Reconsidering the logit: the risk of individual names

Zoltan, Varsanyi

June 2007

Reconsidering the logit: the risk of individual names

Zoltan Varsanyi[♣]

In this paper I examine whether the probability of default (PD) of an obligor estimated by a logit model can really be considered a good estimate of the true PD. The general answer seems to be no, although in this paper I don't carry out a large scale (simulation) analysis. With a simple set-up I show that the logit has a high potential of 'mixing' probabilities, that is, as[1]signing similar scores to obligors with quite different PDs. I demonstrate how this situation is reflected in the convexity that can often be observed in empirical ROC curves. I think that the results have important implications in the pricing of individual exposures and raise the question of the stability of estimated PDs when the value-combinations of the risk factors underlying the portfolio change. This latter issue also relates to capital calculation, model building and validation as required by the new Basel capital rules. For example, because of the concavity of the risk weight formula a bank may want to avoid PD mixing thereby reducing its capital requirement.

I. Introduction

Logit models are standards in the credit risk assessment of banks' customers. It is widely used in the retail and corporate segment where data scarcity is generally not a problem. The purpose of using these models is – through the assessment of riskiness of obligors – to assign obligors/exposures to homogenous risk-grades which, subsequently, forms (or may form) the basis of capital calculation, pricing, limit setting, etc.

One interesting thing in this process is that while we are interested in the riskiness of *individual* names we (can) assess it using the actual *portfolio* that those names are assigned to.[2] On the other hand, even using the whole portfolio doesn't mean that we could determine the loss-contribution of the names in question (which changes from portfolio to portfolio) – logit models only assess the probability of default (PD) irrespective of concentrations in the portfolio, for example.

That is, there are interesting 'interactions' between the single name and the portfolio levels. This relationship can be further explored by the following question: does a well-fitting logit model really estimates the risk (PD) of a single name correctly? I arrived at this question in connection with a very general problem in economics: when someone wants to sell something he/she can be expected to have some 'informational advantage' over the buyer, so that the latter one may think that the reason of the intention to sell is a 'hidden' problem with the item that is to be sold. This situation can be described by 'informational asymmetry' and may lead to 'adverse selection' in the portfolio of bought items of the buyer, i.e. the buyer will collect goods of inferior quality, for example; the danger of this happening may lead the buyer to offer less for the item than it is actually worth. The problem can be mitigated if the buyer has the opportunity to carry out a thorough examination of the item before the purchase.

The above described situation may appear in banking, for example, when a bank wants to sell its loan(s). It has an informational advantage over the potential buyers, since it has collected

♣ Economist, Magyar Nemzeti Bank (the central bank of Hungary). This paper has not yet been referenced nor discussed. Please send your comments to: varsanyiz@mnb.hu

[2] This can be compared to rating agency ratings where the assessment takes place on a stand-alone basis.

information before and during the life of the loan. Moreover, why would a bank want to sell a loan unless there is a problem with it – a potential buyer could ask.
The question that I examine in this paper is not related to the buyer's side (e.g. how it could make sure it doesn't make a bad deal), but to the seller's, first of all: whether it is able to assess the real riskiness of a single loan. I try to show that while logit models can work well (in terms of fit, discriminatory power, etc.) when there is a need to assign exposures to rating grades, their PD assessment for single names might be incorrect, even if estimated grade PDs are correct. Probably the most important reason why it may happen is that when fitting logit models it is usually taken as granted that the underlying data is actually generated by a logit model and there is a 'bias' towards accepting the results that seem to be good. In fact, this model imposes a strict structure on the data that may not always be appropriate.

II. Logit models – the basics

Instead of giving an in-depth discussion of logit models in this part I describe their basics and highlight those important features that are relevant for our discussion (e.g. how their appropriateness for a data set is judged).
A loan can basically be in two status: either it defaults or not; and it has a certain probability of being in the default state (which variable I denote by *PD* – using italics to differentiate it from the generally used abbreviation for the probability of default). If we assume (as is general) that loans (obligors) are independent from each other then the probability of observing *k* defaults over a horizon in a portfolio of *n* obligors is:

$$P\left(\frac{k}{n}\right)=\binom{n}{k}PD^{k}\left(1-PD\right)^{n-k} \qquad (1)$$

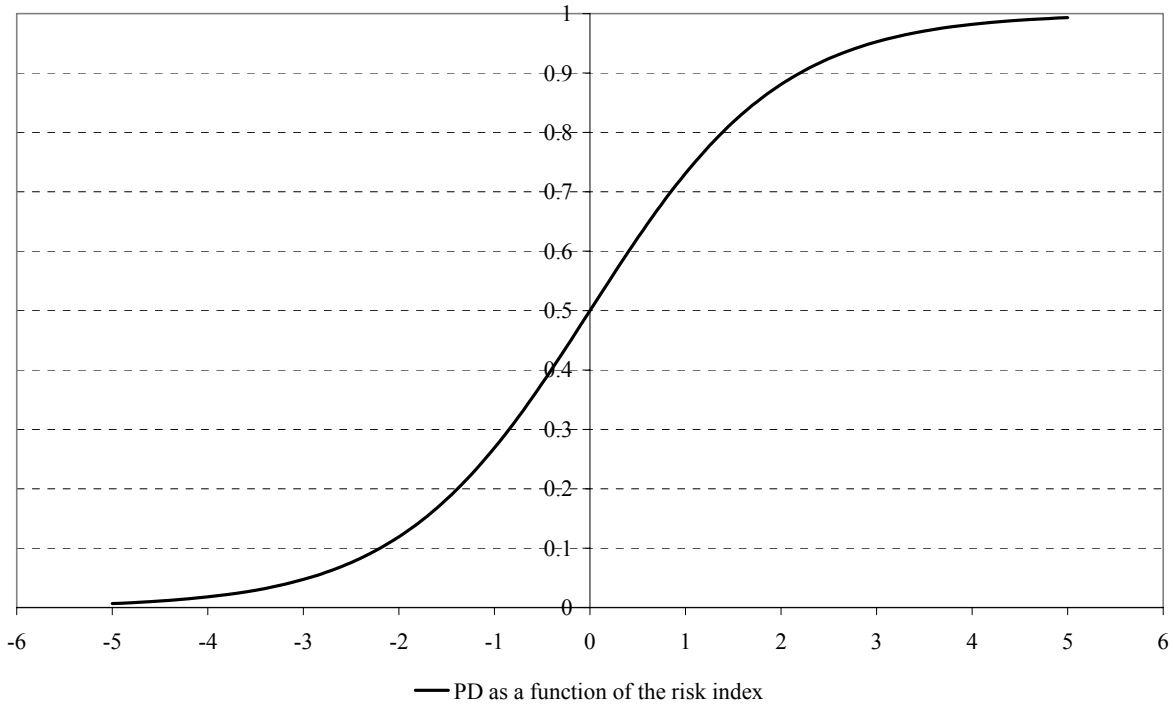
Having a sample of *n* observations with *k* defaults the maximum likelihood principle dictates to estimate *PD* by maximising (1) with respect to *PD*. In the product of the right hand side of (1) we insert *PD* for the *k* defaulted exposures and (1-*PD*) for the rest. Differentiating the right-hand side and leaving the constant term out and equationg the resulting expression with zero leads to the well-known estimation:

$$\hat{PD}=\frac{k}{n}$$

If *PD* is a function of more variables (risk factors) then the principle is the same, except that this time the parameters of the function (the sensitivity of *PD* to the factors) has to be adjusted in the optimisation. This way the estimated *PD* will be a function of the factor values; and the *PD* is usually called a score.
In the logit case the functional form that links the factors and *PD* is the logistic function, that is:

$$\hat{PD}=\frac{1}{1+e^{-\beta x}}. \qquad (2)$$

Subsequently, I will refer to βx as the 'risk index'. The following figure shows the *PD* as a function of the risk index. It can be seen that the higher the risk index the higher the *PD*.



The figure and (2) give a nice interpretation of what the logit model does: *it tries to find such values for the β parameters that for each value of the risk index the observed default ratio in the sample around that value equals the score*. This is not the only criteria, but this is an important one.

III. Can the logit be misleading?

Given our last observation that the score should give a good estimation of the conditional PD around a risk index the main question of this paper arises: is it possible that the optimal logit incorrectly 'mixes' loans with different PDs to have the same score (that is their average PD)? For example, if we have two types of loans, with equal weights in the portfolio, one with *PD*=0.2, the other with *PD*=0.4 and depending on the factors differently, is it possible that the estimated score for both groups will be 0.3?
If the answer to the above questions were 'yes', that would pose several problems: the risk (in terms of PD) of individual loans would not be estimated correctly and the correct price of such loans could not be determined. Moreover, there would be a potential danger that if the portfolio changes the estimated PDs (the scores) would change significantly.

III.1 Is the logit the data generating process (DGP)?

When testing the hypothesis of 'PD mixing' an important question is how one regards the logit model. To my experience, it is always regarded as the true DGP. If this were true, one could generally expect good estimates even with a model that excludes some explanatory variables. If, for example, the factors are uncorrelated, the estimated coefficients will be unbiased and the effect of the excluded variables will appear in the error term; in the presence of correlation between the included and the excluded factors, some effect of the excluded factors will be assumed by the included ones.

What happens when the DGP is not the logit? Then we will try to force a functional form to the data that is different from the true one. However, even in this case we can expect to be able to estimate a logit form – the question then is how good it fits.

III.2 An example

To demonstrate the above I estimate a logit model on simulated data. For the simulation I take two factors, X1 and X2, and take four sub-portfolios. I create their PDs using the logit, using factor-values as described in the following table and all coefficients equalling 1:

Sub-portfolio \ Factor values, PD	X1	X2	PD
P1	-1	0	0.27
P2	0	1	0.73
P3	1	0	0.73
P4	0	-1	0.27

If I simulated all the sub-portfolios and estimated the logit, the estimated coefficients would be around 1. This is not surprising, since for the estimation I used the model that actually generated the data. More importantly, as a consequence, the estimated scores would give an accurate estimate of the PDs in the table – so there were no 'PD mixing', nor its inverse (i.e. when sub-portfolios with the same true PD get different scores).

Now I create another scenario, according to the following table:

Sub-portfolio \ Factor, PD	X1	X2	PD
P1	-1	0	0.1
P2	0	1	0.2
P3	1	0	0.2
P4	0	-1	0.3

What I did was simply overwriting PDs, so that they do not come from a (single) logit model any more. I estimated the equation with a constant term to allow more flexibility in the fitting. The resulting coefficients of the estimation are $\beta 0$=-1.40 (the constant), $\beta 1$=0.34 and $\beta 2$=-0.33, while the scores are, respectively, 0.15, 0.15, 0.25 and 0.25. That is, the PDs of the first and the last two sub-portfolios have been mixed and, additionally, in such a way that the estimated scores are the averages of the PDs of the respective sub-portfolios. This latter observation indicates to some extent that the estimated model is 'good', at least in that it gives correct default predictions – I return to the 'goodness-issue' in the next section.

IV. Bad model – good fit?

In the above last example two widely used tests support the model. The first is that the single factors should have a strong, at least monotone relationship with the PD. It is obvious that as X1 increases the PD tends to increase, while as X2 increases the PD tends to decrease. The second test has already been mentioned above: default rates in rating categories that are based on the estimates scores is correctly estimated by the score in that respective category.
At this point one may note that my example is not very realistic. First, it has a very simplistic structure and second, in the example each exposure (loan) either depends on X1 or X2 thus it can be questioned whether the same model should be fitted to both. I think that both issues can be treated by complicating the structure by more sub-portfolios and factor-value combinations – this is what I would do anyway, since it is required by another test of model goodness, the ROC curve.

IV.1 Common tests of the logit

In this section I briefly discuss tests that I found to be generally used to judge whether the estimated model can be accepted and used in risk management. These test were already mentioned above.

First, it is usually checked, whether potential explanatory variables (factors) have explanatory power in themselves. This is not really a test of the logit model, but of whether a given explanatory variable should be included in the regression at all. I included this test here because later, for the simulation, we will need to justify that we included relevant variables in the model. The test can be carried out using a form of regression with that single variable. If there is no explanatory power, the variable can generally be omitted. At this stage it also should be checked whether there can be any non-linearity in the relationship between the variable and the PD. If there is – for example, much larger as well as much smaller values go with higher PD, while values around the average go with lower PD – the variable should be transformed.

Second, scores should correctly reflect PDs. While from a practical point of view this issue is much more complicated (cf. the issue of 'rating philosophies', or the Basel II regulation), technically, I think it's simply the case and is very important.[3] If scores do not give a good estimation of the true PD, the model can not be that good – although still can be useful practically.

The third test I highlight is the examination of the ROC curve and the calculation of the AUROC (Area Under the ROC) statistic.[4] In the first step we order the sample of defaults/non-defaults according to the score, starting from the worst score. Next, we move from score-to-score and at each move we take the sample that we left behind (all observations that have worse scores than the actual one). In this sub-sample we count all the defaults and divide it by the number of defaults in the whole sample – thus we get the Hit Rate (HR); then we count all non-defaults and divide it by all non-defaults in the sample – thereby getting the False Alarm Rate (FAR). We repeat this calculation for each score value, and, finally, graph HR against FAR. The better the discriminatory power of the model, the higher the proportion of defaults and the lower the proportion of non-defaults in the bad-score region and, thus, the steeper the ROC curve.

Finally, it is important, that these tests are carried out both on the sample that was used for model estimation ('in-sample test') and on a different sample ('out-of-sample test'). It is usually done by splitting the sample available for the analysis into a larger part (e.g. 70%) directly for the estimation and a smaller part put apart for validation. Here, it can be a question how the sample is split – e.g. randomly or according to some schedule – to which issue I return a bit later.

IV.2 A more complicated example

To examine how the above described tests behave on a sample which is not generated by a logit model I carried out some simulations. As in the example in section III.2, there are two factors, X1 and X2; however, now there are 10 sub-portfolios and the PDs belonging to these sub-portfolios are generated by a specific formula. Namely, to assign PD to a pair of values of X1 and X2 I first transformed these factor values into numbers between 0 and 1 using the normal cumulative distribution function and then I used the Clayton copula to map these two

[3] For the issues mentioned see Heitfield [2005] and Basel [2005], respectively.
[4] See Tasche [2005].

values onto (0,1).[5] That is, the two steps are (θ is a parameter that was set at 2.5, the expected value and the standard deviation of the normal distribution was 1 and 2, respectively):

- $X1, X2 \to N(X1) = p1, N(X2) = p2$
- $PD = (p1^{-\theta} + p2^{-\theta} - 1)^{-1/\theta}$

By this procedure I gave a structure to the data and this structure is quite different from the logit. The following table shows the values of X1 and X2 and the PDs:

Table 1: scenario for the simulation and the estimation

Sub-portfolio \ Factor value, PD	X1	X2	PD
P1	1	-2	0.06
P2	-3	0	0.30
P3	-1	2	0.45
P4	0	-3	0.02
P5	1	3	0.16
P6	2	1	0.07
P7	3	-1	0.02
P8	2	-2	0.05
P9	1	0	0.15
P10	1	-3	0.02

In the next step, I simulated data with the above parameters having all the sub-portfolios the same weight and then I estimated a logit model on that simulated data. Finally, I evaluated the tests described in section IV.1.

IV.3 Applying the tests to the model estimated on the simulated data

In this section I examine how the model performs on a randomly selected sample. At each test that reflects a good performance of the model I highlight the problems that arise and some signs that can suggest there are such problems.

IV.3.1 The relationship between the PD and the individual factors

Examining the relationship between the factors and the PD we can look at the DGP as well as the logit model estimated with a single factor as explanatory variable. The actual data that I used for the analysis in this part fulfils the requirement of individual explanatory power: there is a strong correlation between the factors and the PD (-0.75 and 0.64, respectively) and the factor coefficients in the single-variable models were highly significant.[6] This shows that, indeed, at the end of the model-selection process the factors would have been selected into the final logit model.

IV.3.2 The PD of rating grades

The following table shows the true PDs of sub-portfolios as well as the estimated scores:

[5] A copula is used here purely for technical reasons and not for modelling dependencies.
[6] However, the correlation strongly depends on what pairs of values of X1 and X2 one uses in the calculation of PD: if, for example, in Table 1 I change the X1 value for sub-portfolio P1 from 1 to -3 the correlation grows to -0.5 from -0.72.

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Score	*0.04*	*0.34*	*0.31*	*0.04*	*0.22*	*0.08*	*0.02*	*0.03*	*0.08*	*0.03*
PD	0.06	0.31	0.45	0.02	0.16	0.07	0.02	0.05	0.15	0.02

Comparing the PDs and the scores it can generally be said that the two measures differ significantly. For example, for sub-portfolio P3 the score (that is, the estimated PD) equals 31% whereas the true PD is 45%; for sub-portfolio P5 there is a score of 22% versus a PD of 16%.

However, if one creates rating grades based on the scores, that does not necessarily lead to problems: the score of a grade will give a good estimation of PD in that grade. Let's see, for example, the grade made up of P1, P7, P8 and P10 – those sub-portfolios that have score around 3%: although scores estimate individual sub-portfolio PDs poorly (for example, 3% instead of the true 5% in the case of P8), in the grade comprised of these four sub-portfolios the average PD (4%) is relatively close to the average score (3%). Grouping P2, P3 and P5 (although this latter one has somewhat smaller score than the former two) we have an average score of 29% against an average PD of 31%. Moreover, for the whole sample the average score is 12% and the PD equals 13%.

What we have just seen is the 'theoretical' relationship: I used the 'true' PD as calculated according to section IV.2. Now, in the following table I show the relationship between the PD and the score in the test- and the validation sample:

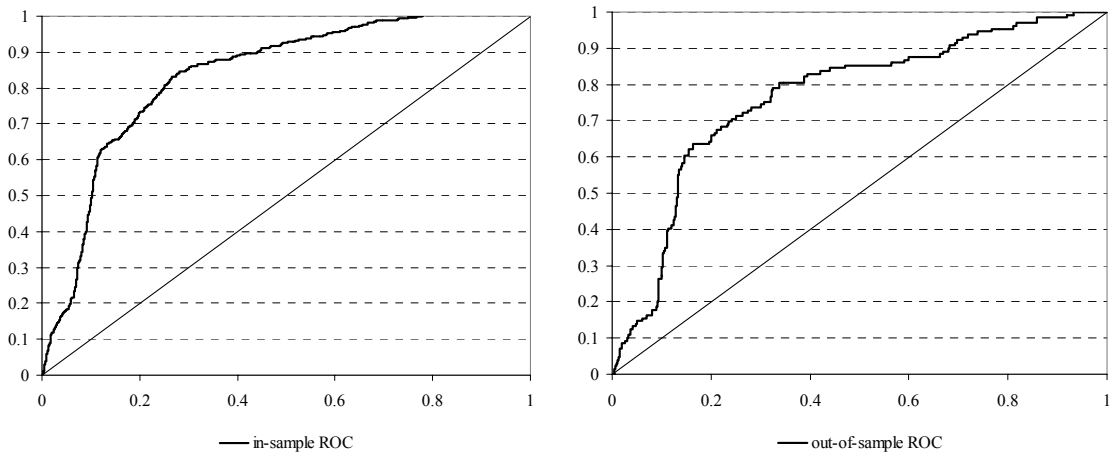
	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Score	*0.04*	*0.34*	*0.31*	*0.04*	*0.22*	*0.08*	*0.02*	*0.03*	*0.08*	*0.03*
PD - in-sample	0.04	0.26	0.46	0.03	0.12	0.05	0.01	0.04	0.14	0.03
PD - out-of-sample	0.03	0.23	0.45	0.00	0.12	0.07	0.05	0.04	0.20	0.02

As regards grades based on the scores, the average PD in the grade containing P1, P7, P8 and P10 has an average score of 3% and an average PD of 3% in-sample and 3.5% out-of-sample. The average in-sample PD in the grade that contains P2, P3 and P5 is 28%, the out-of-sample PD is 26% that compares with a 29% average score.

What these results show is that *the PD of individual names might be rather poorly approximated by the score coming from a logit model; at the same time the logit model can be expected to give a good estimation of PDs at the rating grade level*.

IV.3.3 The ROC and the AUROC

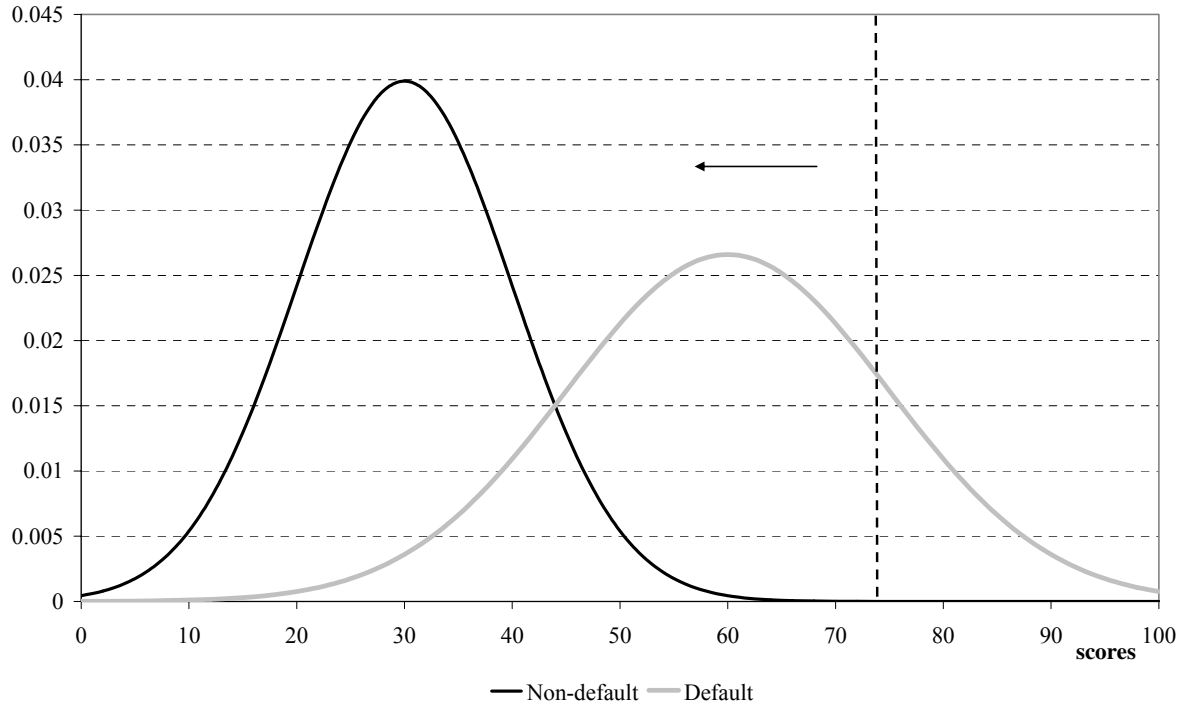
Finally, I calculate FAR, HR and AUROC and check whether these indicators support to accept the estimated logit model. The next two figures show the ROC curve in-sample and out-of-sample:



Evaluating the curves visually at first it can be said that both represent a good discriminatory power.[7] More importantly, the AUROC indicator which I calculated by approximating the ROC curve with a step-function from below equals around 0.79 in-sample and 0.76 out-of-sample!

IV.3.3.1 The ROC should be concave

Our first conclusion is that we estimated a good model to the data. However, if we want to dig into the issue more deeply, we have to notice that the ROC curve is not concave, but there are convex parts in it, for example around the FAR=0.1 region.
Why do we require a concave curve? The following figure shows the distribution (density) of hypothetical non-defaulted and defaulted exposures over the scores.



[7] In fact, we have only 10 points of the ROC curves that can be seen in the figures, since there are only 10 factor-value combinations. For this reason the calculated AUROC measure is not fully accurate, although it can be expected to give a relatively good approximation.

To get the ROC curve we need to calculate the FAR and the HR indicators. The calculation of the FAR amounts to simply moving the dashed black line to the left (towards better scores) and calculating the probability mass that the non-defaults' distribution has to the right of the curve. Similarly, the HR is calculated as the probability mass of the defaults' distribution to the right of the curve. Thus, a change in the HR (FAR) can be expressed as a change in the probability that a given score occurs as we change the score slightly, with the condition of default (non-default). We can express it for the HR, using Bayes' theorem as:

$$P_1(s) \equiv P(s \mid y=1) = \frac{P(y=1 \mid s)P(s)}{P(y=1)} = \frac{sP(s)}{P(y=1)}, \tag{3a}$$

where y is the default indicator and s is the score. For the FAR we have:

$$P_0(s) \equiv P(s \mid y=0) = \frac{(1-P(y=1 \mid s))P(s)}{1-P(y=1)} = \frac{(1-s)P(s)}{P(y=1)}. \tag{3b}$$

It is important to note that $P(s)$ above is the *density* of the score distribution; more importantly, the same is true for $P_0(s)$ and $P_1(s)$ – that is, the formulas express the *change* in the distribution of defaulted and non-defaulted exposures over scores. Now, concavity means that the derivative of the HR with respect to the score is smaller than that of the FAR up to certain scores value, above which it becomes larger; and the ratio of the two measures is increasing in s strictly monotonely.[8] This is the same as saying that the area under the density of non-defaults for certain score values is larger in the case of certain scores (smaller scores where there are more non-defaults) and above a level it becomes increasingly smaller than the area under the density of defaults. The ratio of the change in HR and FAR can be expressed as:

$$\frac{\frac{sP(s)}{P(y=1)}}{\frac{(1-s)P(s)}{1-P(y=1)}} = \frac{1-P(y=1)}{P(y=1)}\frac{s}{1-s},$$

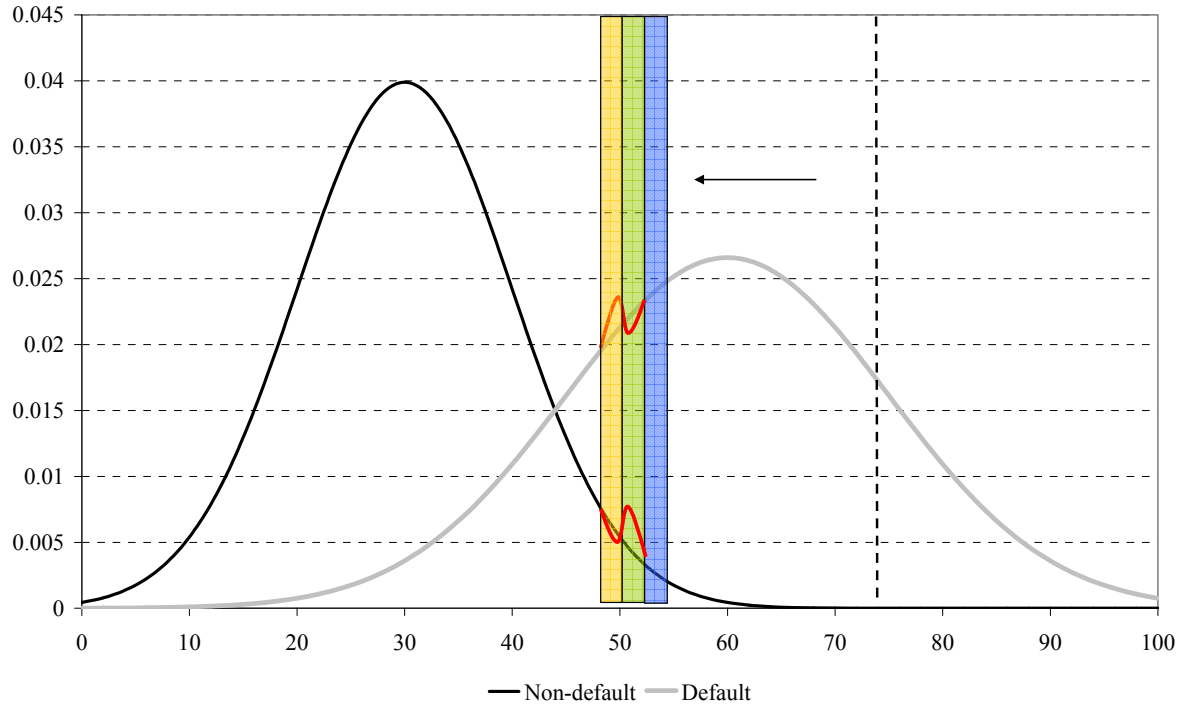
which is an increasing function of s and has values both above and below one – so that, 'theoretically', we should have a concave ROC. Why don't we have it then?

IV.3.3.2 When the ROC has convex regions

One explanation, which is the primary focus of this paper, for the convexity in the ROC curve might be the probability-mixing by the logit, i.e. its tendency to give similar scores to obligors with different PDs. As a result, the last equality in (3a) and (3b) will not hold – the score will not always reflect the conditional (on the factors) PD correctly. To demonstrate the situation I denoted three score-domains on the figure below, each with the same width, with different colours.[9] The changes in the points of a ROC curve are produced by adding the mass under the grey curve in a colour-band to the HR and adding the mass in the same band under the black curve to the FAR. In the base case as we move to the left we add always less and less to the HR and more and more to the FAR. To put it differently, the ratio of the change in the HR and the change in the FAR decreases as we move from the blue band to the green or from the green to the orange – thereby the ROC curve becomes concave, as we saw it above.

[8] Although in the figure we moved from the worse scores to the better ones, in this argument we move in he opposite direction, starting from the best score.

[9] The curve is not linked to the simulation, so the numbers are not comparable.



What happens when the densities are not monotone in the region where they should be (like where the green and orange colour-bands are)? Such deviations are shown in the figure by red lines. For example, in the case of defaulted exposures, there are less defaults in the sample in the green band and more in the orange band than in the base case. The deviations depicted in the figure will have the following effect: as we move from the blue band to the green the quotient of the increase in the HR and the FAR will be even smaller than in the base case. This will increase the concavity of the curve. As we move from the green band to the yellow, however, the quotient increases, since in that region there are more defaults and less non-defaults than in the base case. The increase in the quotients will lead to such increase in the curve that it turns into convex. At the same time, the deviations can be constructed (red lines can be drawn) in such a way that the average PD in the green *and* the orange bands taken together is the same as the average of the scores in this region.

IV.3.4 Out-of-sample test: how should sampling be carried out?

When carrying out the out-of-sample test I split the sample into estimation and validation sub-sample randomly. Without going deeper into the issue, it may be worth splitting the sample according to groups of similar factor-value combinations. Then, if the structure of the data is really different from logit, the model fitted on the estimation sample should not fit well on the validation sample. The question is still open, however, what an institution should do when it encounters such a problem.

V. Conclusion

In this paper I explored some effects of estimating a logit model to data that has an actual structure different from the logit. The main conclusion is that the logit in such a case easily leads to ‘probability-mixing’, that is, the model may assign similar scores to obligors with rather different default probabilities; convexity in ROC curves can reflect such a problem.

References:

Basel [2005]: ‘International Convergence of Capital Measurement and Capital Standards’, Basel Committee on Banking Supervision, BIS, November 2005 (update)

Heitfield, E. [2005]: ‘Dynamics of rating systems’, in: ‘Studies on the Validation of Internal Rating Systems’, BIS Working Paper No. 14, February 2005

Tasche, D. [2005]: ‘Validation of internal rating systems and PD estimates’, in: ‘Studies on the Validation of Internal Rating Systems’, BIS Working Paper No. 14, February 2005, or: www.defaultrisk.com/pp_test_04.htm