**Paper 210-31**

# Receiver Operating Characteristic (ROC) Curves

## Mithat Gönen, Memorial Sloan-Kettering Cancer Center

## ABSTRACT

Assessment of predictive accuracy is a critical aspect of evaluating and comparing models, algorithms or technologies that produce the predictions. In the field of medical diagnosis, receiver operating characteristic (ROC) curves have become the standard tool for this purpose and its use is becoming increasingly common in other fields such as finance, atmospheric science and machine learning. There are surprisingly few built-in options in SAS for ROC curves, but several procedures in SAS/STAT can be tailored with little effort to produce a wide variety of ROC analyses. This talk will focus on the use of SAS/STAT procedures FREQ, LOGISTIC, MIXED and NLMIXED to perform ROC analyses, including estimation of sensitivity and specificity, estimation of an ROC curve and computing the area under the ROC curve. In addition, several macros will be introduced to facilitate graphical presentation and complement existing statistical capabilities of SAS with regard to ROC curves. Real data from clinical applications will be used to demonstrate the methods.

## INTRODUCTION

Receiver operating characteristic (ROC) curves are useful for assessing the accuracy of predictions. Making predictions has become an essential part of every business enterprise and scientific field of inquiry. A simple example that has irreversibly penetrated daily life is the weather forecast. Almost all news sources, including daily newspapers, radio and television news, provide detailed weather forecasts. There is even a dedicated television channel for weather forecasts in United States. Of course, the influence of a weather forecast goes beyond a city dweller's decision as to pack an umbrella or not. Inclement weather has negative effects on many vital activities such as transportation, agriculture and construction. For this reason collecting data that helps forecast weather conditions and building statistical models to produce forecasts from these data have become major industries.

I will give examples from other areas where prediction plays a major role to motivate statisticians from diverse fields of application. Credit scoring is an excellent example: when a potential debtor asks for credit, creditors assess the likelihood of default to decide whether to loan the funds and at what interest rate. An accurate assessment of chance of default for a given debtor plays a crucial role for creditors to stay competitive. For this reason, the prediction models behind credit scoring systems remain proprietary but their predictive power needs to be continuously assessed for the creditors to remain competitive.

The final example is concerned with the field of medical diagnostics. The word "prediction" rarely appears in this literature, but a diagnosis is a prediction of what might be wrong with a patient producing the symptoms and the complaints. Most disease processes elicit a response that is manifested in the form of increased levels of a substance in the blood or urine. There might be other reasons for such elevated levels, and blood or urine levels mis-diagnose a condition because of this. The kind of analysis one would perform for weather forecasts is similarly valid for these blood or urine "markers."

ROC curves provide a comprehensive and visually attractive way to summarize the accuracy of predictions. They are widely applicable, regardless of the source of predictions. The field of ROC curves is by and large ignored during statistics education and training. Most statisticians learn of ROC curves on the jog, as needed, and struggle through some of the unusual features. To make matters worse for SAS users, very few direct methods are available for performing an ROC analysis although many procedures can be tailored with little attempt to produce ROC curves. There is also a macro available from the SAS Institute for this purpose. The goal of this paper is to summarize the available features in SAS for ROC curves and expand on using other procedures for further analyses.

## BASIC CONCEPTS: BINARY PREDICTOR

One of the simplest scenarios for prediction is the case of a binary predictor. It is important, not only pedagogically because it contains the most important building blocks of an ROC curve, but also practically because it is often encountered practice. I will use an example from weather forecasting to illustrate the concepts and at the end of the section mention some situations from other prominent fields.

The article by Thornes and Stephenson (2001) reviews the concepts of assessment of predictive accuracy from the perspective of weather forecast products. Their opening example is very simple and accessible to all data analysts regardless of their training in meteorological sciences. The example relates to frost forecasts produced for M62

motorway between Leeds and Hull in United Kingdom during the winter of 1995/1996. A frost is defined as when the road temperature falls below 0 °C. First, the forecast for each night is produced as a binary indicator Frost or No Frost. Then actual surface temperature for the road is monitored throughout the night and the outcome is recorded as Frost if the temperature dropped below 0 °C and as No Frost if it did not drop below 0 °C. The guidelines provided by the Highways Agency mandate the reporting of results (both forecast and the actual) in a consolidated manner (such as the following 2x2 contingency table) only for the days for which the actual temperature was below 5 °C. The example was concerned with the winter of 1995-1996 when there were 77 nights when the actual road surface temperature was below 5 °C. The results are given Table 1. Such a tabular description is the standard way of reporting accuracy when both the prediction and the outcome are binary. It is visually appealing, simple to navigate and it contains all the necessary information.

There were 29 nights when frost was forecast and a frost was observed and there were 38 nights when no frost was forecast and no frost were observed. Those two cells (the diagonal cells of the 2x2 table, the shaded portion of Table 1) represent the two types of correct forecast. A general terminology is true positives (TP) and true negatives (TN). The roots of this terminology can be found in medical diagnostic studies when a test is called positive if it shows disease and negative if it does not show disease. By analogy we can consider frost to mean "positive" and no frost to mean "negative" in which case we will have 29 true positives and 38 true negatives.

Table 1: Forecast accuracy for road surface temperature example

| Forecast | Observed | | |
|---|---|---|---|
| | Frost | No Frost | Total |
| Frost | 29 | 6 | 35 |
| No Frost | 4 | 38 | 42 |
| Total | 33 | 44 | 77 |

What about the forecast errors? There were 6 nights when a frost was forecast and none were observed. There were 4 nights when no frost was forecast but a frost was observed. You can easily extend the terminology to call these two cells as false positives (FP) and false negatives (FN). The following table is a generic representation of Table 1 using the terminology introduced above.

Table 2: Reporting accuracy for binary predictions

| Forecast | Observed | | |
|---|---|---|---|
| | Positive | Negative | Total |
| Positive | True Positive (TP) | False Positive (FP) | TP+FP |
| Negative | False Negative (FN) | True Negative (TN) | FN+TN |
| Total | TP+FN | FP+TN | TP+FP+FN+TN |

There are a variety of ways that one can summarize the forecast accuracy. One that comes to my mind first, perhaps, is the misclassification rate (MR) which is the proportion of all misclassified nights, the sum of false negative and false positives, out of all nights:

$$MR = \frac{FN + FP}{TP + FN + FP + TN}$$

One minus the misclassification rate is sometimes called "percent correct" or simply "accuracy." MR for the data in table 1 is 10/77=13%.

While misclassification rate is simple to compute and understand, it is sometimes too crude for understanding the mechanism behind misclassification. It is also prone to be biased if one is not careful in how the information is assembled. Suppose, instead of following the Highways Agency's guidelines, the forecast provider decides to include all nights in a calendar year. There are 77 nights reported in Table 1 and by definition all of those nights were those when the actual temperature dropped below 5 °C. Therefore the remaining 288 nights were all above 5 °C (no frost), bringing column marginal totals to 33 nights with frost (unchanged) and 332 nights with no frost. It is very possible that the MR for these 288 nights where the temperature was above 5 °C was much less than the MR for the nights where the temperature was below 5 °C. Suppose for the time being that the misclassification rate for these 288 nights was 5%, resulting in 15 misclassified nights (rounded up). Then there would be a total of 25 misclassified nights out of 365 and the MR will now be 25/265=7%. Table 3 shows several possibilities.

It is clear that MR is sensitive to which nights are included in the sample since the performance of the forecast is not homogeneous for all the nights in a year. It is also clear that as one includes more "easy-to-forecast" nights in the sample, the MR will be smaller. One can safely make a case that for warm days in spring and summer a forecast as to whether the M62 will freeze the following night is quite useless since most people can make the same prediction (quite accurately) without resorting to a scientifically obtained forecast. Therein lies the logic of the Highways Agency to restrict the reporting of accuracy only to those where the actual temperature was 5 °C or less.

Table 3: Putative MR for the 288 nights when the temperature was above 5 °C and the corresponding MR for all the 365 nights

| MR for the 288 nights when the temperature was above 5 °C | 0% | 1% | 2% | 4% | 6% | 8% | 10% | 12% | 13% |
|---|---|---|---|---|---|---|---|---|---|
| MR for all the 365 nights | 2.7% | 3.5% | 4.3% | 5.9% | 7.5% | 9.1% | 10.6% | 12.2% | 13.0% |

In unregulated areas such as credit scoring, where scoring algorithms remain mostly proprietary, there are no such rules, or even guidelines, as to how accuracy should be evaluated. In addition, consumers of predictions are not always diligent or knowledgeable in interpreting the details of forecast accuracy. Therefore it is desirable to propose measures that are more robust.

The most common way of reporting the accuracy of a binary prediction is using the true (or false) positives and true (or false) negatives separately. This recognizes that, a false negative prediction may have different consequences than false positive ones. In our weather forecast example, a false positive is probably less costly because its primary consequence may be more cautious and better-prepared drivers. On the other hand a false negative may end up in under-preparation and may lead to accidents. This suggests reporting false positive and false negative rates separately.

$$FPR = \frac{FP}{FP + TN}, \quad FNR = \frac{FN}{FN + TP}$$

One can extend these definitions to include TPR=1-FPR and TNR=1-FNR. The true positive rate (TRP) is sometimes called *sensitivity* and the true negative rate (TNR) is sometimes called *specificity*. While these are generic terms that routinely appear in the statistics literature, each field has come up with its own terminology. Weather forecasters, for example, use miss rate for FNR and false alarm rate for FPR.

The FPR and FNR for the weather forecast data are 4/33=12% and 6/44=14%. The sensitivity and specificity are 29/33=88% and 38/44=86%. In this instance FPR and FNR are both very close to each other. When this is the case, they will also be very close to MR. In fact one can show that MR is a weighted average of FPR and FNR: MR = w*FPR + (1-w)*FNR and the weight (w) is the proportion of nights with an observed frost. This is sometimes called the "prevalence" of frost.

Note that the denominators for FPR and FNR are the total observed positives and negatives. It is possible to define similar quantities using the forecast positives and negatives. In this case the ratios corresponding to sensitivity and specificity are called positive predictive value (PPV) and negative predictive value (NPV):

$$PPV = \frac{TP}{TP + FN}, \quad NPV = \frac{TN}{TN + FP}$$

It is very important to understand the correct interpretation of sensitivity, specificity, PPV and NPV. I will start by predictive values first. Their denominators are the number of positive and negative forecasts. In the weather forecast example PPV can be interpreted as the probability of an actual frost when a frost was forecast and NPV is the probability of observing no frost when no frost was forecast. In contrast, the denominators for sensitivity and specificity are observed positives and negatives. Therefore sensitivity is the probability that a night with a frost will be correct identified by the forecast out of all nights with a frost during the winter and similarly specificity is the probability that a night without a frost will be correct identified by the forecast out of all nights with no frost (and less than 5 °C) during the winter.

It is easy to imagine the utility of these probabilities as occasional consumers of a weather forecast. If I will drive from Leeds to Hull on M62 only a few nights during a winter, all I care is whether the forecast will be accurate on those few

nights. On the other hand the municipality will be more interested in sensitivity and specificity when deciding whether to pay for these forecasts since that speaks to the entire "cohort" of nights in the upcoming winter.

It is relatively easy to compute all of these measures in SAS®. The following DATA step prepares the data:

```
data m62;
   input forecast $ observed $ weight;
   datalines;
Frost Frost 29
Frost No_Frost 6
No_Frost Frost 4
No_Frost No_Frost 38
;
run;
```

and the following execution of PRC FREQ provides the necessary calculations:

```
proc freq data=m62;
   table forecast*observed;
   weight weight;
run;
```

The output is shown below and contains many useful pieces of information:

```
forecast      observed

Frequency|
Percent  |
Row Pct  |
Col Pct  | Frost   |No_Frost|  Total

Frost    |     29  |     6  |    35
         |  37.66  |  7.79  | 45.45
         |  82.86  | 17.14  |
         |  87.88  | 13.64  |

No_Frost |      4  |    38  |    42
         |   5.19  | 49.35  | 54.55
         |   9.52  | 90.48  |
         |  12.12  | 86.36  |

Total          33       44       77
            42.86    57.14   100.00
```

As a quick reminder to the PROC FREQ output, the key to the four numbers in each cell is found in the upper left portion of the table. The first number is the frequency, i.e. TP, FP, FN and TN. The second number uses the table sum (sum of column sums or sum of row sums) as the denominator. You can compute the misclassification rate using the numbers in the numbers in the second row. The third number is the row percentage, i.e. the proportion that use row sums as the denominator so these will include the PPV and NPV. Finally, the fourth number is the column frequency, using column sums as the denominator. Sensitivity, specificity, FPR and FNR will be among the third row numbers.

For example, the MR is the sum of the second numbers in the off-diagonals: 7.79+5.19=12.98%. PPV and NPV are 82.86% and 90.48%. The FPR and FNR are 13.64% and 12.12%.

It is easy to recognize that these are all binomial proportions if one considers the corresponding denominators as fixed. The technical term is conditioning on the denominators. This gives easy rise to the use of binomial functionality

within PROC FREQ to compute interval estimates.  The following example uses BY processing and the BINOMIAL option in the TABLE statement to produce confidence intervals and hypothesis tests for sensitivity and specificity.

```
proc sort data=m62;
   by observed;
run;

proc freq data=m62;
   by observed;
   table forecast / binomial;
   weight weight;
run;
```

The first part of the output for observed=Frost is given below. Here the denominator is observed positives so the report is focusing on sensitivity and FPR. The listing tells us that there were 29 true positives (both forecast and observed were "Frost") and 4 false positives (frost was forecast but not observed).

```
observed=Frost

The FREQ Procedure

                                  Cumulative     Cumulative
forecast     Frequency     Percent     Frequency     Percent
─────────────────────────────────────────────────────────────
Frost              29       87.88            29       87.88
No_Frost            4       12.12            33      100.00


      Binomial Proportion
      for forecast = Frost
─────────────────────────────────────
Proportion                0.8788
ASE                       0.0568
95% Lower Conf Limit      0.7674
95% Upper Conf Limit      0.9901

Exact Conf Limits
95% Lower Conf Limit      0.7180
95% Upper Conf Limit      0.9660

  Test of H0: Proportion = 0.5

ASE under H0              0.0870
Z                        4.3519
One-sided Pr >  Z        <.0001
Two-sided Pr > |Z|       <.0001
```

The output under "Binomial Proportion" contains the key information. Sensitivity was 87.88% and the asymptotic standard error (ASE) for the estimate of standard error is 5.68%. You should remind yourself at this point that the standard error for the estimate of a binomial proportion p is given by

$$ASE = \sqrt{\frac{p(1-p)}{n}}$$

where *n* is the denominator for the binomial proportion. The confidence limits appearing under the ASE are based on asymptotic theory. If *n* is large, then the 95% confidence interval can be calculated using

$$p \pm 1.96 \times ASE$$

Exact confidence limits, in contrast, are based on the binomial distribution and they have better coverage in small samples and/or rare events. Since they are now calculated by default, there is no need to use the asymptotic confidence limits.

Finally, the last part provides a test of whether the sensitivity is equal to 50% or not. In essence this is a test of whether there is any information for predicting frosts since 50% sensitivity can be obtained using a coin toss. The z-statistic reported is computed by z = p/ASE(null) where ASE(null) means the ASE under the null hypothesis. The difference from ASE above is in the use of observed p (estimate of sensitivity) versus the value under the null (0.5 in this case). This is a typical way of computing test statistics, sometimes referred to as Wald test. In large samples z has a normal distribution under the null hypothesis so a p-value can be obtained by referring to a standard normal table. This results in a p value less than 0.0001, a highly significant result suggesting that weather forecasts in this example are much better than tossing a coin.

The output below, which is the second part of the PROC FREQ output with BY statement, is for the FNR (not specificity). This is because the percentages computed by PROC FREQ use the first level in alphanumeric order which happens to be "Frost" in this case. When observed=Frost this matched with TP and the result was sensitivity but when observed=No_Frost it matches FNR.

```
observed=No_Frost

The FREQ Procedure

                                    Cumulative    Cumulative
forecast     Frequency     Percent   Frequency      Percent
───────────────────────────────────────────────────────────
Frost               6       13.64           6        13.64
No_Frost           38       86.36          44       100.00


       Binomial Proportion
       for forecast = Frost
─────────────────────────────────────
Proportion               0.1364
ASE                      0.0517
95% Lower Conf Limit     0.0350
95% Upper Conf Limit     0.2378

Exact Conf Limits
95% Lower Conf Limit     0.0517
95% Upper Conf Limit     0.2735

  Test of H0: Proportion = 0.5

ASE under H0             0.0754
Z                       -4.8242
One-sided Pr <  Z        <.0001
Two-sided Pr > |Z|       <.0001
```

To compute the specificity one only needs to subtract the estimates from 1. Therefore the specificity is 1=0.1364=86.36% and the confidence intervals are 1-0.2735=72.65% to 1-0.0517=94.83%. In terms of hypothesis testing the only change is in the sign of *z*. Specifically, ASE and the p-values will be the same.

While BY processing is simple and provides all answers, albeit after some minor hand calculation, it is possible to

completely automate the procedure. The following code calls PROC FREQ twice to accomplish this:

```
proc freq data=m62(where=(observed='Frost'));
  table forecast / binomial(p=0.8);
  weight weight;
run;

proc freq data=m62(where=(observed='No_Frost'));
  table forecast / binomial(p=0.75);
  weight weight;
run;
```

The difference from BY processing is the ability to set the null hypothesis through the use P= option in the TABLE statement. Since there are two separate calls to PROC FREQ, two different values of P can be used (0.8 and 0.75 in this example). The output is not repeated as it is identical in form to those above, with the only difference being the null value of the sensitivity and specificity.

### THE ROC CURVE

We are now ready to consider the problem that naturally leads to the use of the ROC curve. Suppose we are trying to predict a binary outcome, as in the weather forecast, but instead of a binary predictor we have a continuous predictor. Most weather forecasts are produced by statistical models which generate a probability level for the outcome. To simplify the use sometimes a pre-set threshold on this probability is used to dichotomize the forecast. Most weather forecasts mention "Rain" in their summary but if one looks at details they also report a "chance of rain." How would we assess the predictive accuracy of these probabilities (or, perhaps more appropriately, the model that produced these predicted probabilities)?

Since we know how to analyze predictive accuracy for a binary predictor (using MR, FPR etc) we may consider transforming the predicted probability into a dichotomy by using a threshold. The results, however, would clearly depend on the choice of threshold. How about using several thresholds and reporting the results for each one? ROC curve is one way of doing this. ROC curves focus only on sensitivity and (one minus) specificity. We will soon see the reason for this.

I was unable to find a reasonable data set on weather forecast to illustrate this point, so I will use an example from another field: medical diagnostics. One way to diagnose cancer is through the use of special scan called positron emission tomography (PET). PET produces a measure called standardized uptake value (SUV) which is an indicator of how likely the part of the body under consideration has cancer. SUV is a positive number. After SUV is measured the patients undergo a biopsy whereby a small piece of tissue from the suspected area is removed and examined under the microscope to evaluate whether it is cancer or not. This is called pathological verification and considered the gold standard.

These data are reported by Wong et al (2002). There are 181 patients 67 of whom are found to have cancer by the gold standard. Since we are dealing with a continuous predictor it is no longer possible to report the data in a tabular form. Instead I use the following side-by-side histograms from PROC UNIVARIATE that can be obtained with the simultaneous use of CLASS and HISTOGRAM statements.
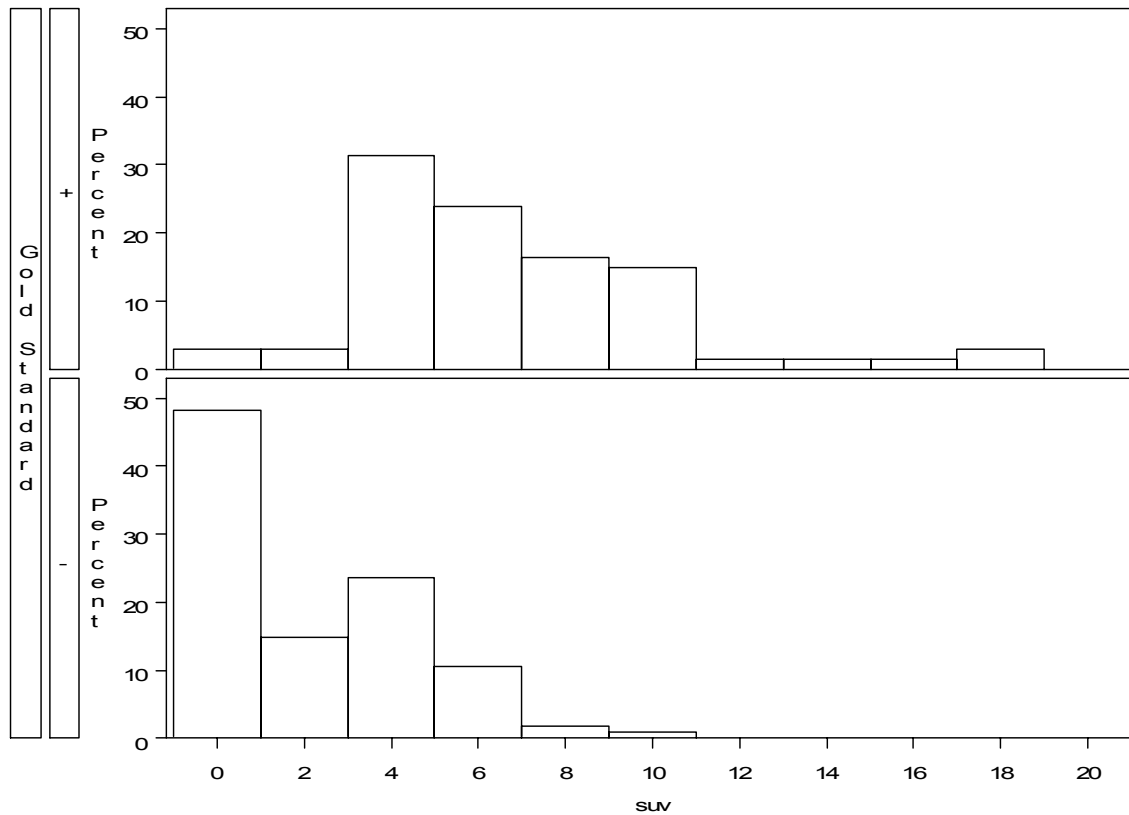
The upper panel of the figure is the histogram of SUV for the gold standard positive patients (those with cancer) and the lower panel is for gold standard negative patients (those without cancer), as denoted by the vertical axis. Patients with cancer tend to have higher SUV values; there are only a few that have very low SUVs. A very high SUV (say, 10 or more) almost certainly implies cancer. There is, however, some overlap between the distributions for SUVs in the middle range (roughly 4-10). So extreme values of SUV strongly imply cancer or lack of cancer, but there is a gray area in the middle. This is in fact a very common picture for many continuous predictors. How accurate, then, is SUV for diagnosing cancer?

One crude way of approaching the problem is to compute the sensitivity and specificity for various thresholds. The following graphical depiction is for using a threshold of 7:
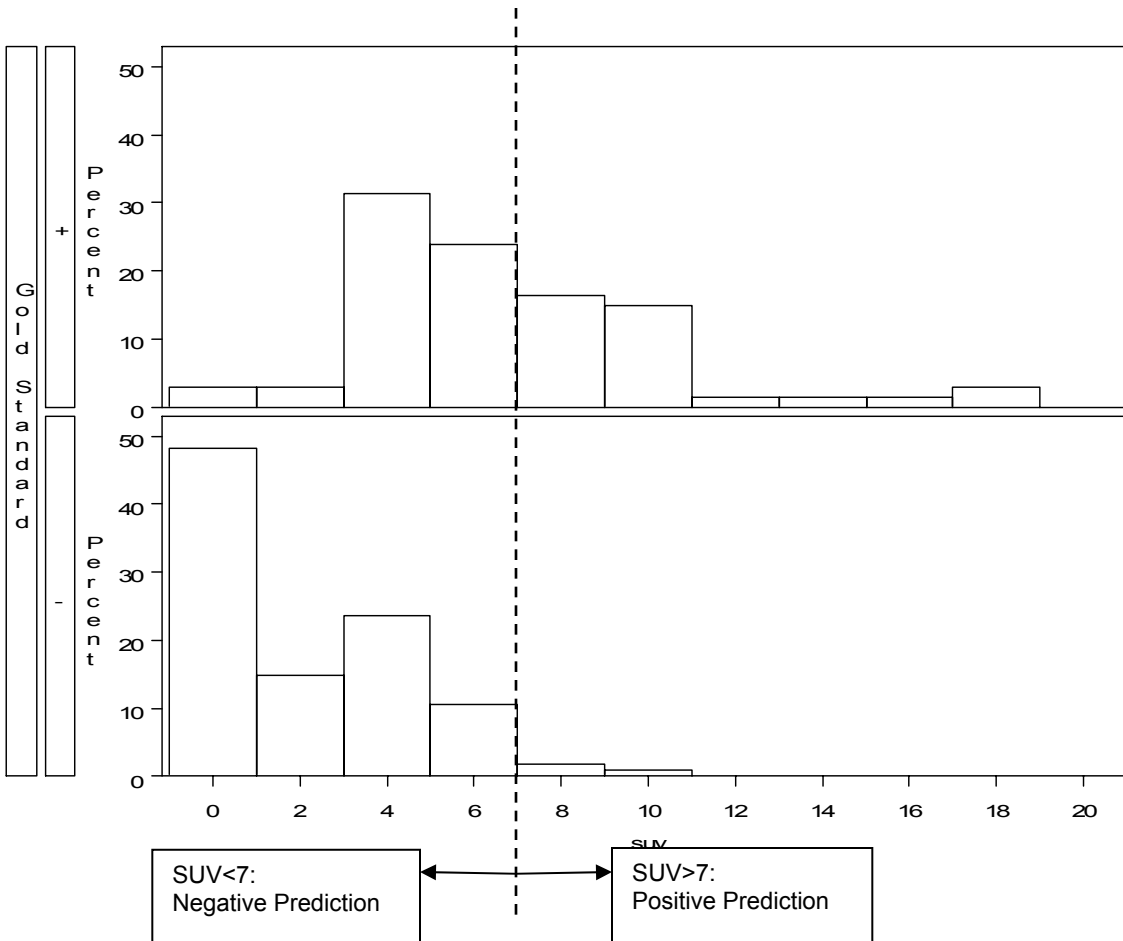
Dichotomization of SUV can be accomplished using the following data step command

```
suv7=(suv>7);
```

which can produce undesirable behavior if the variable SUV contains missing values in which case IF-THEN statements should be employed. Using this variable one can easily obtain the kind of 2x2 table considered in the previous section for the weather forecast example. For this data set it would look like the following:

Table 4: Accuracy of SUV>7 in diagnosing cancer

| SUV>7 Diagnosis | Gold Standard Diagnosis | | |
|---|---|---|---|
|  | Cancer | No Cancer | Total |
| Cancer | 25 | 3 | 28 |
| No Cancer | 42 | 111 | 153 |
| Total | 67 | 114 | 181 |

We can now estimate the sensitivity and specificity to be 25/67=37% and 111/114=97%. This analysis can be repeated for various thresholds. One way to report the results would be in tabular form, such as the following:

Table 4: Accuracy of SUV in diagnosing cancer for various thresholds

| Threshold | 1 | 3 | 5 | 7 |
|---|---|---|---|---|
| Sensitivity | 97% | 93% | 61% | 37% |
| Specificity | 48% | 65% | 88% | 97% |

This table makes clear the wide range of sensitivity and specificity that can be obtained by varying the threshold. It also makes clear the inverse relationship between the two measures: as one increases the other decreases and vice versa. This can be visually seen from the histogram above. If one moves the dashed line to the left (to, say 5, instead of 7) more patients will be classified as positive: some of these will be gold standard +, increasing the sensitivity, but some of them will be gold standard –, decreasing specificity. So the relationship we observed in Table 5 is universal: It is not possible to vary the threshold so that both specificity and sensitivity increase.

A tabular form is limited in the number of thresholds it can accommodate. One can plot the sensitivity versus specificity, keeping in mind that each point on the plot corresponds to a specific threshold. For reasons that will become clear later the traditional ROC curve plots sensitivity on the vertical axis and the specificity on the horizontal axis. Figure 1 is the ROC curve corresponding to this data set.
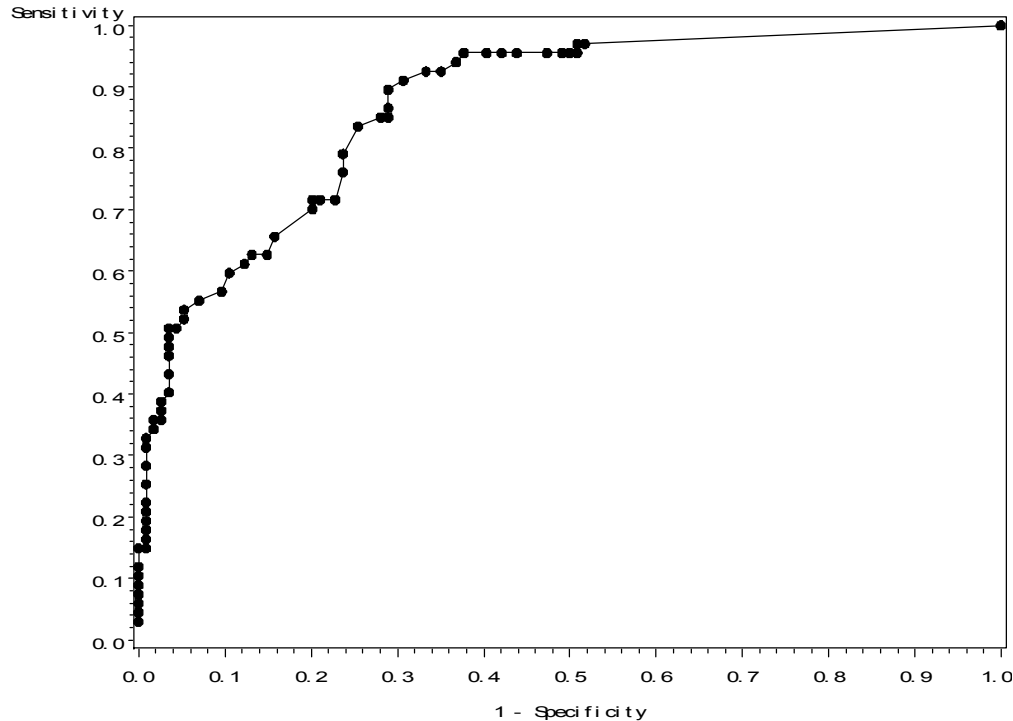


Each point on the ROC curve corresponds to a specific threshold. In this graph the points are connected, leading to this so-called empirical ROC curves. Each point corresponds to a threshold, although the value of thresholds is not evident from the graph. The ROC curve above can be plotted using the OUTROC option of PROC LOGISTIC:

```
proc logistic data=HeadNeck noprint;
  model gold_standard=suv / outroc=ROCData;
run;

symbol1 v=dot i=join;
proc gplot data=ROCData;
  plot _sensit_*_1mspec_;
run;quit;
```

It is often useful to enhance ROC curve plots in a variety of ways. The most common enhancement is the inclusion of a 45-degree line. This line represents the ROC curve for a useless test and a visual lower bound for the ROC curve of interest to out-perform, by including reference lines and highlighting selected thresholds. Another useful enhancement is the inclusion of a few selected thresholds on the curve, along with reference lines that indicate the corresponding sensitivity and specificity. Finally it is customary to have the horizontal and vertical axes have the same length, resulting in a square plot, as opposed to the default rectangular plot of SAS/GRAPH. The following plots the ROC curve of SUV again, with the mentioned enhancements. The macro %ROCPLOT, available from the SAS® Institute can be useful. The following was generated by %PLOTROC, available by e-mail from the author, was used to generate this plot.
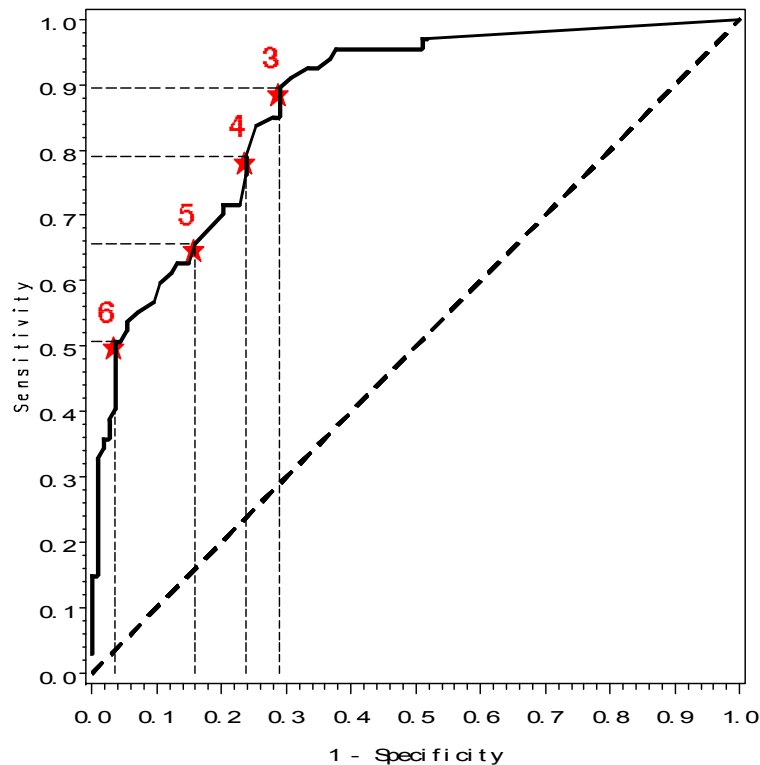
There is a convenient mathematical representation of the ROC curve that yields further insight to many of its properties. Consider the (complement of the) distribution of the SUV in patients with cancer, to be denoted by $SUV_+$: $P(SUV_+ > t)=F(t)$ and the (complement of the) distribution of the SUV in patients with no cancer, similarly denoted by $SUV_-$: $P(SUV_- > t)=G(t)$ where F(.) and G(.) are unspecified functions. Since all patients in F(.) are by definition gold standard positive, F(t) describes the proportion of positive patients whose SUV exceed y out of all positive patients. This is nothing but the sensitivity when t is used as the threshold. Similarly, 1-G(t) would be the specificity, hence G(t) represents one minus the specificity. Therefore F and G are the functions that we need to represent the ROC curve. Now, let x represent a point on the horizontal axis of the ROC curve (that is a fixed level of one minus the specificity). The threshold corresponding to this point will be given by $t=G^{-1}(x)$ and the sensitivity corresponding to t is F(t). So the sensitivity corresponding to x is given by $y=F(G^{-1}(x))$. This representation will be helpful in the sequel. It also explains why the ROC curves focus on sensitivity and specificity instead of NPV and PPV. The denominator of the latter two change with the threshold and do not lend themselves to notation and analyses by the use of cumulative distribution functions of the predictor variable.

## AREA UNDER THE ROC CURVE

While the ROC curve contains most of the information about the accuracy of a continuous predictor, it is sometimes desirable to produce quantitative summary measures of the ROC curve. The most commonly used such measure by far is the area under the ROC curve (AUC). In an empirical ROC curve this is usually estimated by the trapezoidal rule, that is by forming trapezoids using the observed points as corners, computing the areas of these trapezoids and then adding them up. This may be quite an effort for a curve like the one in Figure 1 with many possible thresholds. Fortunately, AUC is connected to a couple of well-known statistical measures that facilitates comparison and improves interpretation.

The first of these measures is the probability of concordance. If, in a randomly selected pair of patients, the one with the higher SUV has cancer and the one with the lower SUV has no cancer than this pair is said to be a concordant pair. A pair where the patient with the higher SUV has no cancer but the one with the lower SUV has cancer is said to be a discordant pair. Some pairs will have ties for the SUV. Finally some pairs will be non-informative, for example both patients may have cancer or both may have no cancer. It is not possible to classify these pairs as concordant or discordant. The probability of concordance is defined as the number of concordant pairs plus one-half the number of

tied pairs divided by the number of all informative pairs (i.e. excluding non-informative pairs). In essence, each tied pair is counted as one-half discordant and one-half concordant. This is a fundamental concept in many rank procedures, sometimes referred to as "randomly breaking the ties." An equivalent way to express concordance is $P(SUV_+ > SUV_-)$ where $SUV_-$ indicates the SUV of a gold standard negative patient and $SUV_+$ indicates the SUV of a gold standard positive patient. It turns out that AUC is equal to the probability of concordance.

AUC is also closely related to the Wilcoxon rank-sum test. The rank-sum statistics is the sum of the ranks in one of the groups (here group is defined by the gold standard). If you take the sum of the rank is the gold standard positive group, then the number of concordant pairs is equal to the rank sum minus a constant that depends on the size of the two groups only.

All of these relationships highlight a fundamental property of the ROC curve that it really is a rank-based measure. In fact the empirical ROC curve is invariant to monotone transformations of the predictor. So one would get the same ROC curve if log(SUV) or $SUV^{1/3}$ is used as the predictor.

PROC LOGISTIC reports the area under the curve in its standard output under the heading "Association of Predicted Probabilities and Observed Responses." This part of the output for the cancer example looks like the following:

```
Association of Predicted Probabilities and Observed Responses

Percent Concordant      86.1     Somers' D      0.743
Percent Discordant      11.8     Gamma          0.758
Percent Tied             2.1     Tau-a          0.348
Pairs                   7638     c              0.871
```

We see that out of the 7638 (informative) pairs, 86.1% were concordant and 2.1% were tied. Proportion of concordant pairs plus one-half the proportion of concordant pairs is 87.1% and is separately reported as "c" for concordance. Also reported is Somers' D statistic which is also related to concordance via $D = 2*(c-0.5)$, Somers' D is simply a rescaled version of concordance that takes values between -1 and 1, like a usual correlation coefficient instead of 0 and 1.

AUC is estimated from the data, so there has to be a standard error associated with this estimation. Unfortunately it is not available in PROC LOGISTIC. For this we turn to a macro provided by the SAS® institute. Macro %ROC, currently in version 1.5 is available at http://support.sas.com/samples_app/00/sample00520_6_roc.sas.txt. The following is the call that can be used for this example:

```
%roc(version,data=HeadNeck,var=SUV,response=gold,contrast=1,details=no,alpha=.05);
```

First one needs to create a gold standard variable takes on values of 0 or 1 only, the macro does not have the capability of re-coding the observed binary classes. The variable GOLD takes on the value 1 for gold standard positive patients and 0 for negative patients. VAR specifies the continuous predictor. CONTRAST will always be 1 for a single ROC curve, it is useful for comparing several curves. Alpha pertains to confidence limit estimation and details control the amount of output printed. The output from this call is given below. The area under the curve is 87.13% (same as the concordance from PROC LOGISTC, as it should be). But this output has the standard error for this estimate (2.63%) and the associated asymptotic confidence intervals (81.98%-92.28%). This is a more complete analysis, as it is possible to judge the effects of sampling variability on the estimate of AUC.

```
The ROC Macro

          ROC Curve Areas and 95%
           Confidence Intervals
      ROC Area Std Error Confidence Limits


SUV      0.8713    0.0263      0.8198       0.9228
```

## THE BINORMAL ROC CURVE

So far our focus was on the empirical curve, the one that is obtained by connecting the observed (sensitivity, 1-specificty) pairs. The empirical curve is heavily-used because it makes no assumptions regarding the distribution of the individual predictors. But there are occasions where one would prefer a smooth curve. The most common way of smoothing an ROC curve is using the binormal model.

The binormal model assumes that SUVs of the patients with cancer follow a normal distribution with mean $\mu_1$ and variance $\sigma_1^2$ and  SUVs of the patients with no cancer follow a normal distribution with mean $\mu_0$ and variance $\sigma_0^2$. Then using the notation from the empirical ROC curves, $G(t)=\Phi((\mu_0-t)/\sigma_0)$. It follows that the threshold t can be written as a function of x as follows: $t = \mu_0 - \sigma_0 \Phi^{-1}(x)$. Since a threshold t corresponds to the sensitivity F9T0 we can write the functional form of the ROC curve as

$$F(t) = \Phi\left(\frac{\mu_1 - t}{\sigma_1}\right) = \Phi\left(\frac{\mu_1 - \mu_0 + \sigma_0 \Phi^{-1}(x)}{\sigma_1}\right) = \Phi\left(a + b\Phi^{-1}(x)\right)$$

where

$$a = \frac{\mu_1 - \mu_0}{\sigma_1}, \quad b = \frac{\sigma_0}{\sigma_1}$$

and are often referred to as "binormal parameters." Sometimes they are called intercept and the slope since plotting on normal probability paper would yield a straight line for the binormal curve with intercept a and slope b. This practice is not common anymore, but the nomenclature continues.

The area under the curve for the binormal model also has a closed-form expression:

$$AUC = \Phi\left(\frac{a}{\sqrt{1 + b^2}}\right)$$

To estimate a binormal model from data one needs to estimate the means and variances separately. This can simply be accomplished using PROC MEANS with a CLASS statement. It can also be obtained using PROC NLMIXED as follows:

```
proc nlmixed data=hn(drop=m);
  parameters m1=0 m0=0 s1=1 s0=1;
  if gold=1 then m=m1;else if gold=0 then m=m0;
  if gold=1 then s=s1**2;else if gold=0 then s=s0**2;
  a=(m1-m0)/s1;
  b=s0/s1;
  model highs ~ normal(m,s);
  estimate 'a' a;
  estimate 'b' b;
  estimate 'AUC' probnorm(a/sqrt(1+b**2)));
run;
```

The advantage of using NLMIXED is that, via the ESTIMATE statement, one can directly obtain estimates, and perhaps more importantly, standard errors for a, b and AUC. The following output from NLMIXED execution, under the title of "Additional Estimates," is all that  is needed:

```
                              Additional Estimates


                      Standard
    Label    Estimate     Error     DF    t Value   Pr > |t|   Alpha     Lower     Upper

    a          1.3096    0.1776    181       7.37     <.0001    0.05     0.9591    1.6600
    b          0.6604   0.07189    181       9.19     <.0001    0.05     0.5185    0.8022
    AUC        0.8628   0.02935    181      29.40     <.0001    0.05     0.8048    0.9207
```

Using this output the binormal ROC curve is given by $\Phi(1.31 + 0.66\ \Phi^{-1}(x))$ where x takes on values from 0 to 1. That the ROC curve is defined for all points between 0 and 1 is a feature of the model. Although not all thresholds are observed, the binormal model "interpolates" and makes available a sensitivity for all possible values of one minus specificity.

The following is a graphical comparison of the empirical curve (solid line) and the binormal curve (dotted line) using cancer example. Early on binormal ROC curve is above the empirical one, suggesting that sensitivity is overestimated when compared with the observed values. Later, curves cross and binormal curve now underestimates sensitivity. Most of the time the difference between the two curves is less than 10%. This would be considered a reasonable fitting binormal ROC curve.
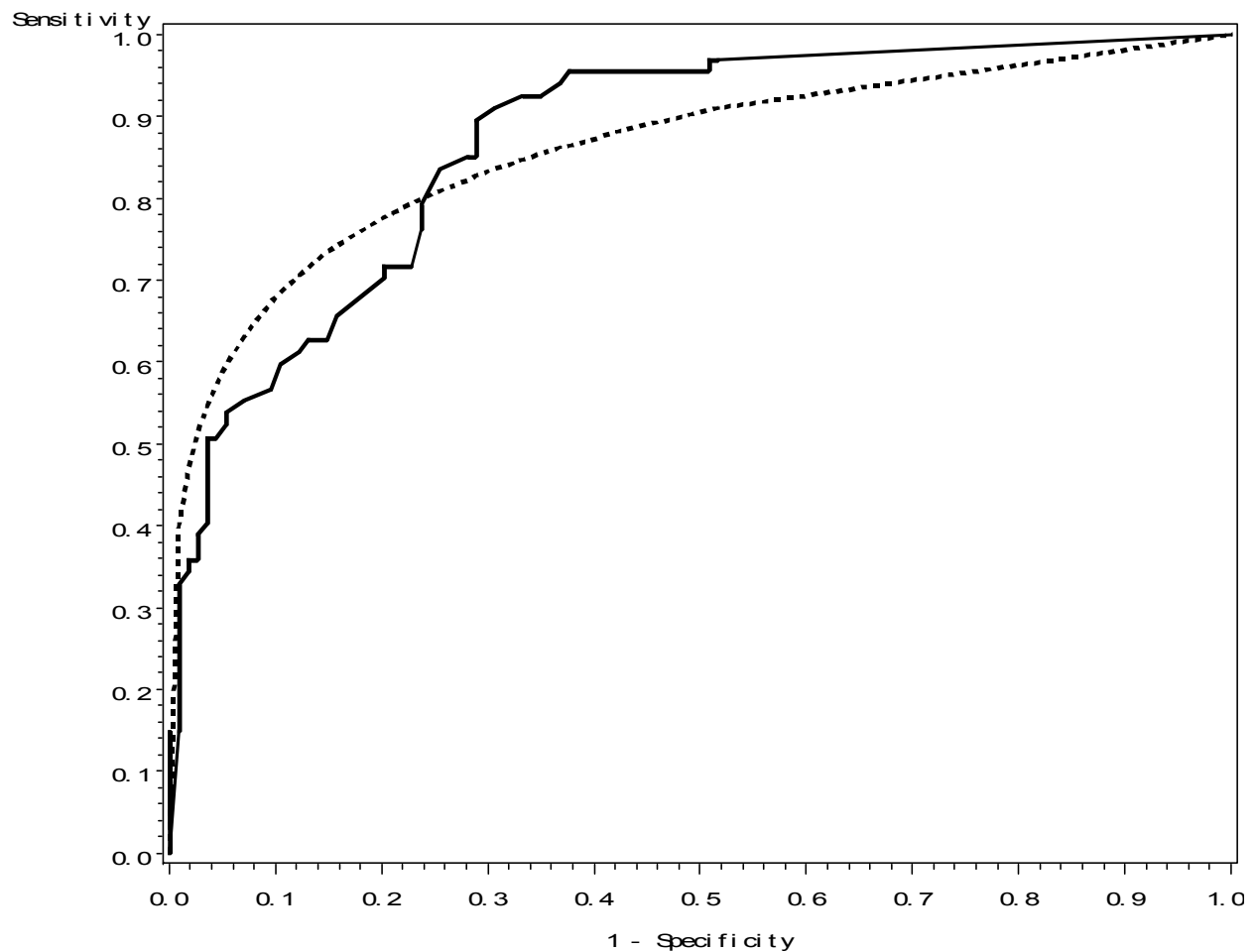
The area under the empirical ROC curve was 87%, slightly higher than 86% which is the area under the binormal ROC curve. The confidence limits are also very similar: (82%, 92%) versus (81%, 92%). By most accounts the binormal analysis yields similar results to that of the empirical curve.

## COMPARING TWO ROC CURVES

It is often of interest to compare predictions. How does an easy- and inexpensive-to-obtain medical diagnosis differ from one that requires an invasive procedure or a radiologic scan? One can construct ROC curves for each and compare them. Visual comparisons usually reveal several useful features, but statistical significance between two ROC curves can also be obtained.

Continuing with the cancer example, all of the patients who had their SUV measured also have additional information available: size of the lesion under consideration (called tumor or T stage, an ordinal variable with four categories, 1 through 4), whether lymph nodes are involved or not (called nodal or N stage, an ordinal variable with three categories, 0 through 2) and a binary variable which is the overall assessment of the physician which is affected by SUV, T stage and N stage, but other, perhaps unquantifiable factors.

When multiple variables need to be combined to make a prediction, a reasonable way is to build a logistic regression model. We have simply used the main effects in the following code, more careful modeling might produce a better prediction model:

```
proc logistic data=hn;
   class tstage nstage overallint;
   model gold_standard=suv tstage nstage overallint;
   score out=predprobs outroc=combinedroc;
run;
```

The output from the LOGISTIC procedure indicates that the area under the curve for this combined prediction rule is 90.2%, a slight increase from 87.1% with SUV alone. Without access to standard errors it is not possible to judge whether this increase can be explained by sampling error or not. In addition this is a paired-data situation, since every patient receives two predictions, one with SUV only, and the other with the combined rule. Therefore the two ROC curves are not independent and any statistical procedure evaluating their significance needs to assess the covariance between the two curves.

%ROC macro that we used in above to obtain estimates of standard error can also be used to compare two ROC curves. First one needs to merge the PREDPROBS data set created by the execution of PROC LOGISTIC above with the original data set that contains SUV. This data set is called MERGED here and P_ is the variable that contains the predictions from the logistic model. CONTRAST macro variable expects input similar to the CONTRAST statement in PROC GLM, so 1 -1, as below would refer to the difference of the two variables stated in the VAR macro variable.

```
%roc(version, data = merged,var = p_ suv, response = gold, contrast = 1 -1);
```

The output from this call is given below. We first see the area under the curve estimates for each of the ROC curves, along with their standard errors and (marginal) confidence intervals. Then the contrast information is printed. The next piece of output is the one that compares the two ROC curves. The two AUCs differ by about 2% with a standard error of 1.3%. The confidence interval for this difference is -0.4% to 4.6%, covering 0 and indicating that the two predictions are not statistically distinguishable. The p-value for the corresponding significance test is 0.1021.

```
The ROC Macro


          ROC Curve Areas and 95%
            Confidence Intervals
       ROC Area Std Error Confidence Limits


P__      0.9023    0.0267    0.8499     0.9547
highs    0.8815    0.0293    0.8240     0.9389



 Contrast Coefficients
              P__       highs


Row1            1         -1



       Tests and 95% Confidence Intervals for Contrast Rows
      Estimate  Std Error Confidence Limits    Chi-square Pr > ChiSq


Row1  0.0208     0.0127    -0.0041     0.0458      2.6732  0.1021



      Contrast Test Results
Chi-Square    DF       Pr > ChiSq


    2.6732    1         0.1021
```

Same comparisons can be performed within the context of a binormal model. It is easier to write the binormal model as a regression:

$$E(W) = \alpha_0 + \alpha_1 G + \alpha_2 P + \alpha_3 GP + U + \exp(\beta G)\varepsilon$$

$$U \sim N(0, \tau^2)$$

$$\varepsilon \sim N(0, \sigma^2)$$

In this model G=0,1 represents the gold standard and P==0,1 is the type of predictor (SUV vs combination). U is a random effect modeling the covariance between SUV and combination results due to the fact that they were obtained on the same patients.

$$\sigma_0^2 = \sigma^2$$

$$\sigma_1^2 = \exp(2\beta)\sigma^2$$

The ROC curves induced by this model for each type are binormal with the same slope, b = $\sigma_0^2$ / $\sigma_1^2$. The intercept is different and given by a$_0$ = $\alpha_1$ / $\sigma_1^2$ and a$_1$ = ($\alpha_1$ + $\alpha_3$) / $\sigma_1^2$. Therefore it is $\alpha_3$ that represents the difference between the two ROC curves.

This model is fit using PROC NLMIXED with the following statements. The data set required for PROC NLMIXED contains two rows per patient, one for P=0 and one for P=1. It needs to have a subject indicator, SUBID in this data set. The binormal parameters and the AUCs are computed along with their standard errors using the ESTIMATE statements.

```
proc nlmixed data=long;
  parameters s1=1 s0=1 sr=1 _g=0 _gt=0 _a=0;
  if gold=1 then s=s1**2;else if gold=0 then s=s0**2;
  mm=_a+_g*gold+_t*type+_gt*gold*type+u;
  a1=_g/s1;a2=(_g+_gt)/s1;b=s0/s1;
  auc1=probnorm(a1/sqrt(1+b**2));auc2=probnorm(a2/sqrt(1+b**2));
  model result ~ normal(mm,s);
  random u ~ normal(0,sr) subject=subid;
  estimate 'a1' a1;
  estimate 'a2' a2;
  estimate 'b' b;
  estimate 'AUC1' auc1;
  estimate 'AUC2' auc2;
  estimate 'AUC1-AUC2' auc1-auc2;
run;
```

The relevant output from NLMIXED is given below:

```
                           Additional Estimates

                    Standard
Label      Estimate     Error    DF   t Value   Pr > |t|   Alpha    Lower     Upper

a1           1.8494    0.2396    142     7.72    <.0001     0.05    1.3758    2.3229
a2           1.3307    0.2188    142     6.08    <.0001     0.05    0.8981    1.7633
b            0.5677   0.06834    142     8.31    <.0001     0.05    0.4326    0.7028
AUC1         0.9461   0.02014    142    46.97    <.0001     0.05    0.9063    0.9859
AUC2         0.8764   0.03620    142    24.21    <.0001     0.05    0.8048    0.9480
AUC1-AUC2   0.06970   0.03163    142     2.20    0.0292     0.05   0.007174   0.1322
```

The AUCs are estimated to be 94.6% and 87.6%. Note that the estimate for the AUC of SUV is not identical to the one obtained from the binormal model that use only SUV. The reason for that has to do with the variance parameters. In particular the slope now is common to both predictors and different from the slope in the previous model. The difference between the two AUCs is given to be 7% and the p-value for this difference is 0.0292, suggesting a significant improvement over SUV alone by the combination score.

It is not uncommon that the binormal model and the empirical model results in different conclusions. The difference between them is similar to the difference between rank tests and t-tests. If the assumptions underlying the binormal model hold then the birnomal model has more power and this might explain the significant result. On the other hand if the model assumptions do not hold then the Type I error may be inflated which would manifest itself in the form increased false positive results. It is really not possible to conclude which one is the driving force here. Exploratory graphical analyses, like the one performed above remain the best way of checking the assumptions of binormality.

More than two predictors can be compared using dummy variables in the model. One can also adjust for covariates as needed. In summary binormal regression models are very powerful and flexible but they may be sensitive to departures from model assumptions.

We caution that despite its enormous popularity, the binormal models make some strong assumptions. It is no longer rank-based, every single observation contributes to the estimation of the binormal slope and intercept.

## CONCLUSION
Predictive models are used by virtually every field of science and almost all business enterprises. ROC curves provide a way of assessing the accuracy of these predictions. Predictions are increasingly being produced by statistical models implemented by computers and most of them generate predicted probabilities (from a logistic regression model, for example).  ROC curves are most useful when predictions are continuous, as they would be in predicted probabilities. The methods we have considered in this article can be applied to any continuous prediction, regardless of whether they are directly observed variables (such as SUV) or model-generated abstractions (such as a credit score or a predicted probability).

ROC curves can be generalized to ordinal predictions as well. These are usually subjective assessments of experts, or an empirical combination of various variables (such as cancer staging). The empirical ROC curve can be generated using the principles outlined above and the binormal can be extended using latent variables to obtain a smooth curve. PROC LOGISTIC and PROC NLMIXED would be the engines of these analyses and code is available from the author for these analyses. Comparing two ordinal predictors or adjusting for covariates can all be done within the regression framework that is easily implemented in PROC NLMIXED.

Another data structure that has received interest in the context of ROC curves is clustered data. This corresponds to multiple observations for each patient, encountered in medical diagnostic contexts. For example, a PET scan of the abdominal region may be evaluated for stomach, colon and liver separately yielding clustered data. One can implement a hierarchical model in NLMIXED, although since NLMIXED is limited to one RANDOM statement it is not possible to fit a model to this data set.

It is important to remember that graphical methods, such as side-by-side histograms remain the best way of communicating your results as well as checking your assumptions.

## REFERENCES

Thornes JE and Stephenson DB (2001). How to judge the quality and the value of weather forecast products. *Meteorological Applications*, 8:307–314.

Wong RJ et al. (2002). Diagnostic and prognostic value of 18F-fluorodeoxyglucose positron emission tomography for recurrent head and neck squamous cell carcinoma. *Jornal of Clinical Oncology*, 20:4199-4208.

## RECOMMENDED READING

There are two books that cover ROC curves extensively from a statistical perspective: "The Statistical Evaluation of Medical Tests for Classification and Prediction" published by Oxford Press, written by MS Pepe and "Statistical Methods in Diagnostic Medicine" published by Wiley, written by XH Zhou et al. In addition both SAS Press and Springer are in preparation to publish books on ROC curves within the next year.

## CONTACT INFORMATION

Your code requests, comments and questions are valued and encouraged.  Contact the author at

Mithat Gönen
Memorial Sloan-Kettering Cancer Center
1275 York Avenue Box 44
New York, NY 10021
(646) 735-8111
gonenm@mskcc.org
http://www.mskcc.org/mskcc/html/3185.cfm