1 Introduction

A banking industry standard approach to estimating future unexpected losses to credit portfolios relies on three modeled risk measures: probability of default, loss given default, and exposure at default. In this paper we focus on recent developments in the modeling of exposure at default (EAD) which is defined in Qi (2009) as:

“... a bank’s expected gross dollar exposure (including net accrued but unpaid interest and fees) for a facility upon the borrower’s default. For fixed exposures, such as bullet or term loans, EAD is simply the amount outstanding at the time of capital calculation (plus accrued but unpaid interest and fees). For variable exposures, such as lines of credit, EAD is the current outstanding plus an estimate of additional drawdowns and accrued but unpaid interest and fees up to the time of default.”

As was also stated in Qi (2009) the modeling of EAD lags behind probability of default and loss given default in academic and industry research coverage. There have, however, been several recent papers published in the area of EAD modeling including Qi (2009), Moral (2006), Jimenez et al. (2007), Jimenez et al. (2009). A common theme amongst these papers is the use of factors related to EAD, such as loan equivalent (LEQ) value or credit conversion factors (CCF), to model EAD itself. That is, EAD is rarely modeled directly as the dependent variable in a regression or other general linear model. Instead a factor such as LEQ or CCF, which are both functions of EAD and other loan characteristics, are usually modeled as the response variable in a regression. Predicted EAD values are then obtained based upon the fitted factor values and the algebraic relationship between EAD and the particular specification of the factor that was employed in the regression. We will discuss the details of the various conversion factor and commonly used factors for modeling in the next section.

In slight contrast to this approach, Moral (2006) does in fact explore using the dollar change in exposure from a reference time point, such as one year prior to default, to the default date. That is, he uses dollar EAD minus the dollar drawn amount at the reference time point as a potential response variable. In this research here we take this concept one step further and directly model dollar EAD as the response variable. This paper will focus on some of the potential modeling pitfalls of using derived factors such as LEQ and CCF to indirectly model EAD, which include limitations on choice of model specifications amongst others. By directly using dollar EAD as a response variable these pitfalls can be avoided. We will also provide empirical evidence that modeling EAD directly may be the most parsimonious as well as predictively powerful approach. The rest of the paper is organized as follows: Section 2 provides additional background on LEQ and other derived EAD factors and discusses the modeling constraints they impose. The data and summary statistics used of our empirical modeling comparison is provided in Section 3. Results of our modeling comparison are found in Section 4. Finally Section 5 concludes.

2 EAD and Derived Factors

There are several factors derived from EAD that have been proposed as response variables in the research literature. The most common of these factors is LEQ which was used as the primary modeling variable in (ref:Aratan and Jacobs). We define LEQ in the context of EAD in the following fashion.

\[
EAD(T) = Drawn(t) + LEQ \times [Commitment(t) - Drawn(t)]
\]

\[
= D(t) + LEQ \times [C(t) - D(t)]
\]

(1)

Note that we will consider the EAD to be observed at time \( T \), i.e. the time of default, whereas the drawn amount on the loan and the commitment level (or credit limit) are observed at time \( t \) in general. Usually, time \( t \) is some time point prior to default, such as one year, thus, \( t < T \). Note that in general we can also allow the LEQ value to vary over time as well, i.e. \( LEQ(t) \) for multiple observations of the loan equivalent values across time. This would imply a time series of LEQ values for each facility, i.e. loan or line of credit. However, from an estimation standpoint this raises the question of whether to use traditional ordinary least squares (OLS) regression techniques, time series techniques or even a panel data approach. Moreover, if there is in fact a time series of LEQ values rather than one static LEQ value this raises the question of whether each facility has one true LEQ value or rather we only observe realizations over time of a LEQ random variable. For the moment, let us assume just a single observation of the LEQ value for each facility. Thus, we do not include a time component in our notation for the LEQ values.

Simple algebraic manipulation of (1) yields the calculating formula for the observed realizations of the LEQ values as a function of \( EAD(T), C(t) \) and \( D(t) \),

\[
LEQ = \frac{EAD(T) - D(t)}{C(t) - D(t)} = \frac{EAD(T) - U(t)}{1 - U(t)}
\]

(2)
where \( U(t) = D(t)/C(t) \) is the proportion of utilization at the reference time point. To begin, we see from Equation (2) that when the dollar drawn amount is equal to the dollar commitment amount then \( U(t) = 1 \) and that LEQ is undefined. This situation is by no means a rarity in loan characteristics. Business intuition should probably dictate that predicted EAD should simply be equal to the current drawn amount, which in this case is actually equal to the commitment level. Nonetheless, the LEQ value is not defined and creates difficulties when using LEQ values as the response variable.

In addition, as the utilization rate approaches one from below the LEQ value will asymptote to either plus or minus infinity depending upon whether \( EAD(t) > D(t) \) or \( EAD(T) < D(t) \), respectively. It could also potentially be the case that \( U(t) > 1 \) if the facility is overdrawn at time point \( t \). This presents further complications. Furthermore, when \( U(t) \) is close to one small changes in the proportion of utilization result in quite drastic changes in the associated LEQ value. Such instability is undesirable in a response variable.

Moreover, we have started with a variable that, in theory, is supposed to exist on the unit interval but in actuality exists on the entire real line. In practice, this is usually dealt with by censoring the observed LEQ values to exist on the unit interval. If ordinary least squares (OLS) is then utilized on a censored response variable the resulting regression coefficients will be biased and will not converge in probability to the population parameter values.

There is a further concern here with respect to using LEQ values as a response variable. It is important to keep in mind that the most often the end goal of such an EAD estimation exercise is to forecast predicted EAD which is a required component in calculating economic and Basel capital levels. Usually economic and Basel capital is forecast over a one year time horizon and therefore EAD is as well. Therefore, as practitioners we would want to use any available information at the reference time point in order to predict future EAD values. However, LEQ values are a function of observable variables at both the reference time point as well as the time of default. This is less than ideal when using LEQ values a response variable since we will have necessarily included observable variables at the reference time point in our response variable. Rather than including variables that are observable at the reference time point in our response we should consider using them as explanatory variables instead.

Clearly these concerns present problems when using LEQ values as a response variable. Despite these shortcomings, the main point to be taken away is the fact that the LEQ value can be expressed as a deterministic function of \( C(t) \), \( D(t) \) and thus \( U(t) \). Will focus on this point and illustrate additional pitfalls later in the paper. Before doing so though, we provide a brief overview of some of other potentially factors used.

Various other factors have been proposed in the literature as alternatives to LEQ for modeling EAD. (ref:Moral) suggests a credit conversion factor (CCF) designed to capture EAD’s proportion of the total dollar credit limit. More formally this is

\[
CCF(t) = \frac{EAD(T)}{C(t)}. \tag{3}
\]

(ref:Jacobs) further proposes a slight variation of the CCF, which he also calls the CCF, with the current utilized dollar amount replacing the dollar credit limit in the denominator. To avoid confusion will call this ratio as the utilization conversion factor (UCF) and define it as

\[
UCF(t) = \frac{EAD(T)}{U(t)}. \tag{4}
\]

Both CCF and UCF are more stable that LEQ at points where \( U(t) \) is close to one, and are both well defined when \( U(t) = 1 \). However, both factors are similar to LEQ in that they can be expressed as a deterministic function of \( C(t) \) or \( U(t) \). The discussion that follows demonstrates some potential problems inherent with using LEQ as a modeling variable for EAD, however, the problems inherent with using LEQ are more generally relevant to CCF, UCF, or any factor related to EAD that is a deterministic function of \( C(t) \), \( D(t) \), or \( U(t) \).

To better understand the problems that arise in using the above factors as response variables for EAD modeling we will look at a simple example using LEQ values as a response variable in a regression set up. Let us assume for a moment the following regression set up where we regress LEQ values on a number of explanatory variables including the utilization rate, which we have denoted as \( U(t) \). This approach is not uncommon in practice and can be found in (ref:Qi) and (ref:Araten and Jacobs) for example. Our set up is as follows. Here for ease of exposition
we drop the time component in the notation and just use \( U_i \) for the \( i^{th} \) observation.

\[
LEQ_i = \beta_0 + \beta_1 U_i + \sum_{j=2}^{p} \beta_j x_{ij} + \epsilon_i \quad \text{for} \quad i = 1, \ldots, n
\]

Note for the moment we assume that the other explanatory variables, \( x_{i2}, \ldots, x_{ip} \), do not involve the commitment level nor the dollar drawn.

What we are proposing here is extremely unconventional from a regression set up viewpoint. In particular, note that we have already demonstrated that observed realizations of the LEQ values are calculated based on a deterministic function involving the utilization rate. This is highly unorthodox and potentially violates some of the key underlying assumptions of a regression set up and specifically traditional OLS estimation. In this set up the predicted LEQ values can be obtained based upon estimated values for the regression coefficients.

\[
\hat{LEQ}_i = \hat{\beta}_0 + \hat{\beta}_1 U_i + \sum_{j=2}^{p} \hat{\beta}_j x_{ij} = k_i + \hat{\beta}_1 U_i
\]

Here \( k_i = \hat{\beta}_0 + \sum_{j=2}^{p} \hat{\beta}_j x_{ij} \) is a constant that does not involve the utilization rate, the commitment level nor the dollar drawn, but it is contingent upon the particular other explanatory variables included many of which may in fact be facility level specific, such as obligor risk rating and age of the facility to name just two. Thus, each facility will have a different value for \( k_i \). This fact is important and will be addressed further below. Turning to predicted EAD values based on the fitted LEQ values we have,

\[
\hat{EAD}_i = D_i + \hat{LEQ}_i \times (C_i - D_i) = k_i C_i + \left(1 + \hat{\beta}_1 - k_i \right) D_i - \frac{\hat{\beta}_1}{C_i} D_i^2
\]

where we have made use of the fact that \( U_i = D_i / C_i \). Again, here we have omitted the time component in order to simplify the notation. Note that we obtain a quadratic function in the dollar drawn amount. For a fixed commitment level, \( C_i \), as well as holding the other explanatory variables constant we clearly see that we will have a parabolic relationship with predicted EAD on the vertical axis and the dollar drawn amount on the horizontal axis.

The parabola will either open upwards or downwards depending upon the sign of \( \hat{\beta}_1 \). This then implies that over a certain range of dollar drawn values the predicted EAD value will be decreasing in the dollar drawn amount. We can determine the exact value of the range by finding the extrema associated with the predicted EAD value. Taking the partial derivative of predicted EAD with respect to the dollar drawn amount and setting the result equal to zero yields the following extrema.

\[
D_i^* = \frac{C_i \left(1 + \hat{\beta}_1 - k_i \right)}{2\hat{\beta}_1}
\]

If the sign of \( \hat{\beta}_1 \) is negative then the parabola will open upwards since the entire coefficient of \( D_i^2 \) will be positive, provided of course that \( C_i > 0 \) as it should be. We need to ensure for \( D_i \in (0, C_i) \) that predicted EAD is a monotonic increasing function and predicted EAD is in fact positive. For the case when \( \hat{\beta}_1 < 0 \) and the parabola opens upwards, these requirements are equivalent to \( 1 \) \( D_i^* < 0 \) and \( 2 \) predicted EAD evaluated at \( D_i = 0 \) is positive. Please refer to Figure 1 to help visualize the various cases that can arise.

The two conditions stated above will ensure that predicted EAD, as a function of \( D_i \), will be a positive and monotonic increasing function for \( D_i \in (0, C_i) \). Thus, we have the following two conditions when \( \hat{\beta}_1 < 0 \).

\[
EAD_{i|D_i=0} = k_i C_i > 0 \quad \Rightarrow \quad k_i > 0 \quad \forall \quad i = 1, \ldots, n
\]

\[
D_i^* = \frac{C_i \left(1 + \hat{\beta}_1 - k_i \right)}{2\hat{\beta}_1} < 0 \quad \Rightarrow \quad k_i < 1 + \hat{\beta}_1 \quad \forall \quad i = 1, \ldots, n
\]

If on the other hand, \( \hat{\beta}_1 > 0 \) then the parabola will open down. Please refer to Figure 2 for an illustration of the various cases that can arise. We still require that \( EAD_{i|D_i=0} = k_i C_i > 0 \), which again implies that \( k_i > 0 \) assuming \( C_i > 0 \) as it should be. However, the second condition that predicted EAD, as a function of \( D_i \), is a monotonic increasing function for \( D_i \in (0, C_i) \) is given by the following.

\[
D_i^* = \frac{C_i \left(1 + \hat{\beta}_1 - k_i \right)}{2\hat{\beta}_1} > C_i \quad \Rightarrow \quad k_i < 1 - \hat{\beta}_1 \quad \forall \quad i = 1, \ldots, n
\]
Thus, under the case when $\hat{\beta}_1 > 0$, the above conditions imply that $k_i \in \left(0, 1 - \hat{\beta}_1\right)$ for $i = 1, \ldots, n$. Note that we can combine both of these cases, i.e. when $\hat{\beta}_1 < 0$ and $\hat{\beta}_1 > 0$, into the following succinct condition $k_i \in \left(0, 1 - |\hat{\beta}_1|\right)$ for $i = 1, \ldots, n$.

What is important here, is that if facility level explanatory variables are included in $x_{i2}, \ldots, x_{ip}$ then the above conditions need to hold for all facilities considered in the portfolio. This would have to be checked for all facilities both in the data set used for estimation as well as the portfolio of facilities for which this model would be used to predict future LEQ and EAD values.

It is the case, however, that this problem with inclusion of the utilization proportion as an explanatory variable in the model can be overcome with a simple transformation. In the above situation if we want to include the utilization proportion as an explanatory variable, yet still obtain a monotonic function for predicted EAD versus dollar drawn, we need to consider the following transformation.

$$LEQ_i = \beta_0 + \beta_1 (1 - U_i)^{-1} + \sum_{j=2}^{p} \beta_j x_{ij} + \epsilon_i \quad \text{for } i = 1, \ldots, n$$
Figure 2: Predicted EAD versus current dollar drawn amount when $\hat{\beta}_1 > 1$. Only case (d) is acceptable.

To see why this is the case consider the predicted LEQ values and predicted EAD.

$$\hat{\text{LEQ}}_i = \hat{\beta}_0 + \hat{\beta}_1 (1 - U_i)^{-1} + \sum_{j=2}^{p} \hat{\beta}_j x_{ij} = k_i + \hat{\beta}_1 (1 - U_i)^{-1}$$

$$EAD_i = D_i + \hat{\text{LEQ}}_i \times (C_i - D_i) = \left(k_i + \hat{\beta}_1\right) C_i + (1 - k_i) D_i$$

Thus, we obtain a linear, and necessarily monotonic, function in $D_i$ that will be either increasing or decreasing depending upon the sign of $1 - k_i = 1 - \hat{\beta}_0 - \sum_{j=2}^{p} \hat{\beta}_j x_{ij}$, which would need to be checked for all facilities.

A secondary solution to the problems that arise with using the utilization rate, drawn amount, or undrawn amount in a model with LEQ as the dependent variable is to simply not include those explanatory variables. However, it can be easily demonstrated that by using LEQ as our response variable we cannot avoid including these in the model . . . .

The overwhelming take away from a statistical and econometric viewpoint is that it can be very difficult to produce a clean and easily interpretable model using LEQ, CCF, or UCF values as a response variable to model EAD for at least three key reasons.

1. Firstly, for LEQ at least, the inherent instability of the LEQ values is undesirable. Small changes in the utilization rate can lead to drastic changes in the value of the LEQ, especially when the utilization rate is close to one. Furthermore, the LEQ values are supposed to exist on the unit interval. In reality they exist on the entire real line since they asymptote to plus and minus infinity. This is usually handled by using censored
LEQ values and applying OLS. It is quite easy to demonstrate that applying traditional OLS to censored data produces both biased and inconsistent estimators.

2. Secondly, as we continue to pursue more and more complicated models that take in additionally explanatory variables, such as utilization rate, dollar drawn or dollar undrawn for instance, we would have to continue to use these contrived predetermined transformations in order to maintain the monotonic relationship between predicted EAD and the current dollar drawn amount. This necessarily precludes us from investigating a richer class of transformations that may more appropriately capture the true relationships.

3. Finally, even if we do use the transformations outlined above for the utilization rate and the dollar undrawn amount, the direction of the relationship between predicted EAD and the dollar drawn amount is not guaranteed to be positive, i.e. monotonically increasing as we expect it should be. It will in fact be dictated by the estimated regression coefficients and the values of the other explanatory variables included in the model.

One particular solution is to instead directly use dollar EAD as the response variable or possibly some transformation such as log dollar EAD as the response variable. This avoids the instability issue with LEQ values as well as allows for us to enter any and all explanatory variables or any particular transformation of the explanatory variables with a clear interpretation of how those variables impact EAD. For instance consider the following regression set up,

\[
\log (EAD_i) = \beta_0 + \beta_1 \log (C_i - D_i) + \beta_2 \log (C_i) + \sum_{j=3}^{p} \beta_j x_{ij} + \epsilon_i
\]

where the explanatory variable \( DU_i = C_i - D_i \) is equal to the dollar undrawn amount and \( x_{i3}, \ldots, x_{ip} \) represent other potential explanatory variables to be included in the model. Such a model is often referred to as a Log-Log regression, since both the response variable and at least some of the explanatory variables are used on the log scale. A particular accessible reference text for such a model is (ref: Gujarti). One of the concerns in the past with modeling EAD directly is the large variance of EAD in many loan portfolios (ref: ??). As we will see in Section 3 the data set we will use for our model comparisons has EAD values ranging from xx to yy. By modeling EAD on the log scale we can avoid some of the large variation in the value. Furthermore, the log transformation of dollar EAD is indicated based upon a Box Cox transformation analysis. It is also quite easy to show, as in (ref: Gujarti), that the regression coefficients of the explanatory variables that have been logged represent elasticities. That is, for a one percent change in the dollar undrawn amount, say, dollar EAD will respond by a \( \beta_1 \) percent change.

\[
\frac{d \log (EAD)}{dEAD} = \beta_1 \frac{d \log (DU)}{dDU}
\]

This is the exact mathematical definition for the elasticity between EAD and the dollar undrawn amount. Similar calculations follow for EAD and the commitment level. Such a model has a nice econometric and business interpretation. For instance, end users can easily obtain the predicted percent change in EAD for a one percent change in the dollar commitment amount. This can help to provide guidance to end users when determining a commitment level for a given set of facility characteristics.

3 Data for Empirical Analysis

Description of Commercial EAD data

4 Modeling Comparison

In this section we compare the results of modeling EAD through LEQ, UCF, and CCF with modeling it directly using a log log model in equation (6). Below we present the regression summary results from just such a model based upon the Commercial Credit Risk data set. We began by simply regressing log dollar EAD on log dollar undrawn and log dollar commitment.

Notice that all the \( p \)–values are numerically zero indicating that the estimated regression coefficients are statistically different from zero. Furthermore, the adjusted \( R^2 \) value of 83.99% is quite large. To further corroborate the validity
Table 1: The adjusted $R^2$ value is equal to 83.99% and the p-value of the $F$ statistic is numerically zero.

of using the log transformation of the response variable we consider a so called semi-log model of the following specification,

$$EAD_i = \beta_0 + \beta_1 + \log(DU_i) + \beta_2 \log(C_i) + \beta_3 \log(ORR_i) + \beta_4 \log(Age) + \beta_5 \log(DR_i) + \beta_6 Committed\ Facility$$

where for the $i$th facility $DU_i$ is the dollar undrawn, $C_i$ is the dollar commitment amount, $ORR_i$ is the obligor risk rating (an internal scorecard risk rating), $DR_i$ is the overall Bank default rate, which serves as proxy to calculate downturn EAD under stressed economic conditions, and finally the committed facility indicator variable, which represents whether or not the particular facility is a committed or advised line of credit. It is usually the case that committed facilities, which are legally binding, have larger LEQ and EAD values in comparison to advised lines of credit (ref:????). From this semi–log model we performed a Box Cox transformation analysis (ref:???) in order to determine the most appropriate transformation for the response variable given this model specification. Please refer to Figure 3. From Figure 3 we see that the estimated value of $\lambda$ is fairly close to zero, which implies that a logarithmic transformation is in fact appropriate for the response variable.

Therefore, we consider a Log–Log model of the form of Equation (6) with the additional explanatory variables described above. Similar comments apply here as above. The $p$–values of all explanatory variables are numerically zero and the adjusted $R^2$ value actually increased slightly in comparison to the base model.

Also note that the signs of the estimated regression coefficients make business sense. Starting with the log dollar undrawn explanatory variable we see that the sign is negative. That is, as the undrawn amount increases we expect EAD to decrease. Or as the undrawn amount decreases predicted EAD will increase. This makes business sense since as the obligor draws down on the credit line we expect EAD to increase. Thus, there is an inverse relationship between dollar undrawn and EAD as we would expect there to be. With respect to dollar commitment level the relationship with EAD is positive as it should be.

Obligor risk rating (ORR) also has an intuitive sign as well. It is important to keep in mind that for the commercial model an ORR rating of 1 is the best rating and 10 is an obligor in default. It is widely believed that LEQ values and the ORR have an inverse relationship. This is because obligors with a relative worse rating usually have additional constraints placed on the loan in order to prevent excessive draw down, whereas, obligors with a relatively good rating, i.e. the Bank’s best borrowers, do not have such covenants placed on the loan. Thus, LEQ and ORR have an inverse relationship. Since EAD and LEQ are positively related it makes intuitive business sense that ORR and EAD also have an inverse relationship. For instance as the ORR increases, i.e. as the obligor becomes more and more of a credit risk, we expect EAD to decrease due to the additional restrictions and controls placed on the facility.

The overall Bank default rate is included in order to obtain stressed or downturn EAD predictions. Notice that the sign for the overall Bank default rate is positive. That is, as the overall Bank default rate increases we expect EAD to increase as well. Furthermore, note that the default rate serves us by provided a method by which we can calculate downturn EAD for the purposes of stress testing.

|                  | Estimate | Std. Error | t value | Pr(>|t|) |
|------------------|----------|------------|---------|----------|
| Intercept        | -0.2298  | 0.0278     | -8.27   | 0.0000   |
| log(Dollar Undrawn) | -0.2591 | 0.0034     | -76.15  | 0.0000   |
| log(Dollar Commitment) | 1.2038 | 0.0040     | 302.86  | 0.0000   |

Table 2: The adjusted $R^2$ value is equal to 84.94% and the p-value of the $F$ statistic is numerically zero.
Finally, the Committed Indicator variable captures the categorical variable for whether or not the facility is a committed facility or an advised facility. Committed facilities are legally binding. It is more difficult for the Bank to quickly decrease the commitment amount if a credit event is potential. However, with advised lines of credit the Bank has more flexibility. Therefore, the change from the base case of an advised facility to a committed facility will increase the EAD as we expect it to and the positive sign of the regression coefficient validates this.

In order to test for the presence of multicollinearity we present the variance inflation (VIF) factors for the above regression model. As a rule of thumb, the literature usually suggests that variance inflation factors in excess of ten may be cause for concern. The largest VIF value is 2.7231. Thus, there does not appear to be a strong presence of multicollinearity.

However, I also explored a QQ-plot for the residuals of such a model and it appeared that the Normality assumption was not justifiable. In particular, it appeared that the residuals exhibited a heavy tail nature. Bear in mind that the traditional OLS estimators are still unbiased and moreover consistent estimators for the true unknown population parameter values even in the presence of heavy tailed errors. However, for the purposes of prediction of the response variable the standard error associated with such a prediction will be unduly, i.e. incorrectly, smaller than it actually should be. Furthermore, the prediction interval will also be incorrectly smaller than it otherwise should be had we taken account of the heavy tail nature of the residuals. Some of this may be overcome though with more sophisticated regression techniques such as quantile regression or generalized linear modelling where we assume the errors follow a Student t distribution instead of a Normal. Or a particularly simple solution may in fact be to use

Figure 3: The estimated value of $\lambda$ is close zero, which indicates a log transformation of dollar EAD is most appropriate.
<table>
<thead>
<tr>
<th>Variable</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>log(Dollar Undrawn)</td>
<td>2.613206</td>
</tr>
<tr>
<td>log(Dollar Commitment)</td>
<td>2.723081</td>
</tr>
<tr>
<td>log(Obligor Risk Rating)</td>
<td>1.170190</td>
</tr>
<tr>
<td>log(Age)</td>
<td>1.027381</td>
</tr>
<tr>
<td>log(Default Rate)</td>
<td>1.105337</td>
</tr>
<tr>
<td>Committed Indicator</td>
<td>1.139755</td>
</tr>
</tbody>
</table>

weighted least squares and or use heteroscedastic robust standard errors. Further investigation of these approaches is ongoing.

References


