UMEÅ UNIVERSITET

Master's Thesis -VT17

Modeling credit risk for an SME loan portfolio: An Error Correction Model approach

Jonathan Lindgren

June 14, 2017

Supervisors:
Niklas Lundström, Umeå University
Bujar Huskaj, the Bank

Examiner:
Markus Ådahl, Umeå University

Master's Thesis, 30 Credits
Industrial Engineering and Management, M.Sc.
Department of Mathematics and Mathematical Statistics
Umeå University

Abstract[en]

Since the global financial crisis of 2008, several big regulations have been implemented to assure that banks follow sound risk management. Among these are the Basel II Accords that implement capital requirements for credit risk. The core measures of credit risk evaluation are the Probability of Default and Loss Given Default. The Basel II Advanced Internal-Based-Rating Approach allows banks to model these measures for individual portfolios and make their own evaluations. This thesis, in compliance with the Advanced Internal-Based-rating approach, evaluates the use of an Error Correction Model when modeling the Probability of Default. A model proven to be strong in stress testing. Furthermore, a Loss Given Default function is implemented that ties Probability of Default and Loss Given Default to systematic risk.

The Error Correction Model is implemented on an SME portfolio from one of the "big four" banks in Sweden. The model is evaluated and stress tested with the European Banking Authority's 2016 stress test scenario and analyzed, with promising results.

Abstract[sv]

Sedan den globala finanskrisen 2008 har flera stora regelverk införts för att säkerställa att banker hanterar risker på sunt sätt. Bland dessa regelverk är Basel II som infört kapitalkrav för kreditrisk som baseras på Sannolikhet för Fallissemang och Förlust Givet Fallissemang. Basel II Advanced Internal-Based Approach ger banker möjligheten att skatta dessa riskmått för enskilda portföljer och göra interna kreditriskvärderingar. I överensstämmelse med Advanced Internal-Based-rating undersöker denna uppsats användningen av en Error Correction Model för modellering av Sannolikhet för Fallissemang. En modell som visat sin styrka inom stresstestning. Vidare implementeras en funktion för Förlust Givet Fallissemang som binder samman Sannolikhet för Fallissemang och Förlust Givet Fallissemang med systematisk risk.

Error Correction Modellen modellerar Sannolikhet för Fallissemang av en SME-portfölj från en av de "fyra stora" bankerna i Sverige. Modellen utvärderas och stresstestas med Europeiska Bankmyndighetens stresstestscenario 2016 och analyseras, med lovande resultat.

Acknowledgements

For their aid and encouragement throughout this thesis and assistance to make this thesis both enjoyable and factual, I would like to gratitude:

To Krister Ahlersten for our discussions and giving me the opportunity to write my thesis at your office, a place that came to be my first job after graduation; to Bujar Huskaj, for your patient guidance and insight; to Niklas Lundström for your encouragement and support throughout my thesis; to Daniel Lindgren for reviewing my paper and your invaluable feedback.

Writing a thesis on my own has proven to be far more difficult then I could imagine, in times ideas just circulated and it felt like I got nowhere. Without you, acting as a sounding board and your guidance throughout, this thesis would not have been what it is today, and for this, I am very grateful.

Contents

1 Introduction

This thesis is written at one of the four major banks in Sweden. Hereinafter referred to as the Bank.

1.1 Background

Financial institutes such as banks, which frequently deal with lending, will always have losses of interest and principal. Given a large portfolio, some borrowers will default on their payments. These losses, referred to as credit risk, vary over time depending on the economic situation of the country.

The Bank of International Settlements define credit risk as "*...the potential that a bank borrower or counterparty will fail to meet its obligations in accordance with agreed terms*". The core components of credit risk evaluation is Probability of Default (PD), Loss Given Default (LGD) and Exposure at Default (EaD). Their relationship is described by the equation:

$$EL = PD \cdot LGD \cdot EaD,$$

where PD is the estimated frequency of default on a one-year time horizon, EaD is an exposure such as a bank loan and LGD is the percentage of the exposure the bank expects to lose on the condition of default.

Credit risk is a central component of a bank's risk management that accounts for up to 80% of a bank's total risk exposure (Assouan 2012). Therefore, it is essential to have models that reflect reality. This would not be a problem if we had hundreds of years of data since time series models, for example, could find a strong relationship between credit risk and macroeconomic variables. However, in reality, a bank may have 30–40 years of data and when we look at specific time periods, such as in downturn, we would be happy to have 10 years of data. Current models become inaccurate when looking at such scenarios and therefore valuing loans and financial instruments become problematic.

1.2 The Basel Committee on Banking Supervision

To assure that banks take appropriate measures regarding risk management the Bank of International Settlements created the Basel Committee after the 1970s financial crisis which destabilized the world's financial markets. The goal of the Basel Committee is to enhance financial stability by improving the quality of banking supervision all over the world. The standards created by the Basel Committee are called the Basel Accords and are implemented as laws in countries all over the world (Bank of international settlements 2016).

In June 2004, the Basel Committee issued the Revised Framework on International Convergence of Capital Measurement and Capital Standards, also known as Basel II. The refined standards define two credit risk measures: Expected Loss and Unexpected Loss, which are shown in Figure 1. Expected loss is the average level of credit loss a bank can expect. Unexpected Loss is the tail risk of a portfolio and is defined as the losses that are higher than the level of Expected Loss. These losses can not be covered by interest rates and risk premium alone. Therefore, banks are required to hold capital in case of unexpected losses (Coen 2000).

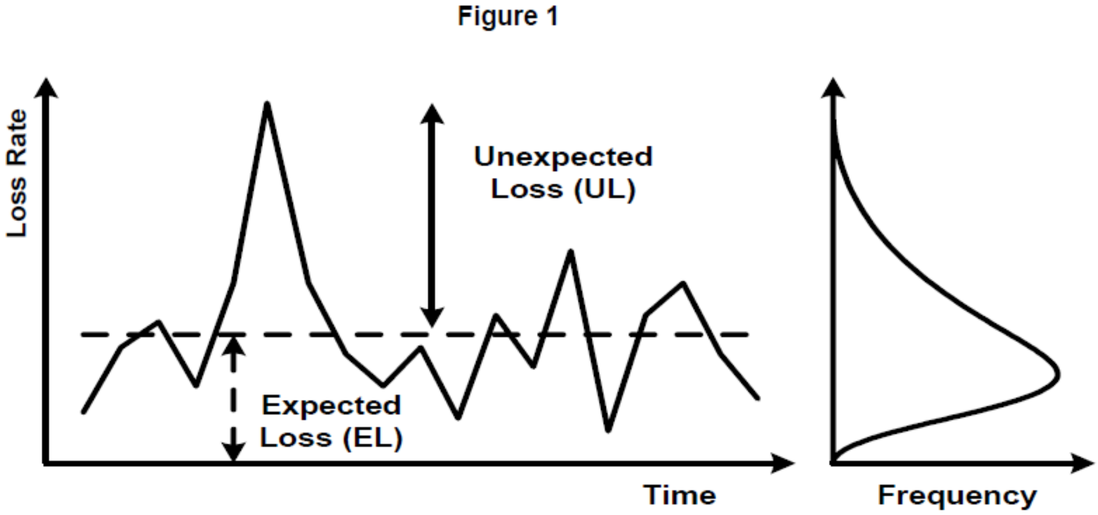

Figure 1: Relation between Expected Loss and Unexpected Loss (Bank of international settlements 2005)

As smaller banks may not have the capability of calculating their own credit risk models the Basel II introduced the standard approach. This allows banks to calculate credit risk from standardized tables of different risk sectors. This method is simple and yields inaccurate assessments of individual portfolios. Therefore Basel II, along with the standardized approach, introduced the Internal Rating-based Approach as a means to allow banks to make their own models of credit risk.

Internal Rating-based Approach

The revised framework introduced the internal ratings-based approach (IRB), which allows banks to create their own models for borrower's PD. Furthermore, banks can apply for advanced IRB to create their own models for the borrower's PD, LGD and EaD. IRB has the advantage of letting banks create models for the actual portfolios instead of utilizing generalized models and thus getting better credit risk evaluations.

Models created under the Advanced IRB approach are required to be *portfolio invariant*, i.e the capital required to offset Unexpected Loss should only depend on the risk of the loan and not the portfolio it is added to. This is achieved by a portfolio that has a large number of small exposures and, by the law of large numbers, the idiosyncratic risks from each individual exposure cancel out one-another. As the portfolio is unaffected by the idiosyncratic risks it can be modeled for systematic risk alone which is referred to as an *asymptotic single risk factor model* (Bank of international settlements 2005).

With the introduction of IRB and the asymptotic single risk factor model methodology banks are allowed to use time series modeling for calculations of credit risk measures on their portfolios and this is the type of methodology implemented in the thesis.

1.3 The European Banking Authority

The European Banking Authority (EBA) is an independent EU Authority which works towards ensuring consistent regulation and supervision within the European banking sector. While their objective is to maintain financial stability within the EU, their main task is to harmonise the prudential rules of the financial institutions in the EU. A part of these tasks are creating stress test scenarios which banks are enforced to conduct to assure they can handle potential stressful financial events (The European Banking Authority 2016*a*).

EBA's 2016 macroeconomic stress test scenario is divided into two scenarios: the baseline scenario which represents a normal economic situation and the adverse scenario which stresses the economic situation as if an economic crash occured (The European Banking Authority 2016*b*). This thesis implements the EBA's 2016 macroeconomic stress test scenario to ensure both a realistic stress test scenario and that the results are commensurable to the Bank's own evaluations.

1.4 Purpose

The purpose of this article is to evaluate the use of an Error Correction Model (ECM) when modeling credit risk. An important aspect of credit risk models is their ability to make realistic predictions as this directly impacts the capital required to offset credit risk. The use of an ECM was introduced by Assouan (2012) which found strong predictability of the model when stress tested. The implementation of an ECM for a credit risk portfolio is an approach that, to my knowledge, has only been tested by Assouan (2012) before. This thesis will further test its power on the Bank's SME portfolio.

The ECM will be used to predict the portfolio's Probability of Default and the strength of the forecast will be compared with a linear regression. Furthermore, the Probability of Default is implemented in a LGD-function derived by Frye & Jacobs Jr (2012) that ties Loss Given Default to Probability of Default with systematic risk.

1.5 Limitations

Since there is a restricted time frame for this thesis the macroeconomic data collection is limited to the macroeconomic variables implemented in the EBA 2016 stress scenario in addition of an indicator of the Swedish market state called State of Economy (Konjukturbarometern). The EBA stress scenario has the advantage of making the resulting forecast realistic and commensurable to the Bank's internal 2016 stress test evaluation.

1.6 Data

The data used in this article consists of two parts. The financial data, namely Default rate and LGD, is historical data, on quarterly basis, based on the Bank's SME portfolio from 2006Q1 to 2015Q4. The macroeconomic data is based on Sweden and is taken from Statistics Sweden and The Organization for Economic Co-operation and Development and are available to the public. The data used can be seen in Table 1. Due to the portfolio data being confidential some data is manipulated and the results can not be presented.

Table 1: Data used for time series modeling

Group	**Variables**	**Level / Growth**
Macroeconomic variables	Real GDP	Growth
	State of Economy	Level
	10 year Government Bond	Level
	Inflation	Growth
	Property Price	Growth
	Unemployment Rate	Level
	OMX30	Growth
Portfolio variables	Loss Given Default	Level
	Observed Default Rate	Level

Observed Default Rate (ODR) is the historical default rate of a portfolio used to estimate PD such as $PD = E[ODR]$.

1.7 Outline

In this thesis we will analyze the usage of an ECM and a Loss Given Default function as a basis for modeling credit risk at a major bank in Sweden. The theory behind the models along with data handling and stress testing is explained in Chapter 2. Chapter 3 explains how we apply the theory for modeling PD and create guidelines for our models. In Chapter 4 the results are presented which are then analyzed and discussed in Chapter 5.

2 Theory

2.1 Linear Regression and Ordinary Least Squares

The Linear Regression Model is given by:

$$Y_t = \alpha + X_t \cdot \beta + \varepsilon_t, \tag{1}$$

where Y_t is the dependent variable to be estimated, X_t are explanatory variables such as macroeconomic variables, α is a coefficient called the intercept, β are coefficients and ε_t is the residual term. The variables have the form:

$$Y = \begin{pmatrix} y_1 \\ \vdots \\ y_T \end{pmatrix}, X = \begin{pmatrix} x_{1,1} & \cdots & x_{1,k} \\ \vdots & \ddots & \vdots \\ x_{T,1} & \cdots & x_{T,k} \end{pmatrix}, \beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_T \end{pmatrix}, \tag{2}$$

where T is the number of observations and k is the number of macroeconomic variables. The coefficients α and β are estimated through Ordinary Least Squares (OLS) such as:

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (X_*' \cdot X_*)^{-1} \cdot X_*' \cdot Y, \tag{3}$$

where X_* includes a one-vector to estimate the α coefficient such as

$$X_* = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{T,1} & \cdots & x_{T,k} \end{pmatrix}. \tag{4}$$

2.2 Error Correction Model

An ECM has the advantage of taking both short-term and long-term effects into account and makes use of cointegrated variables which, as section 2.3 explains, usually requires differentiation where information in the data is lost. The ECM model used in this thesis is the Engel and Granger 2-step ECM approach (Assouan 2012). The first step is to estimate the long-term relationship which makes use of a cointegrated relationship. The residual, or error, from the long-term relationship is then implemented as an error-correction term in the second step which is called the short-term relationship. To describe the model a standard econometric equation such as Equation (1) is first specified between the cointegrated variables seen below:

$$Y_t = \alpha + X_t \cdot \beta + \varepsilon_t. \tag{5}$$

An OLS-estimate on Equation (5) yields the residual, ε_t, which is implemented into the short-term relationship, described in Equation (6), with lag $T-1$. The short-term relationship is defined as

$$\Delta Y_t = \theta \cdot \hat{\varepsilon}_{t-1} + \gamma \cdot \Delta X_t + \mu \cdot Z_t + \nu_t, \tag{6}$$

where Z_t is a matrix of optional stationary explanatory variables and θ, γ and μ are coefficients estimated with OLS.

To be able to stress test the model with an out-of-sample forecast we substitute ε_t with the error-correction term $(Y_{t-1} - \hat{\alpha} - X_{t-1} \cdot \hat{\beta})$ yielding

$$\Delta Y_t = \theta \cdot (Y_{t-1} - \hat{\alpha} - X_{t-1} \cdot \hat{\beta})) + \gamma \cdot \Delta X_t + \mu \cdot Z_t + \nu_t. \tag{7}$$

For a two-step ECM to be valid it is required that:
- Z_t contains only stationary variables
- θ is negative
- Y_t and X_t are integrated of the same order
- ε_t is stationary.

(Assouan 2012)

2.3 Stationarity, Order of Integration and Cointegration

Order of integration, stationarity and cointegration are all essential to time series modeling. Below is an explanation to each one of them.

Stationarity

Stationarity is a property of a variable or process that has no moments depending on time, i.e. its mean, variance and autocorrelation structure do not change over time. A variable or process that is stationary is said to have an order of integration of 0, or I(0) (Kennedy 1998, p. 268-269, 409). Stationarity is important for time series analysis due to the fact that a non-stationary variable will trend, making time series models unreliable unless accounted for. Since stationarity and cointegration are fundamental to ECM this study will use three tests for stationarity; the Augmented Dickey-Fuller Test, the Phillips-Perron Test and the Dickey-Fuller GLS Test.

Augmented Dickey-Fuller Test (ADF)

The ADF test procedure tests if a variable follows a unit-root process and thereby is a non-stationary process. The test uses an autoregressive process of order p, denoted AR(p), to handle serial correlation parametrically. The following presentation is based on Kennedy (1998, p. 275-285).

A series is assumed to have the following form:

$$\Delta y_t = \alpha + \beta \cdot t + \gamma \cdot y_{t-1} + \delta_1 \cdot \Delta y_{t-1} + \cdots + \delta_{p-1} \cdot \Delta y_{t-p+1} + \varepsilon_t, \tag{8}$$

where α is the intercept, β is a coefficient on time trend, p is the lag order of the test and ε is an independent and identically distributed zero-mean error term. The test statistic is calculated as:

$$tstat = \frac{\hat{\gamma}}{s.e(\hat{\gamma})}, \tag{9}$$

where $s.e(\hat{\alpha})$ is the coefficient standard error.

The null hypothesis is that y_t contains a unit root and thereby is a non-stationary process, $H_0 : \gamma = 1$. The alternative is that y_t follows a random walk and is a stationary process, $H_A : \gamma < 1$. Setting $\tau = \gamma - 1$ yields the unit root hypothesis:

$$H_0 : \tau = 0 \text{ against } H_A : \tau < 0.$$

The model is fitted by the OLS method and t-tested for the null hypothesis at 95% significance level.

Phillips-Perron Test (PP)

The PP test procedure tests an AR(1) process for unit root non-parametrically by using a test-statistic made to take serial correlation into account rather than modeling the AR process for serial correlation such as the ADF Test. The following presentation is based on Phillips & Perron (1988).

The series is assumed to have the following form:

$$\Delta y_t = \alpha + \beta \cdot t + \gamma \cdot y_{t-1} + \varepsilon_t, \tag{10}$$

where α is the intercept, β is a coefficient on time trend and ε is an independent and identically distributed zero-mean error term.

The test-statistic is calculated as:

$$Z_\tau = t_\alpha \cdot \sqrt{\frac{\gamma_0}{f_0}} - \frac{T \cdot (f_0 - \gamma_0) \cdot s.e(\hat{\alpha})}{2 \cdot \sqrt{f_0} \cdot s}, \tag{11}$$

where $\hat{\alpha}$ is the OLS estimate of the AR(1) process, t_α is the t-ratio of α, s is the standard error of the test regression and f_0 is an estimator of the residual spectrum at frequency zero. $\hat{\gamma}_0$ is calculated as:

$$\gamma_0 = \frac{(T-k) \cdot s^2}{T}, \tag{12}$$

where k is the number of regressors.

Dickey-Fuller GLS (ERS)

The ERS test estimates stationarity of a process by detrending the explanatory variable, y, and running an ADF test on the detrended data. The method is based on Elliott, Rothenberg & Stock (1992).

The process is based on the model:

$$y_t = d_t + u_t, \tag{13}$$

$$u_t = \alpha \cdot u_{t-1} + \nu_t, \tag{14}$$

where d_t is a deterministic component and ν_t is an unobserved stationary zero-mean error process. First the detrending is defined as:

$$d(y_t|a) = \begin{cases} y_t & \text{if } t = 1 \\ y_t - a \cdot y_{t-1} & \text{if } t > 1 \end{cases} \tag{15}$$

Regression of the quasi-differenced data $d(y_t|a)$ on the quasi-differenced d_t is considered such as:

$$d(y_t|a) = d_t \cdot \alpha(a) + \eta_t, \tag{16}$$

where $\alpha(a)$ is the coefficient from the OLS estimate of the regression. The value of a used in this thesis is the built-in detrender of Eviews (2017), $a = \bar{a}$, where

$$\bar{a} = \begin{cases} 1 - 7/T & \text{if } x_t = \{1\} \\ 1 - 13.5/T & \text{if } x_t = \{1, t\}. \end{cases} \tag{17}$$

The GLS detrended data, ${y_t}^{\mathrm{d}}$, with $\bar{a}$ can now be defined as:

$${y_t}^{\mathrm{d}} = y_t - d_t \cdot \alpha(\bar{a}). \tag{18}$$

The test is conducted on the ADF test equation (8) by substituting y_t with the detrended data ${y_t}^{\mathrm{d}}$ such as:

$$\Delta {y_t}^{\mathrm{d}} = \gamma \cdot {y_{t-1}}^{\mathrm{d}} + \delta_1 \cdot \Delta {y_{t-1}}^{\mathrm{d}} + \cdots + \delta_{p-1} \cdot \Delta {y_{t-p+1}}^{\mathrm{d}} + \varepsilon_t, \tag{19}$$

and tested for the null-hypothesis:

$$H_0 : \gamma = 1 \text{ against } H_A : |\gamma| < 1.$$

The test statistics used is the t-ratio:

$$y_\gamma = \frac{\gamma}{s.e(\gamma)}. \tag{20}$$

Order of integration

The concept of order of integration, denoted $I(d)$, is the number of differentiations needed for a variable to become stationary. Engle & Granger (1987) defines order of integration as "*A series with no deterministic component which has a stationary,* *invertible, ARMA representation after differencing d times, is said to be integrated of order d, denoted $x_t \sim I(d)$"*. Differentiation stabilizes the mean of a time series by removing changes in the level of a time series, and thus eliminating trend and seasonality. But, as data is differentiated, valuable information about the dataset is lost and should therefore be avoided if possible.

Cointegration
Variables that show non-stationarity can together nullify the effect of non-stationarity. If tests for stationarity on two variables independently show *I(1)* but together show *I(0)* they are said to be cointegrated (Kennedy 1998, p. 268-270).

Kennedy (1998) defines cointegration as *"co-integrated of order d, b, denoted $x_t \sim CI(d, b)$, if (i) all components of x_t are $I(d)$; (ii) there exists a vector $\alpha(\neq 0)$ so that $z_t = \alpha' x_t \sim I(d-b), b > 0$. The vector α is called the co-integrating vector"*. The implication is that (n) non-stationary variables showing the same $I(d)$ can nullify trends and thereby becoming $I(0)$. Transformation of data by differentiation is then not needed. The test used in this thesis for cointegration test is the Residual Test.

Residual Test

A collection of time series variables are said to be cointegrated if there exists a linear combination of the variables such as:

$$Y_t = \alpha + \beta \cdot X_t + \varepsilon_t, \tag{21}$$

where ε_t is stationary and both X_t and Y_t have an integration order of *I(d)*.

When testing ε_t for stationarity each of the stationarity tests explained above can be used. In this thesis the Phillip-Perron Test is used, as recommended by Assouan (2012).

2.4 Systematic LGD

Commonly the modeling of PD and LGD has been done separately with historical data and macro variables. This is due to the fact that finding a systematic relation between PD and LGD has proven difficult. Frye & Jacobs Jr (2012) and Altman & Karlin (2010) show that PD and LGD do in fact vary systematically in stressful scenarios. From this Frye & Jacobs Jr (2012) have calculated a relationship that depends on distributions of loss and default.

The model makes four assumptions. The first assumption is that the asymptotic distributions of default and loss are comonotonic. While this is technical, the implication is that the default rate and loss rate take the same quantile, q, within their respective distribution such as:

$$CDF_{loss}[\text{loss rate}] = CDF_{DR}[\text{default rate}] = q, \tag{22}$$

where CDF_{loss} is the cumulative distribution function of the loss rate and CDF_{DR} is the cumulative distribution function of the default rate.

The product of default rate and LGD rate equals loss rate. This gives us the formula:

$$LGD = \frac{CDF_{loss}^{-1}[q]}{CDF_{DR}^{-1}[q]} = \frac{CDF_{loss}^{-1}[CDF_{DR}[DR]]}{DR}. \tag{23}$$

The second assumption is that the credit loss and default distributions have two parameters in the asymptotic portfolio. Frye & Jacobs Jr (2012) tested three two-parameter distributions, namely Vasicek, beta and log-normal. The Vasicek distribution limits the default rate to be less than 100% and will be used in this report. This leads to Frey & Jacob's third assumption: The loss and default rate follows the Vasicek distribution.

The fourth assumption states that the value of ρ in CDF_{loss}, see equation (28), is equal to the value of ρ in CDF_{DR}. This leads to our final equation by substituting the equations (27) and (28) into equation (23):

$$LGD = \frac{\Phi\left[\Phi^{-1}[DR] - \frac{\Phi^{-1}[PD]-\Phi^{-1}[EL]}{\sqrt{1-\rho}}\right]}{DR} = \frac{\Phi\left[\Phi^{-1}[DR]\right] - k]}{DR}, \tag{24}$$

where

$$k = \frac{\Phi^{-1}[PD] - \Phi^{-1}[EL]}{\sqrt{1-\rho}}. \tag{25}$$

The variable k, depending on PD, EL and ρ, is referred to as the LGD risk index.

2.4.1 Vasicek Distribution

The Vasicek Distribution is derived from portfolio losses which shows that the distribution, unlike the normal distribution, has properties of high skewness and thick tails. This is due to the fact that defaults are not independent (Vasicek 2002).

The Vasicek distribution, given $0 < x < 1$, has the following density functions (Frye & Jacobs Jr 2012):

$$PDF[x] = \frac{\sqrt{1-\rho}}{\rho} \cdot \frac{\phi[(\Phi^{-1}[EL] - \sqrt{1-\rho} \cdot \Phi^{-1}[x])/\sqrt{\rho}]}{\phi[\Phi^{-1}[x]]}, \tag{26}$$

$$CDF[x] = \Phi\left[\frac{\sqrt{1-\rho} \cdot \Phi^{-1}[x] - \Phi^{-1}[EL]}{\sqrt{\rho}}\right], \tag{27}$$

$$CDF^{-1}[q] = \Phi\left[\frac{\Phi^{-1}[EL] + \sqrt{\rho} \cdot \Phi^{-1}[q]}{\sqrt{1-\rho}}\right], \tag{28}$$

where ϕ is the standard normal probability density function and Φ is the standard normal cumulative distribution function.

2.5 Forecasting

Dynamic forecasting is used in this thesis as both the Linear Regression Model and ECM are regressed on ΔODR.

Dynamic forecast model

Models that depend on a lagged value of the explanatory variable Y requires dynamic forecasting to make an unbiased forecast of the model.

Given the model:

$$Y_t = \alpha + \gamma \cdot Y_{t-p} + \beta \cdot X_t + \varepsilon_t, \tag{29}$$

where p is the number of lags and the coefficients $\hat{\alpha}$, $\hat{\gamma}$ and $\hat{\beta}$ are OLS estimates on T periods.

The forecast of $\hat{Y}_{T+i}$ is calculated as:

$$\hat{Y}_{T+i} = \hat{\alpha} + \hat{\gamma} \cdot \hat{Y}_{T-p+i} + \hat{\beta} \cdot X_{T+i}, \tag{30}$$

for $i = \{1, 2, 3, \ldots, n\}$.

Mean Absolute Error

To estimate how well the models predict PD we calculate the forecast error through the mean absolute error

$$MAE = \frac{\sum_{t=1}^{N} |E_t|}{N}, \tag{31}$$

where $E_t = Y_t - F_t$ is the difference between the historical value and forecast value.

3 Method

As this thesis revolves around modeling PD, rather than the implementation of a model describing PD, this section will focus on the decisions considering tests and implementation of the model.

3.1 Software usage

During this thesis, two software have been used to create the models. Eviews has been used to conduct data testing and time series modeling and Matlab has been used to forecast the models and analyzing the results.

3.2 Data handling

Before modeling PD and LGD the data is tested for trends, stationarity and integration order. Finding trends in the data series is done visually by inspecting the plotted data. If a trend can be noted in the data series it has to be tested for trend stationarity, i.e. if the trend can be removed and thereby making the data series stationary. This is done by including the trend term in the stationarity tests mentioned in Section 2.3. The parameter selection for stationarity tests can be seen in Table 2 below.

Table 2: Intercepts and trends for the macroeconomic variables

Variable	**Intercept**	**Trend**
ODR	Yes	No
Inflation	Yes	No
Unemployment Rate	Yes	No
10-year Government Bond	Yes	Yes
Real GDP	Yes	No
State of Economy	Yes	No
OMX30	Yes	Yes
Property Price	Yes	Yes

As the table shows the stationarity test for OMX30, 10-year Government Bond and Property Price requires trend testing by including a trend estimator $\beta * t$ in the stationarity test. Data showing *I(2)* were excluded from the models due to low significance after differencing. Table 3 shows the results of the stationarity tests and we see that the results vary depending on which test is used. This is a common occurrence in econometrics and one of the motivations for using several tests. Cases where the tests show different results, the choice of integration is given to the majority e.g. ODR has an integration order of 1 due to ADF and PP tests showing *I(1)*.

Table 3: Integration order of macroeconomic variables and ODR

Variable	**ADF**	**PP**	**ERS**	**Integration order**
ODR	*I(1)*	*I(1)*	*I(0)*	*I(1)*
Inflation	*I(0)*	*I(1)*	*I(0)*	*I(0)*
Unemployment Rate	*I(1)*	*I(1)*	*I(0)*	*I(1)*
10-year Government Bond	*I(1)*	*I(1)*	*I(1)*	*I(1)*
Real GDP	*I(0)*	*I(0)*	*I(0)*	*I(0)*
State of Economy	*I(0)*	*I(1)*	*I(0)*	*I(0)*
OMX30	*I(1)*	*I(1)*	*I(1)*	*I(1)*
Property Price	*I(0)*	*I(0)*	*I(0)*	*I(0)*

From Table 3 we have Unemployment Rate, 10-year Government Bond and OMX30 as possible explanatory variables in the long-term relation of the ECM. The remaining macroeconomic variables including the first difference of long-term relation variables can be used as explanatory variables in the short-term relationship. The regression of ODR will have ΔODR as the dependent variable and the same explanatory variables as the short-term relationship of the ECM.

3.3 Statistical Significance versus Economic Theory

One of the big challenges in modeling time series models is the trade-off between economic theory and statistical significance. While it might make sense intuitively to follow test statistics that show the highest coefficient significance or adjusted coefficient of determination (Adjusted R^2), it is important to consider the economic theory behind the choice of variables. As an example, consider the simple linear regression:

$$PD_t = \beta \cdot \text{Unemployment Rate}_t + \varepsilon_t, \tag{32}$$

where the Probability of Default depends on Unemployment Rate alone.

From economic theory, we consider an increase in unemployment as a sign of a struggling economy. If the coefficient β is negative the model would predict the probability of default to decrease if the unemployment increases. This goes against economic theory even though it might show strong statistical significance. This is a common occurrence when modeling time series with a limited set of data. In the restricted time horizon it finds a relationship between the dependent and explanatory variables, despite the fact that the relationship is flawed. It can be considered as a short-term relationship given the exact scenario tested. But since it is important to find models that reflect reality, and predict the future, we need a model that follows sound economic theory while still maintaining statistical significance. This is not always achievable, but is the focal point of this thesis. From this discussion, and by applying the same logical reasoning to the other macroeconomic variables, we get the following guidelines for the models:

- Unemployment Rate and 10-year Government Bond should have a positive coefficient,
- GDP, OMX30, Inflation, Property Price and State of Economy should have a negative coefficient.

3.4 Modeling approach

The first step in time series modeling is to consider what components to take into consideration when making the model. From the ECM specification described in Section 2.2, we have two major restrictions: a long-term relationship where $I(1)$ macroeconomic variables, X_t, are required and a short-term relationship where the first difference of the long-term variables, ΔX_t, and stationary macroeconomic variables, Z_t, are considered. The Linear Regression Model described in Section 2.1 is only restricted to the use of stationary variables .

Once the structure of the model is decided we consider serial correlation on the dependent variable i.e, the behavior of sequential points in the time series affecting each other. If the time series show serial correlation we can include a lagged value of the dependent variable in the model. This will help strengthen the model and its forecasts.

Some macroeconomic variables have delayed response to economic events. Figure 2 shows Unemployment Rate, Inflation and GDP during the global economic crisis of 2008. As expected, GDP and Inflation are directly impacted by the crisis while Unemployment Rate steadily increases during 2009-2010. If the dependent variable has a direct response to the crisis and Unemployment Rate does not it would show low statistical significance in our model. Therefore, lagged values are considered on the macroeconomic variables to account for delayed economic effects.

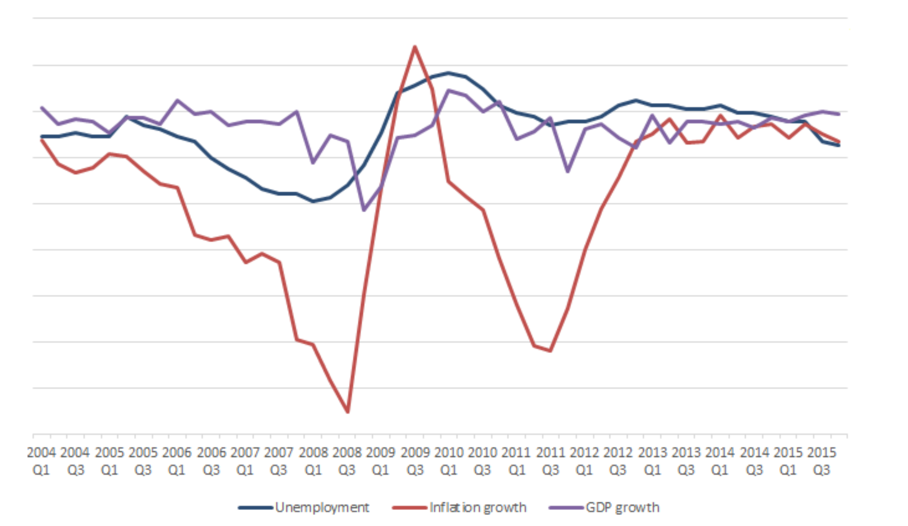

Figure 2: Comparison of Unemployment Rate, Inflation and GDP

Once the final model specification is done the model is tested for statistical significance. Both stationarity tests and coefficient significance are set to a 10% significance level; Models that fall outside this limit are not considered. Furthermore, the coefficients and lags are checked to ensure they follow the guidelines explained in section 3.3. The models are then tested for Adjusted R^2 to find out how well they explain the dependent variable.

Model analysis and prediction

The models that are statistically significant and follow the guidelines mentioned earlier are all viable models for PD within the time frame given by the ODR observations. However, this does not necessarily make them good at predictions which is an important factor of the models. To find out how well the models explain ODR the in-sample forecast is implemented and to see how well they predict future PD an out-of-sample forecast is implemented. The forecasts predict 12 quarters (3 years) i.e:
$i = \{T-12, T-11, T-10, \ldots, T\}$ for the in-sample forecast and
$i = \{T+1, T+2, T+3, \ldots, T+12\}$ for the out-of-sample forecast.

The accuracy of the predictions is analyzed mainly through visual inspection. In the in-sample forecast a good prediction maintains, for each time step, the vertical movements of ODR. In other words the trajectory of PD should mimic the movements of ODR. Furthermore, the in-sample forecast is analyzed through the mean absolute error which calculates the mean distance of PD to ODR. This measure is then compared between the linear regression and the ECM. The out-of-sample forecast is analyzed on the predictions between the linear regression and ECM as there is no ODR to compare with. The predictions should make realistic predictions which, in this stress test, means that the baseline scenario should yield PD comparable to ODR during normal economic conditions and the adverse scenario should yield PD comparable to crisis, such as the financial crisis of 2007-2008.

4 Results

The following chapter shows the models created on the Bank's SME portfolio. This includes an ECM, Linear Regression Model and the LGD-function. Due to confidentiality reasons, some tables do not show numbers and some graphs are censored. To give the reader an idea of how the models act the coefficients are replaced with the signs of the coefficients and the graphs are discussed briefly.

4.1 Stationarity and cointegration results

Before deciding on a model structure the data is tested for integration order. The variables that showed $I(1)$ were differenced and tested again to assure that they were indeed stationary. The results are shown in Table 4.

Table 4: Integration order of macroeconomic variables and ORR

Variable	**ADF**	**PP**	**ERS**	**Integration order**
ODR	*I(1)*	*I(1)*	*I(0)*	*I(1)*
ΔODR	*I(0)*	*I(0)*	*I(0)*	*I(0)*
Inflation	*I(0)*	*I(1)*	*I(0)*	*I(0)*
Unemployment Rate	*I(1)*	*I(1)*	*I(0)*	*I(1)*
Δ Unemployment Rate	*I(0)*	*I(0)*	*I(0)*	*I(0)*
10-year government bond	*I(1)*	*I(1)*	*I(1)*	*I(1)*
Δ 10-year government bond	*I(0)*	*I(0)*	*I(0)*	*I(0)*
Real GDP	*I(0)*	*I(0)*	*I(0)*	*I(0)*
State of Economy	*I(0)*	*I(1)*	*I(0)*	*I(0)*
OMX30	*I(1)*	*I(1)*	*I(1)*	*I(1)*
Δ OMX30	*I(0)*	*I(0)*	*I(0)*	*I(0)*
Property Price	*I(0)*	*I(0)*	*I(0)*	*I(0)*

ODR shows $I(1)$ while its first difference is $I(0)$ which allows us to use ODR as an explanatory variable for the ECM. Unemployment Rate, 10-year Government Bond and OMX30 are the macroeconomic variables that show $I(1)$ and their respective first difference show $I(0)$. Thereby the core of the ECM can be established with explanatory variables and its long-term relationship. Stationary variables ($I(0)$) can be used as explanatory variables in the model to further describe the ODR. For clarity, the data groups that may be used for the long-term relationship are shown in table 5 below.

Table 5: Variables that can be included in the cointegrated relationship

Variable	**Integration Order**
Unemployment Rate	*I(1)*
OMX30	*I(1)*
10-year government bond	*I(1)*

The cointegrated relationship used as the ECM's long-term relationship was tested for variants of the data in Table 5. The long-term relationship that gave the best results were:

Table 6: The long-term relationship

Explanatory variable	**Lag**
Coefficient	—
Unemployment Rate	T-4
OMX30	—
10 year Government Bond	T-5

with an Adjusted $R^2 = 0.65$ and stationary residuals according to the Philips-Perron Test.

4.2 The models

The SME portfolio's ODR has been tested for several models. The resulting models were chosen with the criteria given by the guidelines in Section 3.2, Adjusted R^2 and their predictability in the in-sample forecast and out-of-sample forecast. All coefficients in the models are statistically significant.

The Error Correction Model

The structure of the ECM can be viewed in Table 7. The Unemployment Rate is, as expected, lagged due to showing slow response to economic events while GDP and Inflation both have near direct impact on ODR. The 10-year Government Bond shows delayed impact on ODR which as well is expected from economic theory as an increase in the rates will stress company's finance and in the long-term they may not be able to pay interests and loans. The coefficient of the residual is negative which means that it is indeed *correcting* the long-term relationship. GDP, Inflation and OMX30 all have positive coefficients, which means the model values them as indicators of a good economy and that they follow sound economic theory. The 10-year Government Bond and Unemployment Rate have negative coefficients which further strengthens the model's credibility. However, the State of Economy has a positive sign. This implies that recession would decrease the probability of default which is contradictory. The model has an Adjusted R^2 of 67.9%.

Table 7: The Error Correction Model

Variable	Lag	Coefficient	P-value
Long-term relationship			
Coefficient	—	$C_1 > 0$	***
Unemployment Rate	T-4	$C_2 > 0$	*
OMX30	—	$C_3 < 0$	***
10 year Government Bond	T-5	$C_4 > 0$	*
Short-term relationship			
Long-term residual	T-1	$C_5 < 0$	***
GDP	T-1	$C_6 < 0$	**
Inflation	T-1	$C_7 < 0$	***
State of Economy	T-4	$C_8 > 0$	***

The Linear Regression Model

As ODR is non-stationary the regression is done on ΔODR. The Linear Regression Model can be seen in Table 8 below. The model makes use of GDP, Inflation, ΔOMX30 and ΔODR as explanatory variables. The ΔODR is implemented to account for serial correlation in the explained variable. All coefficients are negative and follow sound economic theory. The lag of GDP and ΔOMX30 shows that the model values them as near direct impact on PD while Inflation has a 5 quarter delay. The model has an Adjusted R^2 of 58.9%.

Table 8: The Linear Regression Model

Variables	Lag	Coefficient	P-value
Constant	—	$C_1 > 0$	**
ΔODR	T-1	$C_2 < 0$	***
GDP	T-1	$C_3 < 0$	***
Inflation	T-5	$C_4 < 0$	***
$\Delta OMX30$	T	$C_5 < 0$	***

4.3 Model prediction

Both the ECM and Linear Regression Model make use of ΔODR and are therefore forecasted with dynamic forecasting as explained in Section 2.5. Due to confidentiality reasons the predicted values cannot be shown.

In-sample Forecast

The in-sample forecast of the ECM and the Linear Regression Model is estimated on 2005Q4 to 2012Q4 and forecasted on 2013Q1 to 2015Q4. To make a fair comparison of the models their predictions are compared with the ODR. This simplifies the comparison of how the models behave and how their trajectories relate to the portfolio's ODR.

The behavior of the ECM follows the ODR's movements rather well in the first two years of the forecast. After 2014Q1 the model drops below the ODR while still maintaining the movements of the ODR but with a greater amplitude. The Linear Regression Model follows the ODR's movements well, and throughout the forecast the PD is slightly above ODR. The ECM and the Linear Regression Model have an Mean Absolute Error of $7.32 \cdot 10^{-4}$ and $6.10 \cdot 10^{-4}$ respectively.

Out-of-sample Forecast

The out-of-sample forecast of both the ECM and the Linear Regression Model are estimated from 2005Q4 to 2015Q4 and forecasted from 2016Q1 to 2018Q4 using the EBA's 2016 stress scenario. The results discussed below are separated by the baseline and adverse scenario.

The baseline scenarios of the Linear Regression Model and the ECM differ considerably. The Linear Regression Model stays close to linear during the entire forecast with small variations over time and at level with the lowest point of the ODR. This is unexpected as the baseline scenario variables vary considerably over the 12 quarters. The ECM increases to levels comparable with ODR during 2011-2012 with a smooth curve along the time periods.

If we compare the adverse scenario between the two models we see that the Linear Regression Model increases to levels almost 50% above ODR during 2008 crisis while the ECM increases to levels comparable with the 2008 crisis. This is in favor of the ECM as the shock on scenario variables are comparable to the 2008 crisis.

4.4 LGD-function

The LGD-function derived by Frye & Jacobs Jr (2012) is implemented on the PD-estimates of the ECM and Linear Regression Model. The Historical LGD shows very small movements along the time horizon. As the historical LGD can not be shown a horizontal line named *Historical LGD* is shown in the graph, with similar behavior to the real historical LGD, to give the reader a sense of the functions predictability. The result can be seen in Figure 3.

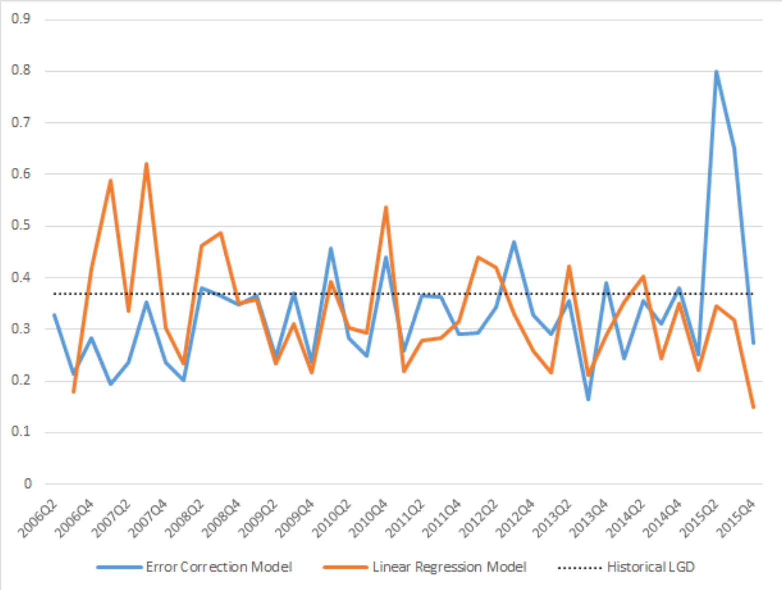

Figure 3: LGD function

As the graph shows, the estimated LGD is very volatile with movements from 20% to 80%. The function was derived on the assumption that both LGD and ODR had correlated movements in stressful scenarios. But as the Portfolio LGD is very stable the estimates turned out to be misleading.

5 Discussion and conclusion

5.1 Review

We have studied the use of time series modeling to estimate Probability of Default and implementation of a Loss Given Default function on a small-to-medium enterprise portfolio. The dataset consists of the portfolios Observed Default Rate and Loss Given Default along with macroeconomic variables restricted to the EBA 2016 stress test scenario. The modeling section consists of an Error Correction Model to estimate Probability of Default as well as a Linear Regression Model on Probability of Default used as a benchmark. Furthermore, both models were stress tested with an in-sample forecast as well as an out-of-sample forecast to evaluate their predictability. The following discussion will discuss the results, if an Error Correction Model can be utilized in credit risk, what went well and what could have been done differently.

5.2 Selection of data

The decision to restrict the macroeconomic variables to the EBA's 2016 stress test scenario is something I, in retrospect, regret. In the beginning it made sense to use the EBA's 2016 stress test scenario as it made the predictions reliable and relatable for the Bank as they implement these stress tests themselves. But as the Error Correction Model has defining restrictions such as stationary variables, non-stationary variables and a cointegrated relationship it set defining restrictions on the models. Among these restrictions are the use of macroeconomic variables at level rather than growth such as OMX30, 10-year Government Bond and Unemployment Rate. As the OMX30 steadily increase over time the model will also continuously decrease the Probability of Default. Therefore it is preferable to use the growth of variables, as growth is very unlikely to continuously increase. However, some variables such as Unemployment Rate are justifiable to use at level since they are restricted between 0% and 100%. The variables mentioned above were all tested integration order on growth, but the results showed that they all would become stationary. This means that there would be no macroeconomic variables of integration order of 1 and thereby no cointegrated relationship to use as the long-term relationship. There would be no Error Correction Model. What should have been done is to collect a large number of macroeconomic variables which would have given greater flexibility when modeling credit risk using the Error Correction Model.

5.3 Model evaluation

When creating a model it is easy to get excited about a good coefficient of determination or a good forecast. But it is as important to assess the structure of the model and see if the model is indeed reflecting the reality, or simply a model that reflects the scenario given. The model's explanatory variables abide to the guidelines except for the State of Economy indicator. The fact that its sign is positive indicates that the relationship described by the model leans towards reflecting the scenario rather than the reality. This is, however, hard to evaluate and requires a further evaluation when more data on the portfolio is available. The model is sensitive to its structure; A mere 1 lag increase in either of the explanatory variables rendered the model statistically insignificant which is a sign that the model is not robust.

The in-sample forecast showed promising results as the direction of change in the Observed Default Rate is very similar to the model's Probability of Default. During the last year the Probability of Default declined slightly below the Observed Default Rate. The linear regression did slightly better showing a small overestimation along the entire forecast period with movements similar to the portfolio's Observed Default Rate. In the out-of-sample forecast the Error Correction Model did far better than the Linear Regression Model. During the baseline scenario the Probability of Default increased to levels comparable with 2011-2013 while the Linear Regression Model decreased to an all time low. In the adverse scenario the Error Correction Model estimates an increase in PD to levels comparable with the 2007-2008 financial crisis. This is to be expected as the adverse scenario showed similar shock on the macroeconomic variables as a strong crisis. The linear regression however increased to roughly 50% higher than the Observed Default Rate during 2007-2008.

Stress testing is central in credit risk as it is required by regulation. The models have comparable predictability in the in-sample forecast but the linear regression model had much more difficulty in the stress test where it got outperformed by the the Error Correction Model. The model has its weaknesses in economic theory but its performance in the stress test can not be neglected.

5.4 LGD evaluation

The LGD-function proved to be very misleading for this portfolio. This is most likely due to the Loss Given Default being very stable throughout the sample while Observed Default Rate was sensitive to economic events. With that being said, the function makes use of several big assumptions to generalize a function tying Loss Given Default to Probability of Default. Therefore, it seems too simplified to give a good estimate. As for this type of portfolio the LGD-function can not be recommended. However, it would be interesting to study on a portfolio where Loss Given Default and Probability of Default have a stronger correlation.

5.5 The Error Correction Model in credit risk evaluation

The Error Correction Model has shown several strengths in this thesis. Among them is its predictability in stress test scenarios where models such as linear regression models tend to give misleading results. According to Assouan (2012) the Error Correction Model's strength lies in its stress testing ability. In my thesis I have gotten the same result. Replicability tends to lend credence to any model's trustworthiness which is why I can conclude that Assouan's claim appears to be sound. The model makes use of an error correction term and seems to do well in datasets with few observations which is a common occurrence in credit risk evaluation. Another strength of the model is that it takes $I(1)$ variables into account which results in differentiation not being required. This is favorable as differentiation changes the data and information in the process is lost.

Before considering using an Error Correction Model it is important to have access to a big dataset of macroeconomic variables as the models structure require both a long-term relationship, their first differences and stationary variables. This thesis made use of 40 observations which felt like a lower limit on the required number of observations.

5.6 Conclusion

The use of an Error Correction Model in credit risk has shown several strength. Among these is its power when modeling a dependent variable of integration order 1. During this thesis several models have been tested and, while the results are discussed on the final models, there was a clear trend between the created Error Correction Models and Linear Regression Models. The Error Correction Models, by a large margin, outperformed the Linear Regression Models in stress testing. This was clearly observed in the models as the Error Correction Model maintained realistic values during the stress test while the Linear Regression Models tend to explode, predicting the Probability of Default to reach levels between 50% to 200% above the Observed Default Rate during the financial crisis of 2007-2008. Furthermore, modeling with the Error Correction Model has the advantage of incorporating non-stationary explanatory variables. This maintains the information in variables which otherwise would be differentiated. But as the model has many advantages, it also has several major restrictions. The model make use of a cointegrated relationship, non-stationary variables as well as stationary variables, and requires the explanatory variables to be of integration order 1. This is troublesome as banks are required to report certain stress tests using macroeconomic data decided on such as in the European Banking Authority's stress test. As for the stress test used in this thesis, the model can not be recommended due to restricted number of macroeconomic variables of integration order 1.

The Error Correction Model created in this thesis shows realistic predictions in stress testing as well as strong in-sample prediction of Observed Default Rate. The in-sample has movements very similar to the Observed Default Rate and makes a small overestimation. During the adverse out-of-sample scenario the Error Correction Model predicts the Probability of Default to increase to levels comparable to the financial crisis of 2008 while the baseline scenario prediction makes a conservative estimate comparable to Observed Default Rate during 2013-2014. All coefficients and lags follow sound economic theory except for the State of Economy. While this is worrisome, it requires future data of the portfolio to further evaluate.

The LGD-function proved to be very misleading in this thesis, showing both unrelated movements between Loss Given Default and Probability of Default as well as unrealistic measures. The portfolio used in this thesis did not show strong correlated movements between Loss Given Default and Probability of Default. Something Frye & Jacobs Jr (2012) based their assumptions on when deriving the function. Therefore, a general usage of the function can neither be recommended nor advised against.

6 References

Altman, E. I. & Karlin, B. J. (2010), 'Special report on defaults and returns in the high-yield bond and distressed debt market: the year 2009 in review and outlook'.

Assouan, S. (2012), 'Stress testing a retail loan portfolio: an error correction model approach', *The Journal of Risk Model Validation* **6**(1), 3.

Bank of international settlements (2005), 'An explanatory note on the basel ii irb risk weight functions'. (Accessed: 2017-04-20).
URL: *http://www.bis.org/bcbs/irbriskweight.pdf*

Bank of international settlements (2016), 'History of the basel committee'. (Accessed: 2017-04-20).
URL: *https://www.bis.org/bcbs/history.htm*

Coen, W. (2000), 'Principles for the management of credit risk - consultative document'. 2017-02-08.
URL: *http://www.bis.org/publ/bcbs54.htm*

Elliott, G., Rothenberg, T. J. & Stock, J. H. (1992), 'Efficient tests for an autoregressive unit root'.

Engle, R. F. & Granger, C. W. (1987), 'Co-integration and error correction: representation, estimation, and testing', *Econometrica: journal of the Econometric Society* pp. 251–276.

Eviews (2017), 'Dickey-fuller test with gls detrending (dfgls)'. (Accessed: 2017-04-20).
URL: *http://www.eviews.com/help/helpintro.html*

Frye, J. & Jacobs Jr, M. (2012), 'Credit loss and systematic loss given default', *The Journal of Credit Risk* **8**(1), 109.

Kennedy, P. (1998), *A guide to econometrics 4th Edition*, The MIT Press.

Phillips, P. C. & Perron, P. (1988), 'Testing for a unit root in time series regression', *Biometrika* pp. 335–346.

The European Banking Authority (2016*a*), 'About us'. (Accessed: 2017-05-20).
URL: *http://www.eba.europa.eu/about-us*

The European Banking Authority (2016*b*), 'Adverse macro-financial scenario for the eba 2016 eu-wide bank stress testing exercise'. (Accessed: 2017-05-20).
URL: *http://www.eba.europa.eu/documents/10180/1383302/2016+EU-wide+stress+test-Adverse+macro-financial+scenario.pdf*

Vasicek, O. (2002), 'The distribution of loan portfolio value', *Risk* **15**(12), 160–162.