Logistic Regression In Cartoons

Russ Lavery, Contractor, Bryn Mawr, PA

ABSTRACT

This paper contains discussions of issues associated with logistic regression where the issues are easier to understand when presented graphically than when they are presented using formulas. The topics discussed are: odds are not probabilities or percents; why use 1 / (1 + exp(-x)); complete and partial separation; why do we use the logit (What do the estimates in SAS printouts mean); the effect of beta values on the percent curve; oversampling – why and how; the ROC curve; the C statistic.

INTRODUCTION

The approach of this paper is to use graphics and to "wander around" some of the issues in logistic regression – and it has reasons for this approach. Firstly, people learn in different ways and sometimes a graphical presentation of material offers insights that are difficult to get from formulas. Secondly, a way to *really* understand an issue is to "walk around and look at the problem from a different angle". More formally, we call that "doing examples" or "what if" explorations. Sometimes, these attempts to "play" with the topic are useful and I hope these attempts to see logistic from a new viewpoint are of use to some readers.

1 / 8) ODDS ARE NOT PROBABILITIES OR PERCENTS

Occasionally, people who are not in the habit of thinking of odds, get confused between odds and percentages. The two statistics look similar, especially when the odds being quoted is between zero and one. Calculating odds and percentages both involve using formulas containing a fraction (a numerator over a denominator) but the denominators are different.

As examples: think of successes (S) and failures (F).
To calculate the percentage of failures the formula is: count of F / total count OR count of F / (count of F + count of S)

To calculate the odds of failures the formula is: count of F / count of S)

If the set of three events is: F S S then:
The percentage of Fs in the set is 1/3 or .33
The percentage of Ss in the set is 2/3 or .66 **(this number being between zero and one is what leads to confusion)
The Odds of Fs in the set is 1/2 or .50 **(this number being between zero and one is what leads to confusion)
The Odds of Ss in the set is 2/1 or 2

2 / 8) WHY DO WE USE THE TRANSFORM 1 / (1 + EXP(-X))?

An S-shaped curve describes many biological and physiological processes. As an example, imagine yourself as a healthcare worker conducting a hearing test where the X axis, in Figure 1, is sound volume. Y is the percentage of people, at that level of sound volume, who press a button indicating they have heard the tone.

At some very, very low volume level nobody hears the tone. Some people have very good ears and hear the tone at very low volume levels. As the tone volume increases, more and more people are able to hear the tone. There typically is a "middle range" where the majority of people start to hear the tone.

Finally, some people are very hard of hearing and need a very loud tone before are able to hear it.

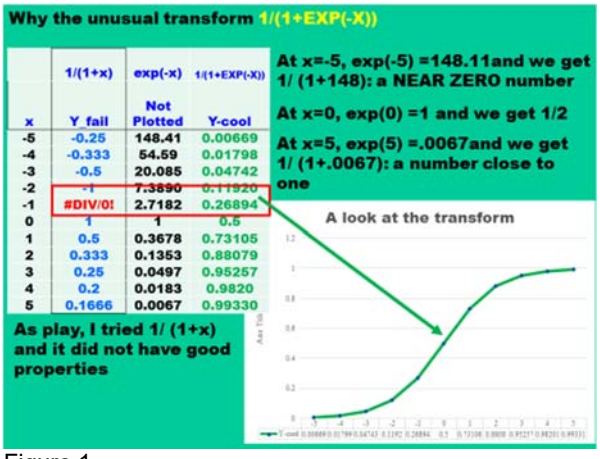


Figure 1

An S-shaped curve is very useful in describing "response to stimuli processes" like these. The formula 1 / (1 + exp(-x)) has the ability to "map"‘ values of X into a S shaped response curve. At x= -5, exp(-5) =148.11 and we get Y= 1/ (1+148): a NEAR ZERO number for large negative values of X. At x=0, exp(0) =1 and we get Y= .5. At x=5, exp(5) =.0067and we get Y= 1/ (1+.0067): a number close to one for large positive values of X. The formula maps Xs to Y values that asymptotically approach zero and one.

3 / 8) COMPLETE AND PARTIAL SEPARATION

Complete separation is a problem in logistic regression because it makes the algorithm used by PROC Logistic fail to converge. It occurs when one X variable, or a linear combination of X variables, can perfectly predict the classification.

If the groups are 100% separable the problem is in the calculation of the beta values associated with the logistic regression, because the values of the beta approach infinity.

If a variable, or a group of variables, can predict with 100% accuracy the effect of the variable (or the group of variables) is infinite.

You can see the problem in Figure 2. All of the green triangles are on one side of the line and all of the blue circles are on the other side of the line. This data set has the problem of "Complete Separation".

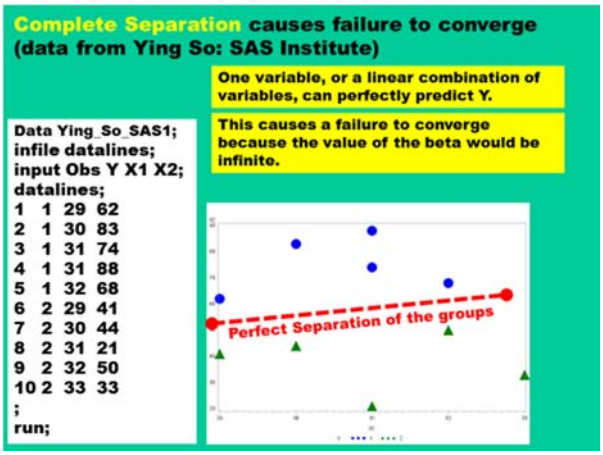


Figure 2

Quasi-complete separation is a similar problem. When a data set suffers from quasi-complete separation, the line "separating" the two groups has observations from both groups ON the separating line. This can be seen in Figure 3.

We need to have some miss-classifications for the algorithms to run to completion. Note that if the green triangle, on the line in Figure 3 were to move "up" just a bit our problem would be solved.

The Firth option instructs SAS to "jiggle" the points a bit – in a random manner. If one were to apply the Firth option to the data set in Figure 3, there is hope that the points on the line that separates the groups would move enough to create some misclassifications and allow the algorithm to converge.

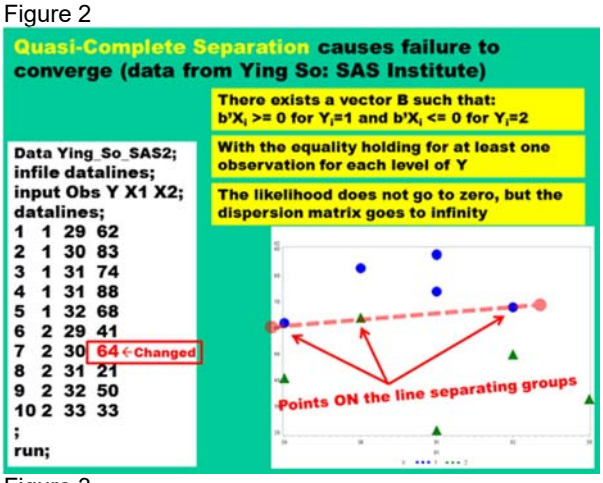


Figure 3

4 / 8) WHY THE LOGIT

Figures 5 is large because it is very important. The pictures on Figure 5 were, for me, worth 1000 formulas.

It would be easier to follow this example if I provided context. In Figure 5, the top chart is a typical S-shaped probability curve. The Y value is a percent and the X value represents levels of some X variable. Imagine our issue is that were trying to kill ants with pesticides. We put 100 ants in each of eleven test tubes and put pesticides, at different levels into each of the test tubes. The pesticide levels (our X values) in test tubes are .1 (zero was not used) 1 , 2, 3 on through 10 ppm. The Y values are really zero or one (an ant can only be alive or dead) but if we take the percentage of ants, in a test tube, who died at a pesticide level we get a number between zero and one. Some ants (probably old or sick) died at very low pesticide levels. Some ants were very resistant to the pesticides and can survive very high levels. This biological process creates an S shaped curve.

The middle chart is the odds chart. There are many formulas for odds, but a commonly used formula is: the probability of the event divided by the probability of the nonevent (probability of death divided by probability of surviving). This odds curve is no longer S-shaped. If the probability curve is S shaped and slopes up for large X, the odds curve goes to infinity for large X.

The bottom chart is the Logit chart. It is created by taking the log of the odds. The three charts in this paper were created in Excel. I manually "eyeballed"/created an S-shaped probability curve. I then used Excel to calculate the odds at each level of X and created the middle chart. I then used Excel to calculate the log of the odds and used those numbers to create the bottom chart – the logit chart. Here is the "ah-hah" for this example. ***The logit chart is a straight line and can meet the assumptions of ordinary least squares.***

Importantly, the beta values, or the estimates column in SAS output (shown in Figure 4 below), from a logistic regression predict the straight line at the bottom of Figure 5 – ***so, the beta values in the SAS printout are used to predict the logit line.***

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Std. Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.5707	0.3513	251.4826	<.0001
dose	1	1.1141	0.0670	276.5385	<.0001

Figure 4

Please look at the formulas in the yellow boxes in Figure 5 and notice the washed out colors associated with B_2 and X_2, The pictures shown in Figure 5 are for a problem with only one X variable but the result (logit line satisfies the assumption of OLS) is generalizable to problems where there are more than one X variable. If you had more than one X variable the formula would be predicting a logit hyper plane rather than a logit line. In **LN(odds of Y) = B_0 + B_1x_1 + B_2 x_2**, the B_2 and X_2, are "grayed out" because I wanted to show that this formula, this insight, would work if you had a logistic regression with more than one X variable. I used a gray color for **LN(odds of Y) = B_0 + B_1x_1 + B_2 x_2** because the pictures on the left side of the slide do not "need" an X2 variable.

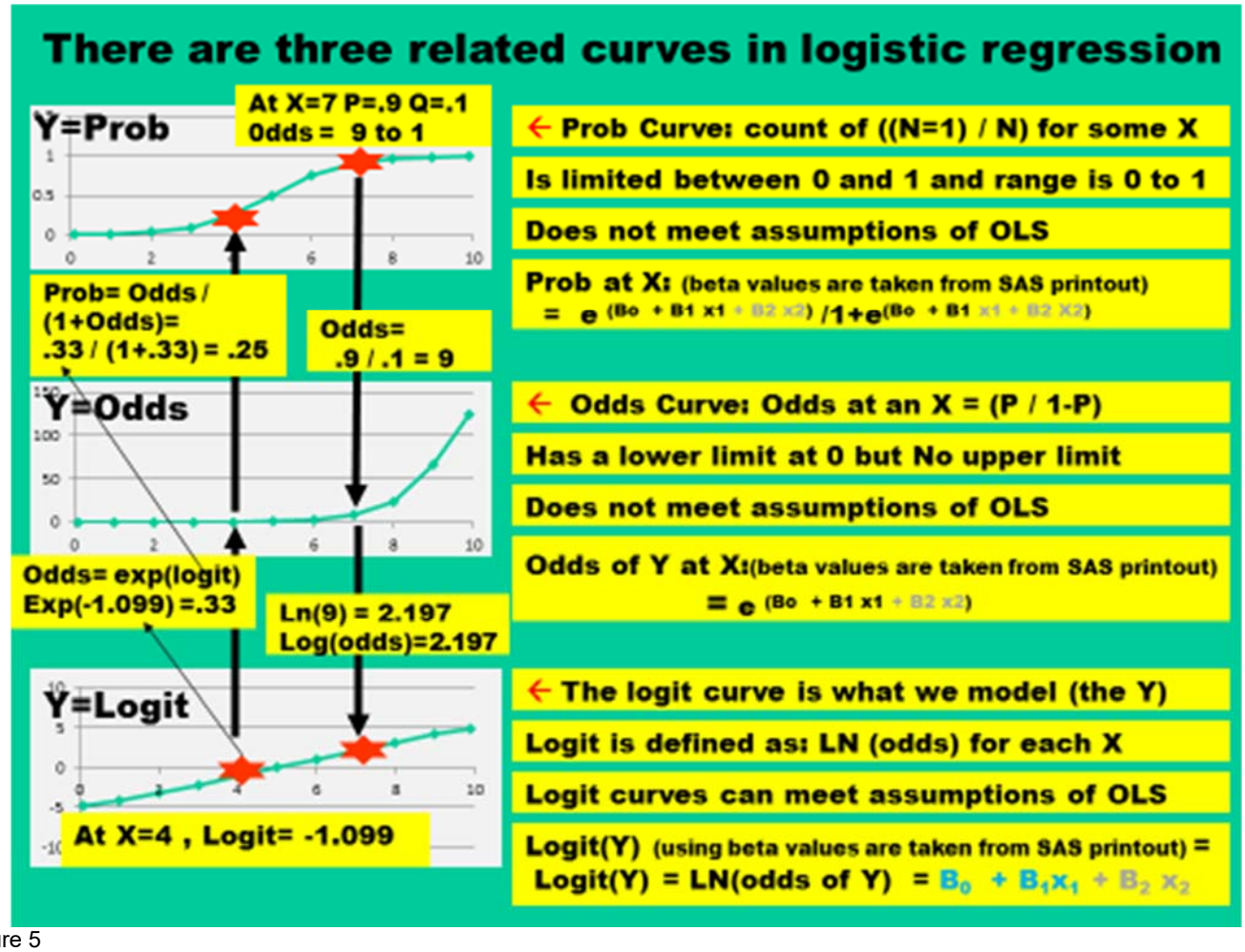


Figure 5

A WORKED EXAMPLE – MOVING BETWEEN THE THREE CURVES IN FIGURE 5

I'd like to work an example that is shown in Figure 5. At 7 ppm 90% of the ants died. To get from the probability curve to the odds curve we calculate the odds at 7 ppm. The odds are .9 /.1 or 9 to 1. The odds curve scaling does not make it easy to see, but the odds at X =7 are 9. To get to the logit curve one takes the natural log of the odds. Ln(9)=2.197 and that is a point on the logit (bottom) curve.

We can use some simple formulas to work the problem in reverse (from logit "up" to %). Slide along the logit line to where X =4. At X=4 the logit is -1.099 (I did this calculation using the Excel spreadsheet and if you use the estimates from the SAS printout you will get -1.05 - call that a rounding error). To get from the logit curve to the odds curve we need to use exponentiation to reverse the effect of taking the natural log. $e^{(-1.099)}$ will give the odds at X equals 4. That value is .33. Many formulas can get you back to the percentage from the odds, but a common one is: % = odds / (1 + odds). The probability of ant death at X=4 ppm is .25.

The important insights to take away from this slide are:

1) The logit of an S shaped curve is a straight line (or a hyper plane) and meets the assumption of ordinary least squares.
2) The beta values (the estimates) on the SAS printout are used in a formula that predicts the logit line (or logit hyper plane).
3) If you plug X values, and estimates from SAS output, into your "model" formula you will get points on the logit line (or logit hyper plane). You can convert those values into points on the odds line (or odds hyper plane) and to the probability line (or probability hyper plane) using a process similar to what we did above.

In the example above I just picked a value (4) on the plotted logit line and used formulas to move "up" to the odds chart and then up to the probability chart. Plugging X values, and estimates from SAS output, into your "model" formula is the same as selecting a point on the logit line (or logit hyper plane).

ODDS RATIOS

Many resources I have read say that $e^{(\beta)}$ is the odds ratio and I think that statement should be expanded. The meaning of $e^{(\beta)}$ depends on the β – the meaning of a β depends whether the β is β_0 or the β for some X.

Imagine a logistic model with one X and the X is binary. If we are thinking of $ß_0$ then $e^{(ß0)}$ is the *odds* of the event for the base case (X=0). If we are thinking of $ß_1$ then $e^{(ß1)}$ is the *odds ratio* of X_1 =1 - relative to the odds at the base case (X=0). Since X is binary, this could be something like the odds of Males to Females.

Imagine a logistic model with one X and the X is continuous. You can think of a continuous variable having a base case at zero. $e^{(ß0)}$ is the *odds* of the event for the base case (X=0). If X_1=1 then $e^{(ß_1*1)}$ is the *odds ratio* of X_1 =1 relative to the base case of X_1 =0. It is also the odds ratio for any two levels of X that are separated by one unit. If X_1=2 then $e^{(ß_1*2)}$ is the *odds ratio* of X_1 =2 relative to the base case of X_1 =0. It is also the odds ratio for any levels of X that are separated by two units.

The odds ratio has a multiplicative, not additive, effect. This is because: Odds ratio = Odds at ($X_{\text{at some level}}$) / Odds (at $X_{\text{at some level-1}}$). So the odds of the event at $X_{\text{at some level}}$ compared to $X_{\text{at some level -1}}$ is equal to Odds ratio * Odds ($X_{\text{at some level}}$).

ODDS RATIOS CONTINUED

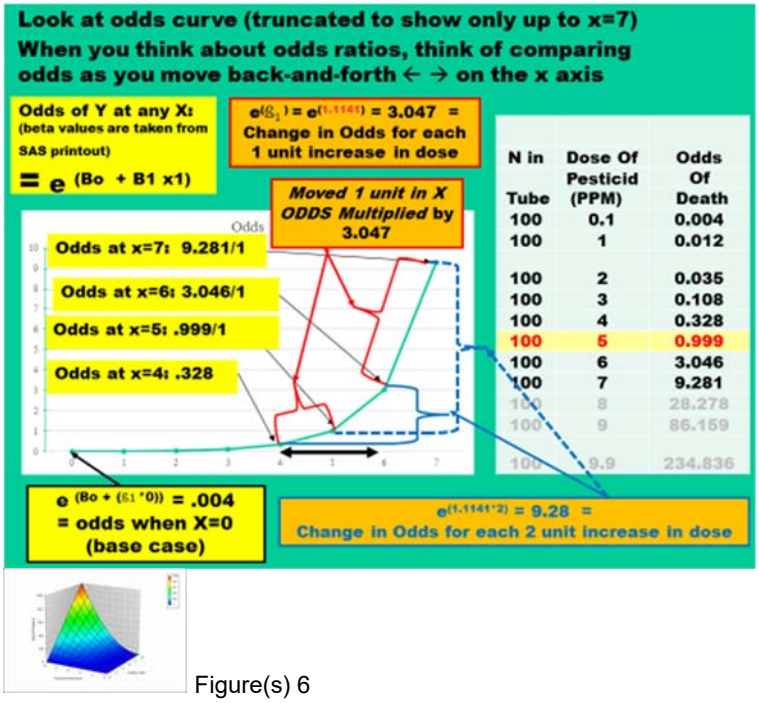
I conceptualize the odds ratio as how the odds change as I slide back-and-forth along the x-axis on the odds chart (middle chart in Figure 5 and to right). Please see the black arrow in Figure 6.

$e^{(ß1)} = e^{(1.1141)} = 3.047$ and this is the ***multiplicative change in the odds*** for each one unit change in X (a 1 unit movement to right or left).

The odds curve looks very flat for small values of X, but if you look at the table in Figure 6 you can see that the odds are three times as large for each one unit increase in X. For small X values the odds are pretty small so a 3-fold increase is not very noticeable. Once the odds of an event get to be above .3 the multiplicative effect caused by a one unit change in X is very apparent on the chart in Figure 6.

Please look at the small chart in Figure 6. If your logistic model has two X variables, you can think of the odds ratio as measuring the change in odds as you move parallel to one of the X axis.

If a model has an interaction term you can calculate a change in odds as you move at an angle to the X axes.



Figure(s) 6

5 / 8) BETA VALUES AND THEIR EFFECT ON THE SHAPE OF THE % CURVES

This paper will only discuss how the betas (column heading is "Estimates" on SAS logistic printouts) affect the % curve for a logistic regression with one X variable.

Figure 5 shows that the logit can be expressed as: logit(Y) = α + βx (Y= a + Bx). This is our high school version of the formula for a straight line – and, remember, the logit line is a straight line.

The next two figures show examples of how changing α and β values *for the logit line* affects the shape of the probability curve.

The 50% point is where half of the subjects "have the event" and is where X equals –α/β.

At that point the percent curve has maximum slope. Note that curves with larger absolute values of β are steeper.

More, and different, curves are shown in Figure 8.

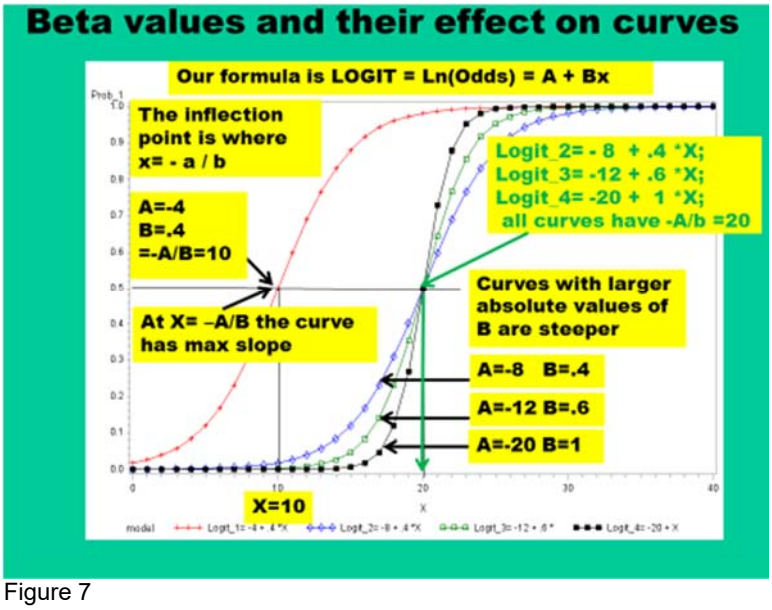


Figure 7

Again, the 50% point is where half of the subjects "have the event" and is where X equals –α/β.

Regardless of whether the probability/percent curve slopes upward as X gets larger, or slopes downward as X gets larger, the 50% point is where x= -α / β.

Regardless of whether the probability/percent curve slopes upward as X gets larger, or slopes downward, the logit of the probability curve will be a straight line (or hyperplane).

Please remember that the α and β, being discussed on these plots, are the values in the "Estimates" column on the SAS printout and are used to calculate the Logit line (bottom chart in Figure 5).

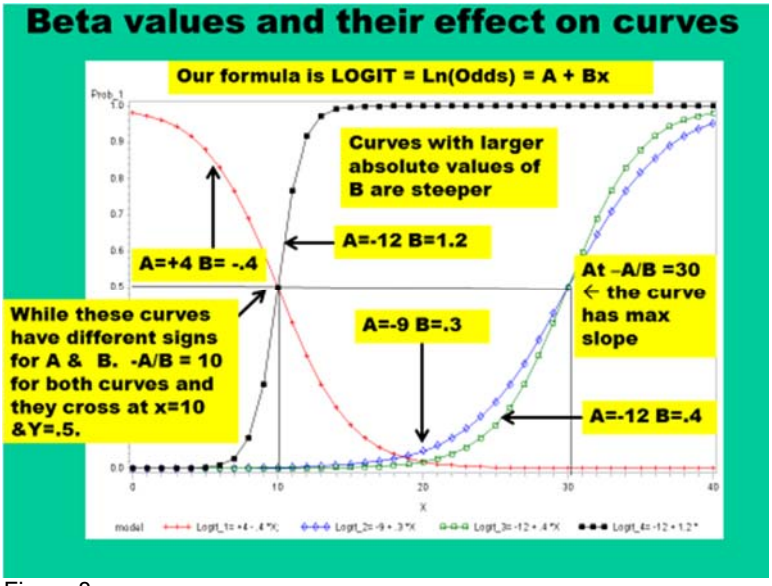


Figure 8

6 / 8) OVERSAMPLING: WHY AND HOW

I think there are two situations when oversampling is justified – but they are very different situations and the benefits to be gained are very different.

The first situation where oversampling can be used is when: the data is all collected, the event being modeled is rare, *and you want to reduce the computer run time for building the model*. Imagine you are on a large bank that has records on every credit card that they issued. You have many cards and many years of activity. You have a few defaults and a great number of non-defaults. In this business situation, you do not have the ability to get any more data and have more than enough data to build a model. You have the all data that you can get, or need, but there is so much data that it is difficult to process.

The real problem is that building the model can take too long – excessive run time. In this case you can build an "over sampled" data set by taking all of the defaults and a random selection of non-defaults. You want to have all of the defaults, because you'd like every variable, and person, to have a chance to affect the coefficients in the model. You're trying to find out how the defaults differ from the rare non-defaults so information about defaults is precious.

We know from the central limit theorem that if you collect four times more observations, generally, the standard deviation is decreased by the square root of four. Adding observations does not give a linear increase in accuracy. Once the number of non-defaults gets to be 2 to 5 times the number of defaults, adding more non-defaults will not improve model performance very much (provided non-defaults were selected randomly). In this situation, it makes sense to reduce run time, by reducing the number of non-defaults, in your data set.

The second situation where oversampling can be used is *when the data yet to be collected and you have a finite amount of dollars/time to spend on the "collecting the data" part of the project*. Imagine you're trying to model buying Lamborghini sports cars in New York City. Buying a Lamborghini is a rare event and if you randomly sampled people in New York you might find very few people in your sample that bought a Lamborghini. This means that any quirks associated with the very few Lamborghinis buyers in your sample might distort the model that you eventually built. This scenario assumes that you don't have enough money in your budget to contact enough people such that a random sample of car owners in New York would produce a data set with a few hundred Lamborghini owners (FYI: If you could use, and did use, a random sample in this project, you would not be spending your money wisely).

In this situation you should spend a higher percentage of your data collection dollars (or at least effort) on contacting Lamborghini buyers than on buyers of regular automobiles. You would like to focus your efforts, and spend your dollars, to ensure that at least 20% of the people in your sample have bought a Lamborghini.

Statisticians have shown that oversampling only affects the β_0 in the estimates column in your SAS printout. If you're building a model with many X variables $B_1 - B_n$ will not be affected.

SAS has an excellent option that makes it easy to model with oversampling. I will illustrate using a data set that ships with SAS Enterprise Miner. In Figure 9, I have built a logistic regression where each person in the data set is one row. There are 1,500 defaults and 45,000 non-defaults. SAS models this very quickly and I consider this to be the gold standard for this model and data set.

```
proc logistic data=CS_ACCEPTS_at_Person_level ;
model GB(event='1')= age TMADD CASH ; run; ;
```

Gold Standard

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.3778	0.0980	197.7233	<.0001
AGE	1	-0.0609	0.00310	386.1673	<.0001
TMADD	1	0.000290	0.000162	3.2050	0.0734
CASH	1	-5.78E-9	5.034E-6	0.0000	0.9991

Figure 9

```
proc logistic data=sampsio.CS_ACCEPTS ;
model GB(event='1')= age TMADD CASH ;
run;
```

OVERSAMPLING TO 50%

Big difference in B0

Please say that again

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.8646	0.1326	197.6015	<.0001
AGE	1	-0.0559	0.00392	202.8637	<.0001
TMADD	1	0.000225	0.000224	1.0063	0.3158
CASH	1	-1.45E-6	5.978E-6	0.0591	0.8079

Figure 10

```
%let PriorProb = 0.0323;  %let PostProb  = 0.5;

data CS_ACCEPTs_W_SampWt;
 set sampsio.CS_ACCEPTS;
sampwt=((1-&priorprob )/(1-&postprob))*(GB=0) + (&priorprob /&postprob)*(GB=1); :

proc logistic data=CS_ACCEPTs_W_SampWt ;
weight sampwt;
model GB(event='1')= age TMADD CASH ;  run; ;
```

Create weighting variable Sampwt

use weighting variable Sampwt

Matches Gold Standard

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.3764	0.3855	12.7478	0.0004
AGE	1	-0.0609	0.0122	24.9452	<.0001
TMADD	1	0.000290	0.000638	0.2070	0.6491
CASH	1	-5.99E-9	0.000020	0.0000	0.9998

Figure 11

Figure 10 is the model that results from oversampling to 50%. There are 1,500 defaults and 1,500 non-defaults in this data set. The numbers in the estimate column are pretty close to those in Figure 9 - except for the intercept. Books say that you can correct the intercept, using a simple formula, if you know the proportions of defaults and non-defaults in both your general population and in your "over sample group". However; for me, the correction did not "correct" the intercept to the exact same number that I got in Figure 9.

The technique shown in Figure 11 did perfectly recover the betas from my "gold standard" and was not difficult to code. As I go forward, in my career, I will use the technique shown in Figure 11.

7 / 8) THE ROC CURVE IN PICTURES

The ROC curve is often used to judge the "predictive power" of a model. Something that is hard to explain on a printed page is that the ROC curve is really concerned with setting classification cutpoints – levels of X where the management decides to "takes action".

You must imagine that we have a different (more expensive/painful test or model) that can very accurately assign people to sick or healthy. Accordingly, we know if people in our training dataset were healthy or sick. We have developed a new, cheaper, less painful test and want to see if it can be used. By "taking action" I mean that management makes a decision (and does something about it). If the logistic is trying to predict disease presence and X is below the cutpoint the doctor says you are healthy (and decides to send you home). If X is above the cutpoint, the doctor decides that you are sick and admits you to the hospital.

In Figure 12, we see histograms of healthy and sick people and their "scores" on our new test (scores are the white numbers on the X – Axis). The yellow people are healthy. The orange people are sick. Both the healthy, and sick, people are plotted on the same X axis – the white numbers going from .1 to .9. We have 25 healthy and 25 sick people in our study.

The test/model produces the prob. of being sick (the X Axis in Figure 12) and is terrible. The distributions of healthy people and sick people plot "right on top of each other". There is no "separation" of the distributions when we use score as X. Admittedly, we will need another picture before the word "separation" makes much sense.

To create a ROC Curve, we need to calculate the cumulative percentage of true positives and the cumulative percentage of false positives for each possible cutpoint. If we set the cutpoint at .15 then we will classify one healthy person as healthy and one sick person as healthy. At .15, cumulative true positives and cumulative false positives are .04. We use .04, .04 as a point on the ROC curve.

The procedure will be to slide the cutpoint to the right and stop every time the cutpoint encounters a subject who is either healthy or sick. You then calculate the cumulative true positives and cumulative false positives and those two numbers become points on the ROC curve. With that as a procedure, we should stop the cutpoint at .25 and do calculations. That slide is not shown in this paper.

Slide 13 shows us setting the cutpoint at .35. With .35 at a cutpoint we classify six healthy people as healthy and 6 sick people as healthy. The cumulative true positives and the cumulative false positives are both .24. We will plot .24, .24 as a point on the ROC curve.

In practice, setting the "final cutpoint" involves balancing the costs of a type I and type II error. One ROC curve evaluates all possible cutpoints.

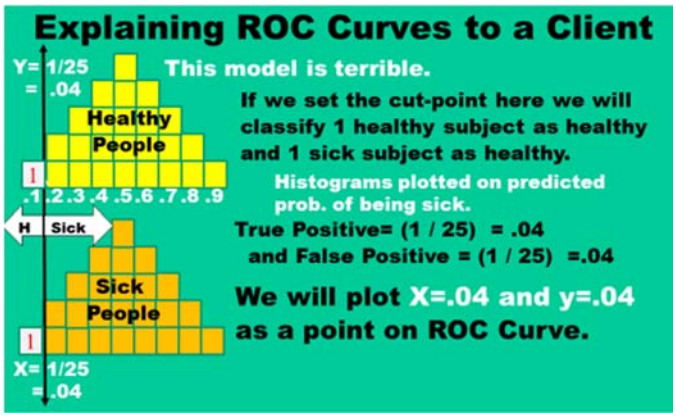


Figure 12

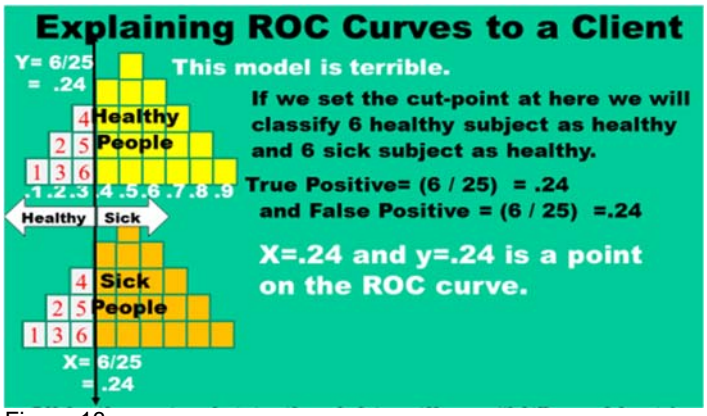


Figure 13

Please look at Figures 12 and 13 to get a better understanding of the concept of a cutpoint.

The chart on the right (Figure 14) shows the calculations for all of the possible cutpoints in this data. Remember this model is terrible. The healthy people and the sick people plot "right on top of each other" and the blood test (or our mathematical model if that's how you'd like to think) does not create any "separation" between the two distributions.

The Excel spreadsheet in the bottom right-hand corner shows the calculations for cumulative true positives and cumulative false positives and the ROC curve is shown in the upper right-hand corner of the slide***. If there is no separation between the two distributions (a terrible model), the plot of the ROC curve will be a 45° line.*** The area under this curve (AUC happens to be equivalent to the c statistic) is .5 and a model that does not predict well has an area under the ROC curve of .5. People describe this model as predicting "only as well as chance".

To understand the logic of the ROC curve we need to contrast this terrible model with a model that is moderately effective and also with a model that's very effective.

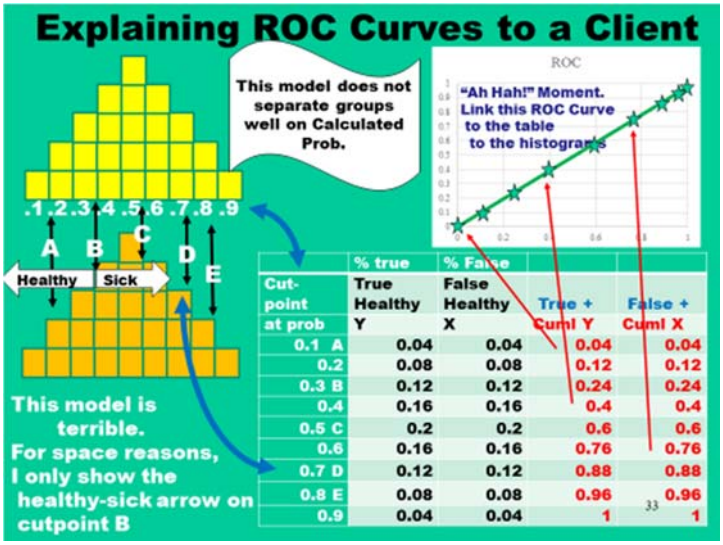


Figure 14

The model in Figure 15 shows a moderately effective model. We plot histograms using modeled prob. of being sick as X. The distribution of sick people is "shifted to the right" because the two groups have different values on the X variable. This shifting of the distributions, by an effective test or model, is the underlying logical process in the ROC curve.

The calculation of cumulative true positives and cumulative false positives is the same as in the previous slide.

In Figure 15, you can see that there is a lack of overlap between the two distributions - both on the right and on the left ends of the axis.

The lack of overlap on the left (small x values) causes a vertical rise in the ROC curve as it starts out from the origin.

The lack of overlap on the right (large X values) causes a horizontal section in the ROC curve as the x-axis moves towards 1.0.

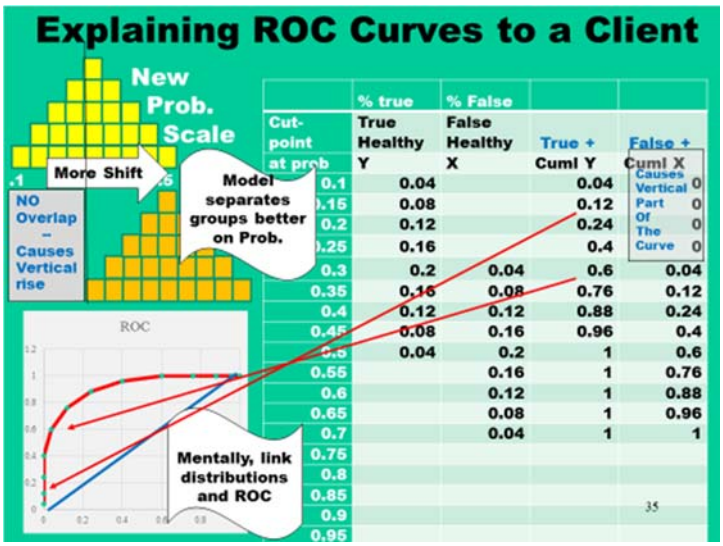


Figure 15

Figure 16 shows the distributions, calculations and ROC curve for a blood test (or model) with excellent predictive power. The distribution of sick subjects has been shifted far to the right and there is little overlap between the distributions. The lack of overlap causes a large vertical rise at the start of the ROC curve. There is a small overlap, at X equals .5, where the model is correctly classifying and mis-classifying at the same rate. The lack of overlap on the right-hand side of the x-axis causes the ROC curve to have a horizontal section. The area under this curve is close to one and the area under a perfectly predicting model is exactly 1.0.

Often people judge the effectiveness of the model only by the area under the ROC curve but Figure 16 suggests that the shape of the curve conveys important information. Look at the chart in the left-hand side of Figure 17. It starts off being an excellent model up until a cutpoint of .25. At that point, with no warning, the model shifts to being no better than random assignment. Between .25 and .55 the model mis-classifies as many subjects as it correctly classifies. In this range the model is terrible. Then the model becomes excellent again. If your score is above .55 the model does an excellent job of classifying subjects as being sick.

Let's look at the effects of the ROC shape. If we set the cutpoint at .25 and say you are healthy, then you are healthy. No sick person ever had a score lower than .25. The model is very valuable, if we set the cutpoint at .25. If we set the cutpoint at .55 and say you're sick, then you are sick. No healthy person ever had a score above .55. So this model can be very useful ***in two limited ranges***.

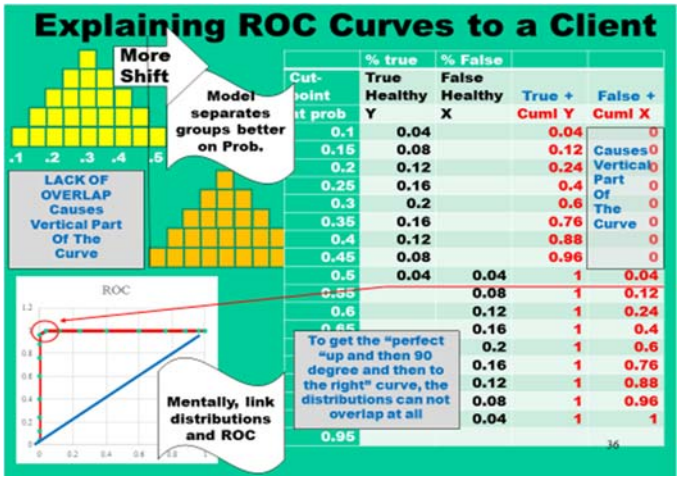


Figure 16

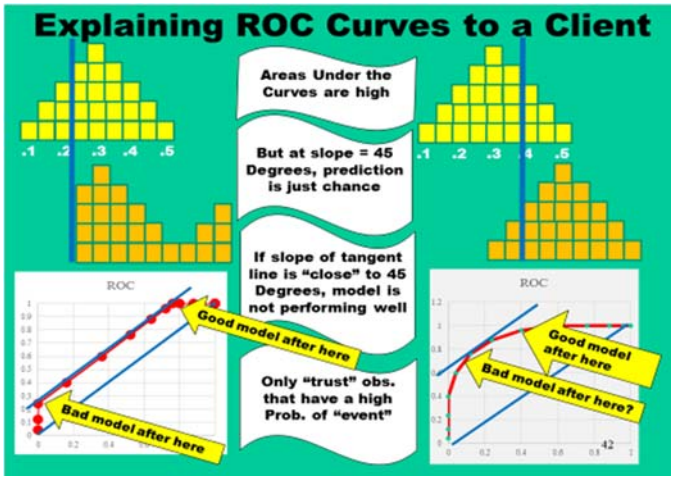


Figure 17

Please look at the chart on the right-hand side of Figure 17. I have read papers where it is suggested that the cutpoint be set where the line tangent to the ROC curve is 45°. I think that suggestion is of limited usefulness. When a tangent to the ROC curve is at 45° the model misclassifies, and correctly classifies, at the same rate. This level of accuracy has limited practical usefulness. If your ROC curve looks like the curve in the right-hand side of Figure 17, I suggest that the cutpoint be moved away from the 45° tangent point and that we recognize the usefulness of the model in answering *different kinds of questions*. If one someone wants to know if they are healthy, we might set the cutpoint at .4 and say "if you're below .4 we think you're healthy". If someone wants to know if there sick we might set the cutpoint it .55 and say "if you're above .55 we think you're sick". If the subject has a score of .48, we should just admit we don't know whether they are sick or healthy. We would recommend using a different (more expensive, more painful) test.

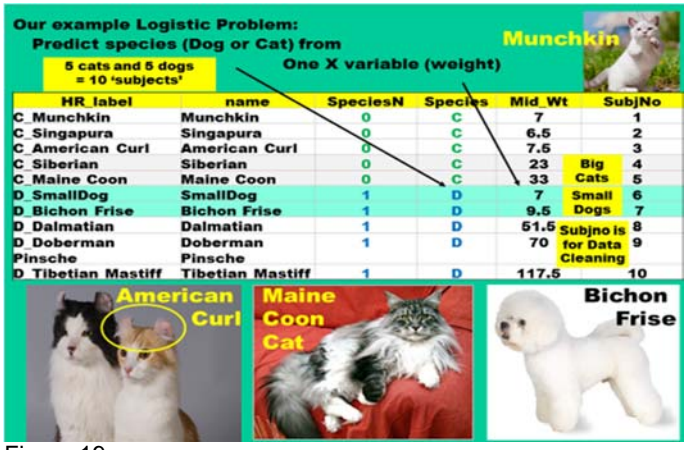
We can link the curves in Figure 17 to ranges on the ROC curve and to the concepts of specificity and sensitivity. Negative results from a highly sensitive test will rule out the disease. Positive results from a highly specific test suggests disease presence. The right hand "model" on Figure 17 is sensitive if the cutpoint is "low" and specific if the cutpoint is set high.

8 / 8) THE C STATISTIC AND THE ROC CURVE

The table of "Association of Predicted Probabilities and Observed Responses" is very common output from a logistic regression. I always found percent concordant, percent discordant and the C statistic difficult to understand from textbook descriptions. In preparation for this paper I found descriptions on the web. I am of the opinion that they are all true but provide very little help and provided no insight. I hope an example with some pictures might be helpful and such an example will be provided below.

The context for this model is that you are trying to build a model that predicts "dog" versus "cat". Weight is the only X variable and you're going to collect information on five cats and five dogs. The data is included in Figure 19. HR label is a human readable label for the observation. ***C***_munchkin means that the munchkin is a ***c***at. The "***D***" in ***D***_Bichon Frise just helps us remember that the silly ball of fluff is a **d**og. We have two very big cats and two very small dogs and these will cause problems for the model. Also; please note that the munchkin and the generically named "small dog" both weigh 7 pounds. Since weight is the only X variable in the model these will be "tied" and the details of interpreting a "tie" will be shown later.

In Figure 20 we see the results of the model. Please focus on the values in the columns named IP_C and IP_D. These columns contained the probabilities of being a cat/dog that were assigned by our model. These columns are the first step in calculating concordance, dissonance and the C statistic. Notice that the Bichon Frise seems much more likely to be a cat than a dog. Notice that the Maine Coon cat is more likely to be a dog than a cat. Notice that the Munchkin and the small dog have the same probabilities. They are both judged to be more likely to be a cat, *but their probability of being a cat is identical at .72994*.



HR_label	name	SpeciesN	Species	Mid_Wt	SubjNo
C_Munchkin	Munchkin	0	C	7	1
C_Singapura	Singapura	0	C	6.5	2
C_American Curl	American Curl	0	C	7.5	3
C_Siberian	Siberian	0	C	23	4
C_Maine Coon	Maine Coon	0	C	33	5
D_SmallDog	SmallDog	1	D	7	6
D_Bichon Frise	Bichon Frise	1	D	9.5	7
D_Dalmatian	Dalmatian	1	D	51.5	8
D_Doberman Pinsche	Doberman Pinsche	1	D	70	9
D_Tibetian Mastiff	Tibetian Mastiff	1	D	117.5	10

Figure 19

out= A02_Scored_Dogs_Cats;

HR_label	SpeciesN	Species	Mid_Wt	Subj No	Formatted Value of the Observed Response FROM	Formatted Value of the Predicted Response INTO	Individual Probability: Species=C IP_C	Individual Probability: Species=D IP_D
C_Munchkin	0	C	7	1	C	C	0.72994	0.27006
C_Singapura	0	C	6.5	2	C	C	0.73461	0.26539
C_American Curl	0	C	7.5	3	C	C	0.72522	0.27478
C_Siberian	0	C	23	4	C	C	0.55781	0.44219
C_Maine Coon	0	C	33	5	C	D	0.43929	0.56071
D_SmallDog	1	D	7	6	D	C	0.72994	0.27006
D_Bichon Frise	1	D	9.5	7	D	C	0.70584	0.29416
D_Dalmatian	1	D	51.5	8	D	D	0.24505	0.75495
D_Doberman Pinsche	1	D	70	9	D	D	0.11854	0.88146
D_Tibetian Mastiff	1	D	117.5	10	D	D	0.01381	0.98619

Figure 20

The second step in calculating concordance dissonance and the C statistic is to, according to several books, do a random matching of all positives to all negatives. Now this would involve a lot of looping and resampling logic and there's a simpler way if we think of the logic for Fisher's Exact Test. Randomly sampling pairs, after enough samples, will converge to the statistics we would get if we paired **all** dogs to **all** cats. There was a reason for building a model on only five cats and five dogs and that reason was that it would allow us to examine all possible combinations and create an exact answer.

The first four rows in Figure 21 show the Munchkin being matched against all the dogs. Since there are five dogs, in our data set the Munchkin "shows up" on rows 1 to 5. Please focus on the columns that are the predicted probability of being a dog.

The logic for concordance, dissonance and ties is to pair up cats and dogs and compare the modeled probability of being a dog. ***If the modeled probability of being a dog is higher for the dog we consider that a characteristic of a good model and increase our account of "concordant pairs" by one.***

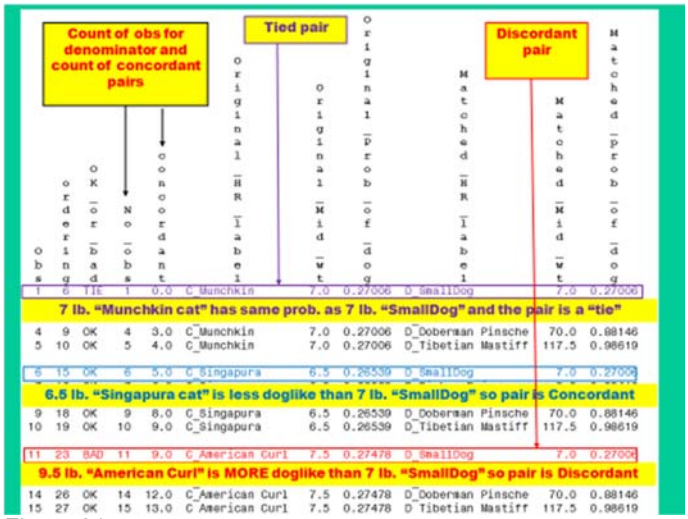


Figure 21

If the modeled probability of being a dog is higher for the cat we consider that a characteristic of a bad model and increase our count of discordant pairs by one. If the cat and the dog both have the same probability of being a dog we consider that a tie. Observation number one in Figure 21 is a tie. Observation number six, in Figure 21, says the small dog is more doglike than the Singapura and this is a concordant pair. Observation number 11, where the American Curl is more doglike than the small dog, is a discordant pair.

In Figure 22 we see the remaining matches for the five cats paired against the five dogs. The matching created a total of 25 comparisons (5 cats paired with each of 5 dogs).

The two big cats (Siberian and Main Coon) are heavier than two of the dogs and are judged to be more doglike. Observations 16, 17, 21 and 22 are discordant pairs. All of the other rows, shown in Figure 22, are Concordant pairs.

At the bottom of the slide you can see calculations that can be made by counting up discordant and concordant pairs.

There were 19 concordant pairs out of 25 pairs, giving us a .76% rate for concordant pairs.

There were five discordant pairs out of 25 pairs, giving us a 5% rate for discordant pairs.

There was one tied pair, giving us a 4% tied rate.

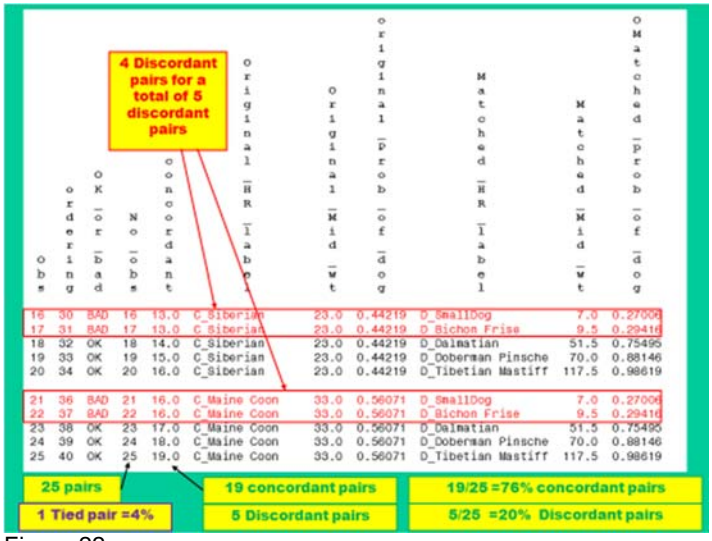


Figure 22

At the bottom of Figure 23 we have retained the calculations from Figure 22 and added both the ROC curve and the table of associated predictive probabilities.

You can see, using the calculations on Figure 22, we have reproduced the percent concordant and the percent discordant and the percent tied. The remaining issue is the C statistic.

To calculate the C statistic, one only has to eliminate ties.

When a pair of observations is tied we increase the count of concordant pairs by .5 and the count of discordant pairs by .5.

Using this new formula the percentage of concordant pairs is .78 and that is the value for the C statistic.

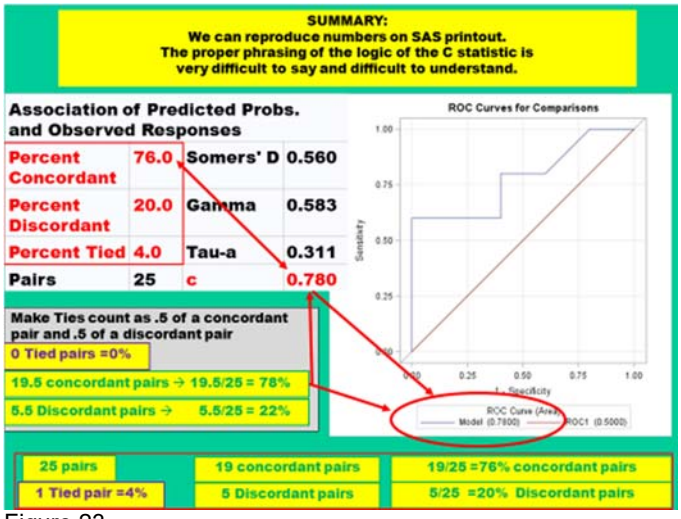


Figure 23

CONCLUSION

This paper presented some thoughts on logistic regression where the concepts, I thought, were better presented using pictures than formulas.

REFERENCES

SAS/STAT 13.1 user's guide: the logistic procedure
a paper by Sing Yo, at SAS institute, that I can no longer find

ACKNOWLEDGMENTS

On yet another paper, I want to thank Peter Flom for his time, his statistical skills and his willingness to have discussions that help other people polish their work.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Russ Lavery Contractor Bryn Mawr, PA E-mail: russ.lavery@veriuzon.net Web: russ-lavery.com