.2	28.2	735	7.0

Note that the ability to classify mortality accurately is decreasing with the decrease of the split; almost all of the observations are classified as non-mortality. The decision tree (Figure 27) is considerably different from that with a 50/50 split. Now, the procedure of Esophagogastroduodenoscopy gives the first leaf of the tree.

Table 11. Misclassification Rate for a 10% Sample

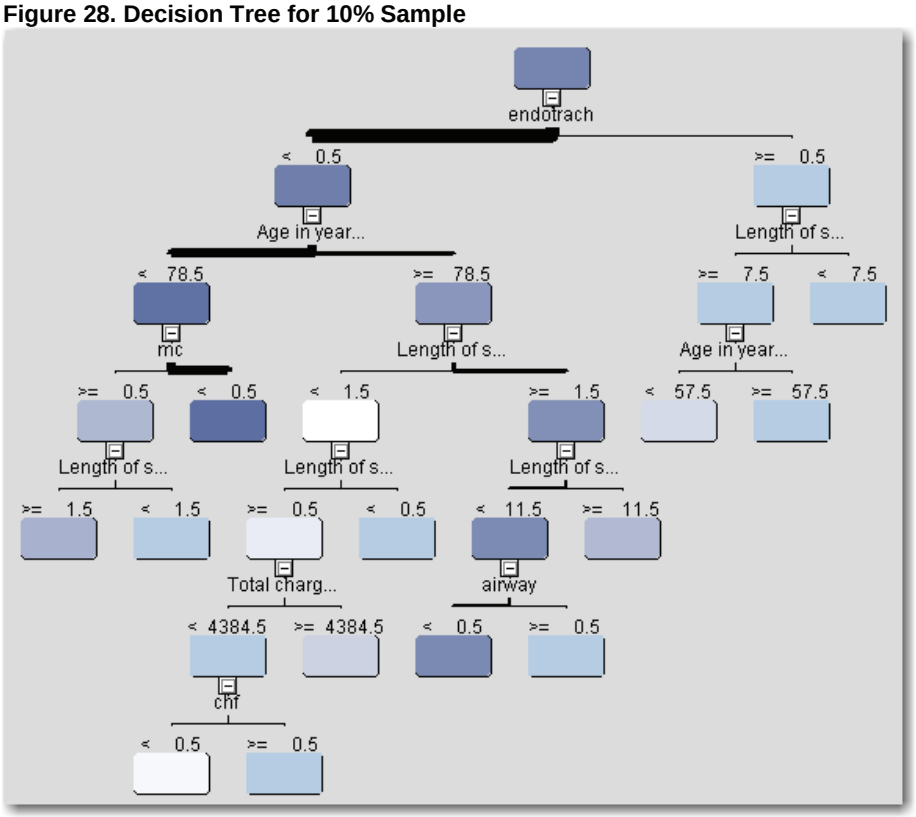
Target	Outcome	Target Percentage	Outcome Percentage	Count	Total Percentage
Training Data					
0	0	91.5	99.3	31030	89.4
1	0	8.5	83.5	2899	8.3
0	1	27.3	0.7	216	0.6
1	1	72.6	16.5	574	1.6
Validation Data					
0	0	91.5	99.2	23265	89.3
1	0	8.4	82.4	2148	8.2
0	1	27.8	0.7	176	0.7
1	1	72.2	17.5	457	1.7

Figure 27. Decision Tree for 25/75 Split in the Data



Note that the trend shown in the 25% is even more exaggerated in the 10% sample. Figure 28 shows that the decision tree has changed yet again. It now includes the procedure of continuous positive airway pressure and the diagnosis of congestive heart failure. AT a 1% sample, the misclassification becomes even more disparate.

Figure 28. Decision Tree for 10% Sample



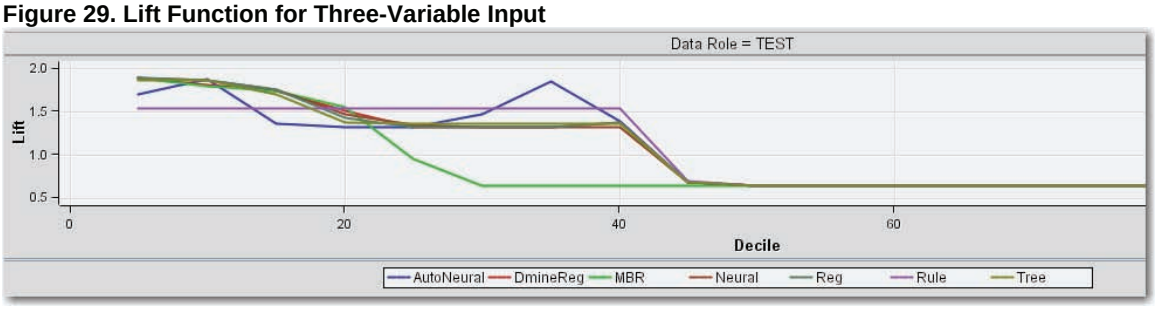
INTRODUCTION TO LIFT

Lift allows us to find the patients at highest risk for occurrence, and with the greatest probability of accurate prediction. This is especially important since these are the patients we would want to take the greatest care for. Using lift, true positive patients with highest confidence come first, followed by positive patients with lower confidence. True negative cases with lowest confidence come next, followed by negative cases with highest confidence. Based on that ordering, the observations are partitioned into deciles, and the following statistics are calculated:

- The *Target density* of a decile is the number of actually positive instances in that decile divided by the total number of instances in the decile.
- The *Cumulative target density* is the target density computed over the first *n* deciles.
- The *lift* for a given decile is the ratio of the target density for the decile to the target density over all the test data.
- The *Cumulative lift* for a given decile is the ratio of the cumulative target density to the target density over all the test data.

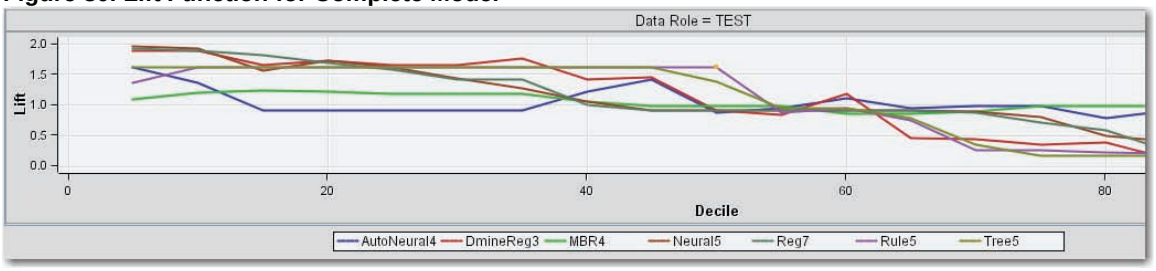
Given a lift function, we can decide on a decile cutpoint so that we can predict the high risk patients above the cutpoint, and predict the low risk patients below a second cutpoint, while failing to make a definite prediction for those in the center. In that way, we can dismiss those who have no risk, and aggressively treat those at highest risk. Lift allows us to distinguish between patients without assuming a uniformity of risk. Figure 29 shows the lift for the testing set when we use just the three input variables of pneumonia, septicemia, and immune disorder.

Figure 29. Lift Function for Three-Variable Input



Random chance is indicates by the lift value of 1.0; values that are higher than 1.0 indicate that the observations are more predictable compared to random chance. In this example, 40% of the patient records have a higher level of prediction than just chance. Therefore, we can concentrate on these 4 deciles of patients. If we use the expanded model that includes patient demographic information plus additional diagnosis and procedure codes for COPD, we get the lift shown in Figure 30. The model can now predict the first 5 deciles of patient outcomes.

Figure 30. Lift Function for Complete Model



Therefore, we can predict accurately those patients most at risk for death; we can determine which patients can benefit from more aggressive treatment to reduce the likelihood that this outcome will occur.

DISCUSSION

Given large datasets and the presence of outliers, the traditional statistical methods are not always applicable or meaningful. Assumptions can be crucial to the applicability of the model, and assumptions are not always carefully considered. The assumptions of a normal distribution and uniformity of data entry are crucial and need to be considered carefully.

The data may not give high levels of correlation, and regression may not always be the best way to measure associations. It must also be remembered that associations in the data as identified in regression models do not demonstrate cause and effect.

Kernel density remains a rarely used technique in the medical literature to investigate population distributions. When it is used listed in a Medline article, it is usually in a technical, non-clinical journal to show improvements in the methodology (Hong, Chen, & Chris J Harris, 2008; Pfeiffer, 1985), or in DNA studies. (Fu, Borneman, Ye, & Chrobak, 2005) Nevertheless, medicine must and will focus more on the study of outlier patients; patients with extreme conditions instead of focusing just on the average or typical patient. Outlier patient costs can often overwhelm the system even when they form just a small percentage of the whole. However, a keyword search of the term "outlier" in Medline returned just 188 articles total. Most of the returned papers had to do with outlier lab results and quality control (Ahrens, 1999; Novis, Walsh, Dale, & Howanitz, 2004) rather than with extreme patients. Some discussed outlier physician performance.(Harley, Mohammed, Hussain, Yates, & Almasri, 2005)

Exactly one paper considered length of stay and the term, 'outlier'. It examined the length of stay of patients in the intensive care unit; all of the patients in ICU can be considered extreme or outlier.(Weissman, 1997) This paper clearly demonstrated the problem of assuming a normal distribution and estimating averages when the data were clearly skewed. The paper also showed that any current method used to define outliers tended to under-estimate the number of outliers, and the extremes in the added costs of outliers. This result was confirmed in a dissertation on costs and outliers.(Battioui, 2007) Therefore, future trends can only go in the direction of more concern for the impact of the outlier, or most severe patients.

REFERENCES

Ahrens, T. (1999). Outlier management. Influencing the highest resource-consuming areas in acute and critical care. *Critical Care Nursing Clinics of North America, 11*(1), 107-116.

Battioui, C. (2007). *Cost Models with Prominent Outliers.* University of Louisville, Louisville.

Fu, Q., Borneman, J., Ye, J., & Chrobak, M. (2005). *Improved probe selection for DNA arrays using nonparametric kernel density estimation.* Paper presented at the Annual International Conference of the IEEE Engineering in Medicine & Biology Society., Shanghai, China.

Harley, M., Mohammed, M. A., Hussain, S., Yates, J., & Almasri, A. (2005). Was Rodney Ledward a statistical outlier? Retrospective analysis using routine hospital data to identify gynaecologists' performance. *BMJ, 330*(7497), 929.

Hong, X., Chen, S., & Chris J Harris. (2008). A forward-constrained regression algorithm for sparse kernel density estimation. *IEEE Transactions on Neural Networks, 19*(1), 193-198.

Novis, D., Walsh, M. K., Dale, J. C., & Howanitz, P. J. (2004). Continuous monitoring of stat and routine outlier turnaround times: two College of American Pathologists Q-Tracks monitors in 291 hospitals. *Archives of Pathology & Laboratory Medicine, 128*(6), 621-626.

Pfeiffer, K. (1985). Stepwise variable selection and maximum likelihood estimation of smoothing factors of kernel functions for nonparametric discriminant functions evaluated by different criteria. *Computers & Biomedical Research, 18*(1), 46-61.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis (Monographs on Statistics and Applied Probability*. Boca Raton, FL: Chapman & Hall/CRC.

Weissman, C. (1997). Analyzing intensive care unit length of stay data: problems and possible solutions. *Critical Care Medicine, 25*(9), 1594-1600.

CONTACT INFORMATION

Patricia Cerrito
Department of Mathematics
University of Louisville
Louisville, KY 40292
502-852-6010
502-852-7132 (fax)
pcerrito@louisville.edu
http://stores.lulu.com/dataservicesonline