

Modelling Loss Given Default

Comparison of 4 regressions models

Jan-Hendrik Stosshoff

readybank in a Nutshell

- Founded in 1929
- Owned by WestLB AG since March 2006
- About 350 employees (including service company)
- Balance sheet total: 750 Mio. Euros
- Headoffice in Berlin, large office in Düsseldorf
- Specialised in consumer lending, main distribution channels:
 - local savings banks (strategic partnership)
 - automotive

Agenda

- Introduction
- Explanation and remarks to data – descriptive statistics
- Regression models / Results
 - Linear regression
 - Logistic regression
 - Tobit regression
 - Beta regression
- Overview / Comparison
- Performance on different data

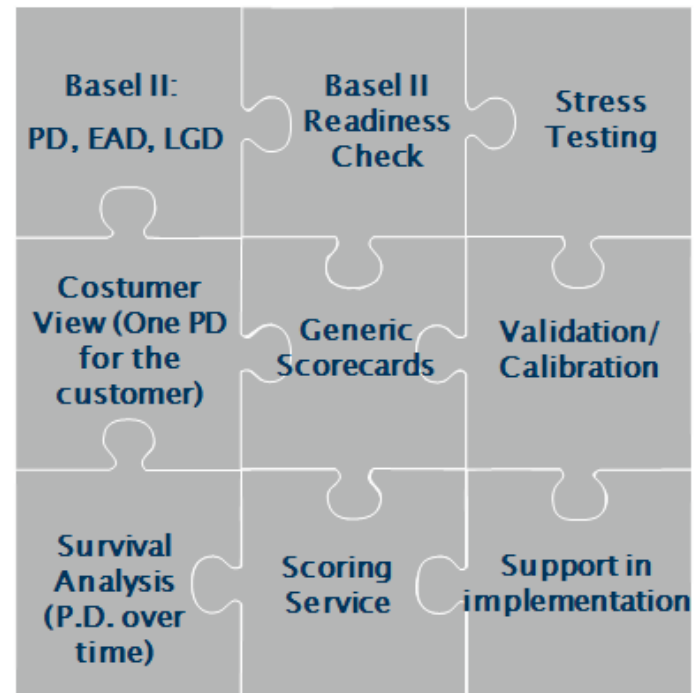
Introduction



➤ This presentation is based on a diploma-thesis in cooperation with arvato infoscoring

- arvato infoscoring provides support on any kind of risk oriented issues and decisions within the customer life cycle. Our offer covers data on creditworthiness of end consumers and B2B-customers, tools (e.g. scorecards, IT platforms) and services for optimizing customer management processes.
- Our clients are market leaders in their industries, namely banking/finance, insurance, telecommunication, mail order & eCommerce.

Service Portfolio Banking (Selection)



Introduction



➤ arvato infoscores list of references (selection):



Introduction

- The new Basel Accord allows to calculate credit risk capital requirements using an advanced internal rating based approach (IRB).
- Credit risk capital requirements depend essentially on:
 - PD (Probability of Default)
 - EAD (Exposure at Default)
 - LGD (Loss Given Default)
- While there is a lot of literature regarding PD models, there is hardly any research concerning LGD models.
 - ➡ Comparison of 4 types of generalised linear models on a credit card portfolio

The data

- Dataset: 12091 observations

- Target value / dependent variable:
 - Loss Given Default: $LGD = Loss / EAD$

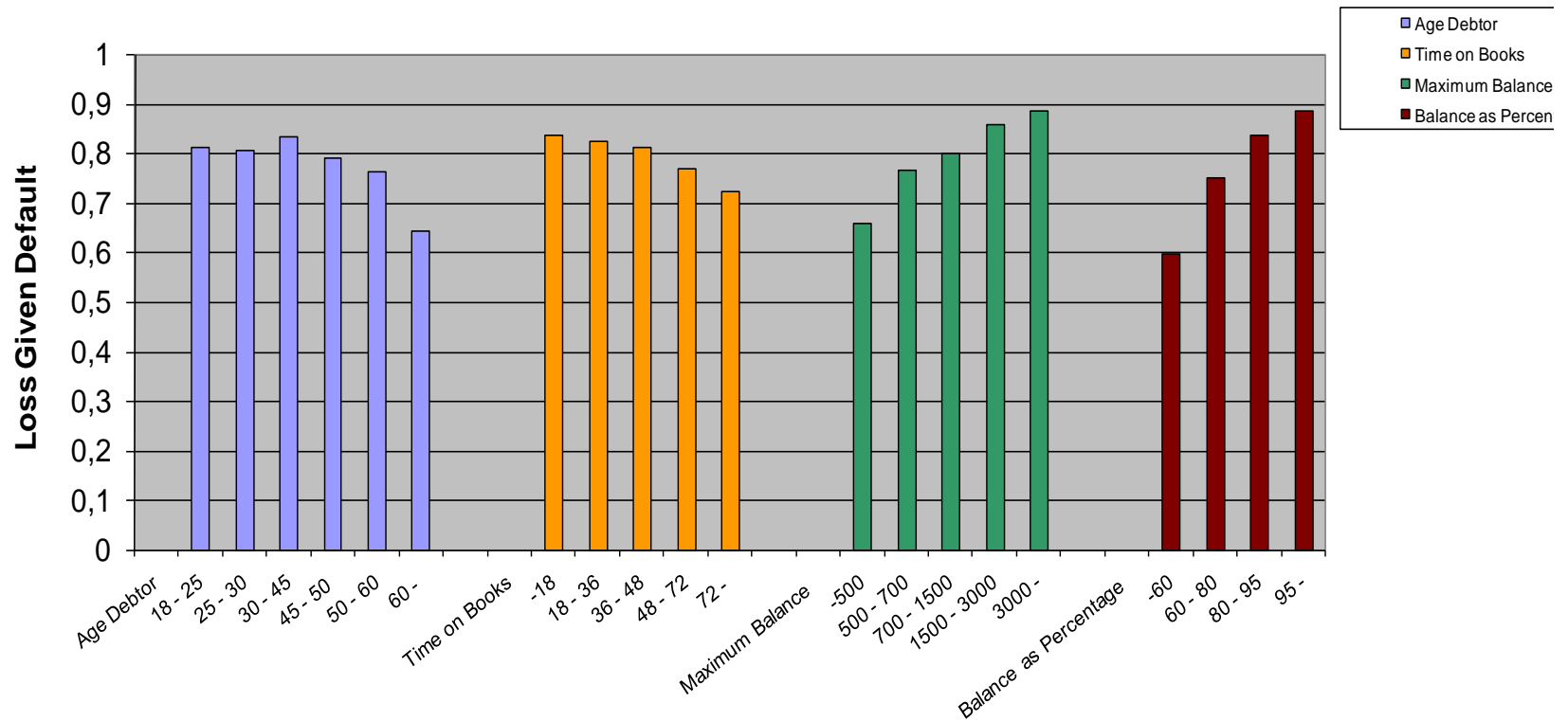
- Covariables / independent variables:
 - Age of Debtor

 - Time on Books of Revolving Account

 - Maximum Balance (of last 12 months)

 - Balance as Percentage of Maximum Balance
(Comparison of „Maximum Balance“ and balance of previous month)

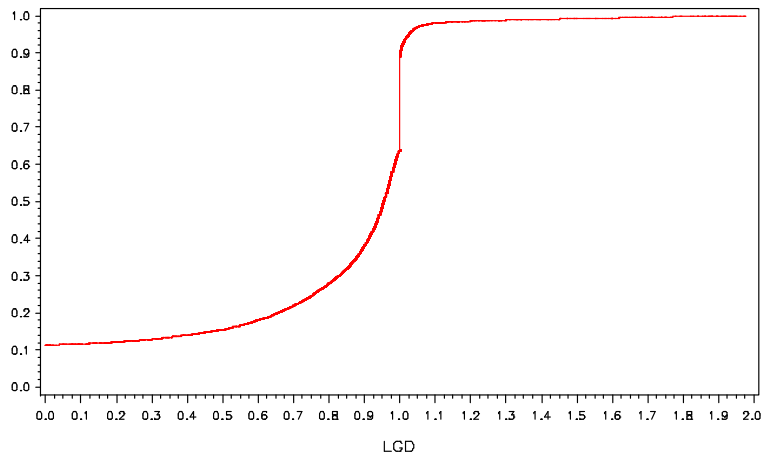
Descriptive statistics: independent variables



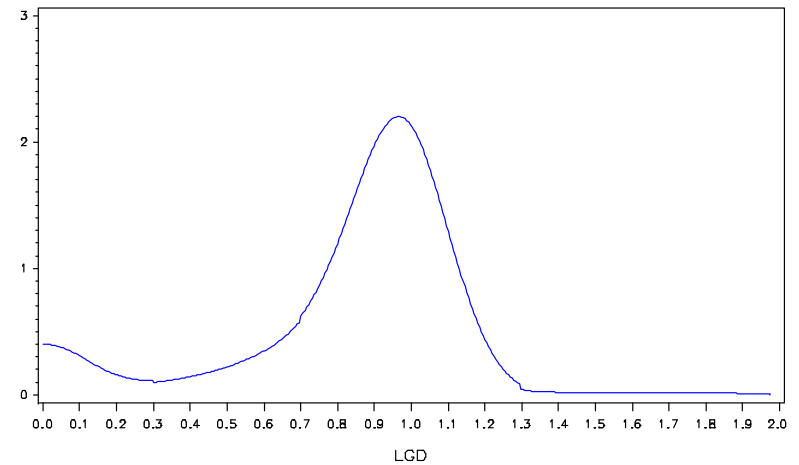
- Higher Age leads to a smaller LGD, same holds for Time on Books
- Higher Maximum Balance leads to a higher LGD, analogous Balance as Percentage
- Exception: Age-Class 30-45 leads to a higher LGD than Age-Class 25-30

Descriptive statistics: dependent variable

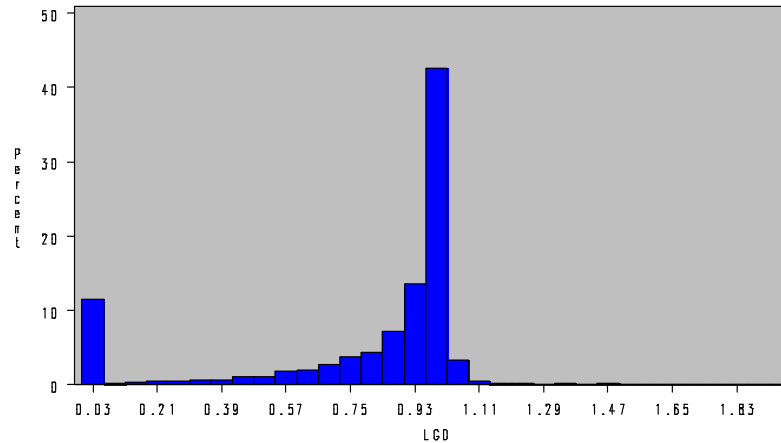
Empirical distribution



Density



Histogram



Accumulation of debtors with LGD=0 and LGD=1 leads to extreme bimodality. (Eye-catching: Peaks at LGD=0 and LGD=1 in all descriptive statistics)

Regression models: Type of models

- We use 4 different types of regression-models:
 - Linear regression
 - Logistic regression
 - Tobit regression
 - Beta regression

- All models belong to the class of Generalized Linear Models (GLM) and use the maximum likelihood estimation.

Regression models: Model quality

- To measure model quality we use a segmentation

Segmentation											
Segment	1	2	3	4	5	6	7	8	9	10	11
% of data	5	15	20	20	10	5	5	5	5	5	5

Segment 1 contains the highest 5% of predicted LGD

Segment 2 contains the next highest 15% ...

...

- On these segments the following measures are calculated:
 - Mean-Square-Error (calibration quality)

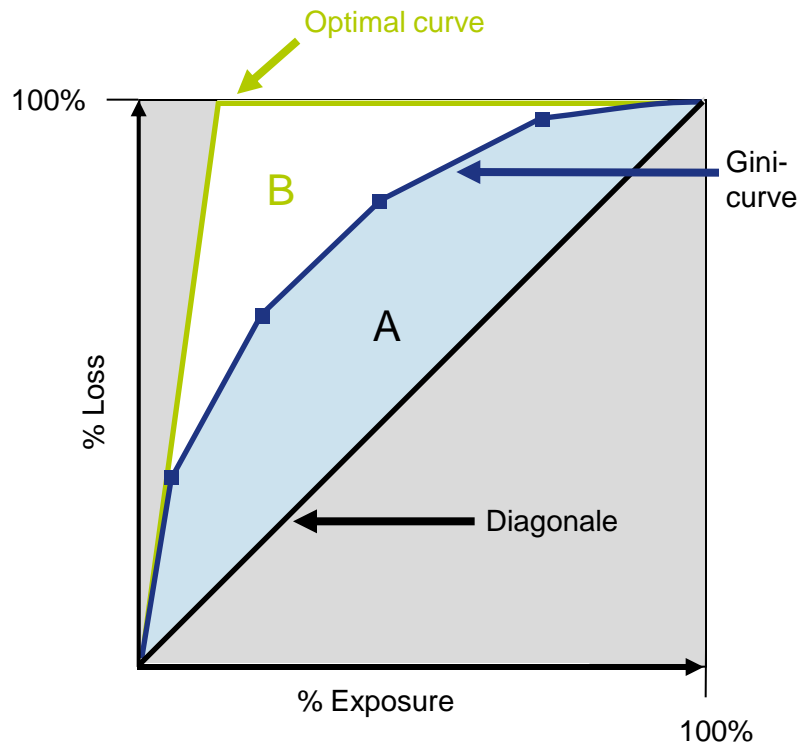
$$MSE_{Segment} = \frac{1}{11} \sum_{i=1}^{11} (\hat{y}_i - \bar{y}_i)^2$$

- Gini coefficient (discriminatory power)
- Furthermore the Mean-Square-Error is calculated independently from segmentation

$$MSE = \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2$$

Gini coefficient

Example: Exposure-weighted LGD



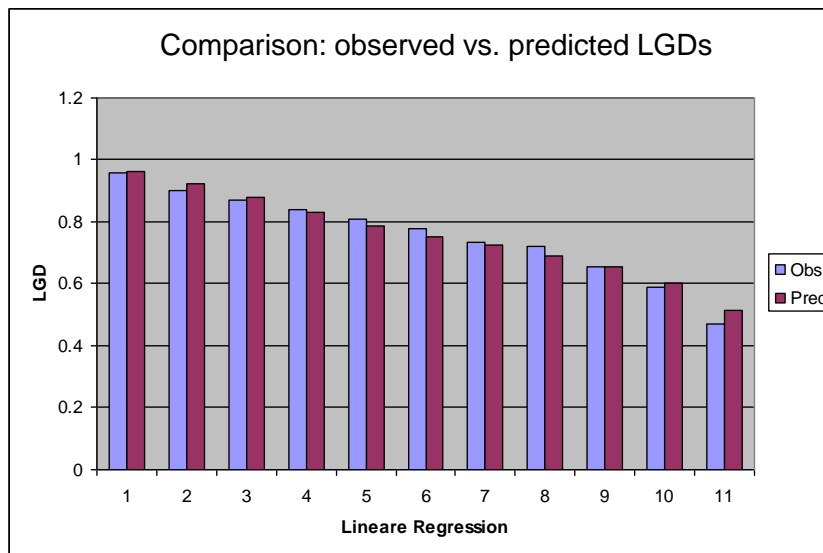
$$Gini = \frac{\text{area } A}{\text{area } A + \text{area } B}$$

Optimal curve The corner of the optimal curve is determined by $\Sigma \text{Loss} / \Sigma \text{Exposure}$

Gini curve cumulative percentage of Exposure (ordered by predicted LGD) vs. cumulative percentage of related Loss.

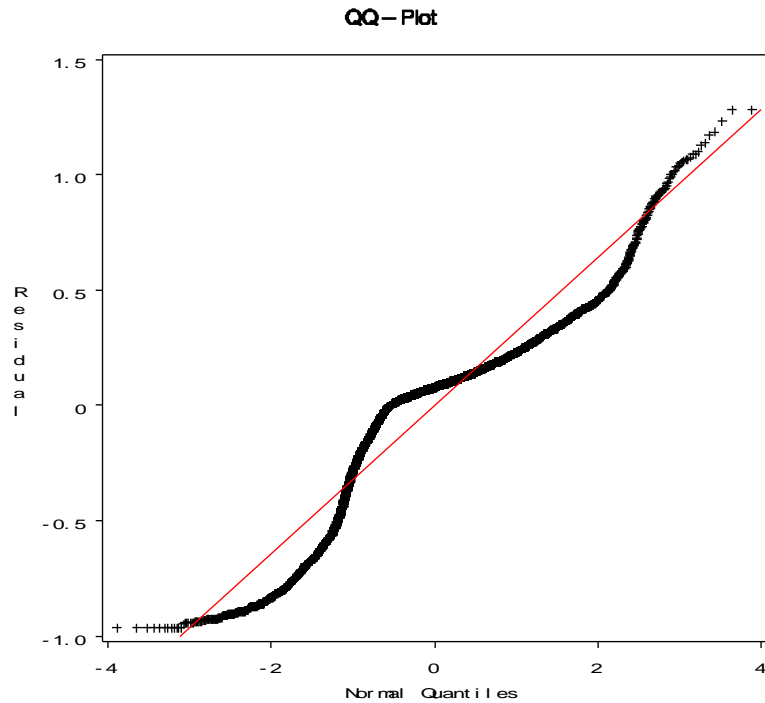
Linear regression

Model	Restriction	MSE	MSE _{Segment}	Gini	Remark
Linear	-	0,103	0,005	0,37	



Observed and predicted values are similar

Linear regression



White Test		
DF	Chi-Square	Pr > ChiSq
111	1341.66	<.0001

Durbin-Watson D	1.978
-----------------	-------

Linear regression underlies some Assumptions:

1) The residuals are normally distributed.

The QQ-Plot shows a violation of this assumption: observed and theoretical quantiles does not match with the line $y = x$.



2) The variance of error term has to be equal.

Violation: The White Test displays heteroskedasticity.



3) Uncorrelated error terms. Usually proved by a

Durbin Watson Test. Result: No correlation ($D \approx 2$).



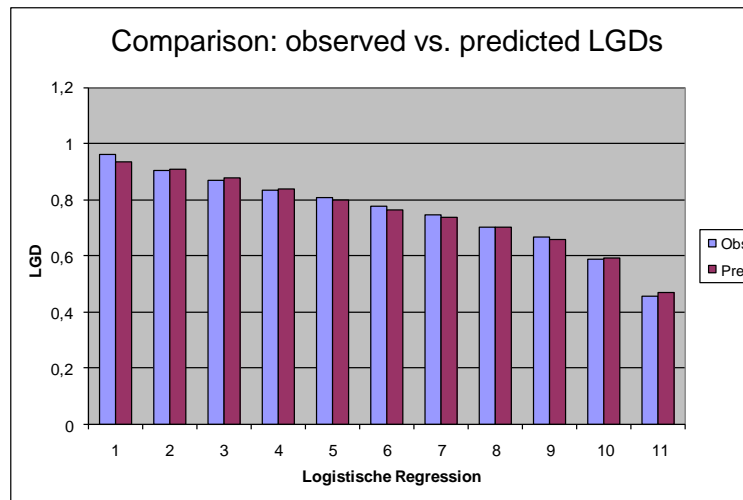
4) Expected Error = 0. Is not violated.



Logistic regression

- Problem: logistic regression requires a dichotomous dependent variable
- interpret LGD as probability to be „good“ or „bad“.
good \rightarrow LGD = 0 bad \rightarrow LGD = 1

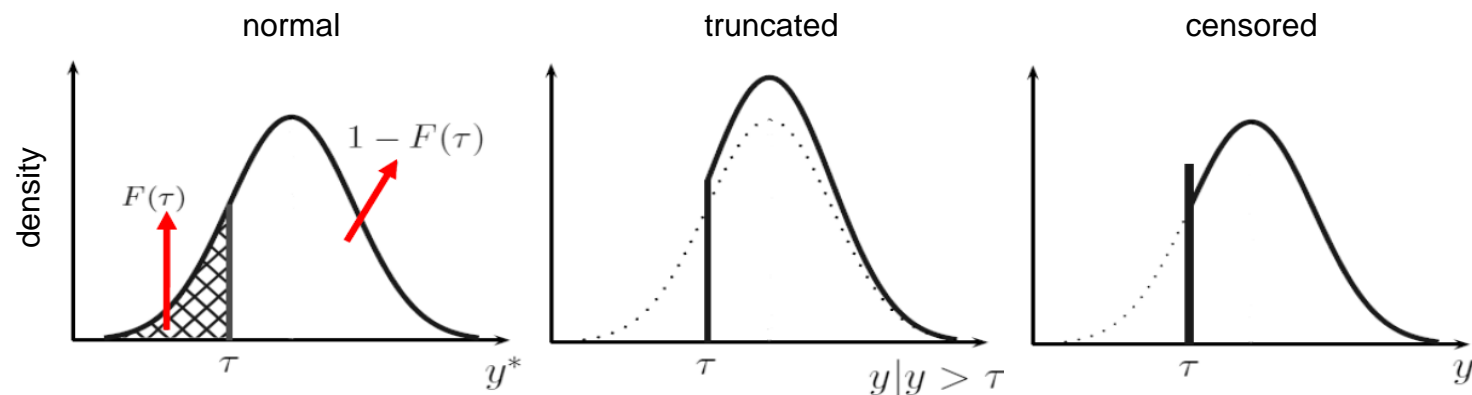
Model	Restriction	MSE	MSE _{Segment}	Gini	Remark
Logistic	-	0,102	0,002	0,38	-



Small difference between
observed and **predicted**
values

Tobit regression

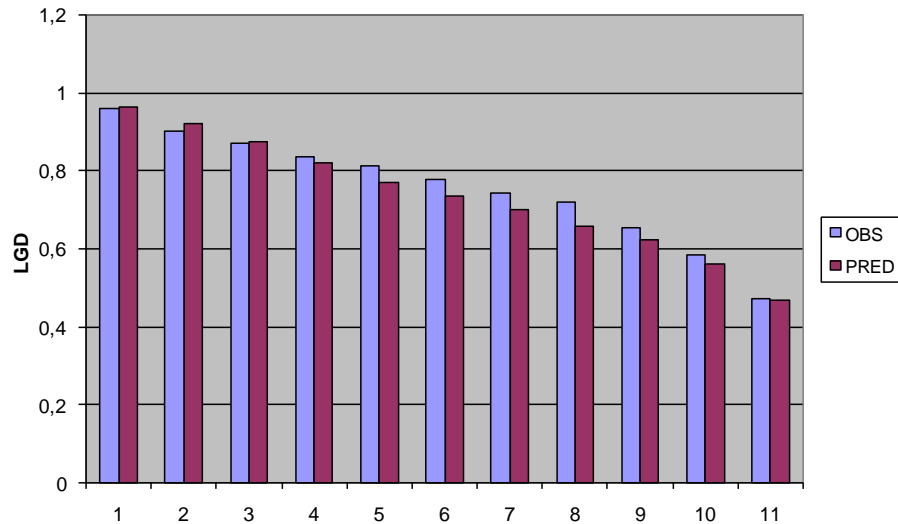
- Described by Tobin (1958)
- For censored or truncated data (assuming a normally distributed error term)
- What means censored? Dependent variable is interval-scaled in a certain interval, but underlies some restrictions in another interval (for example: is not observable or takes only one specific value)
- We use a left censored and an interval censored regression model



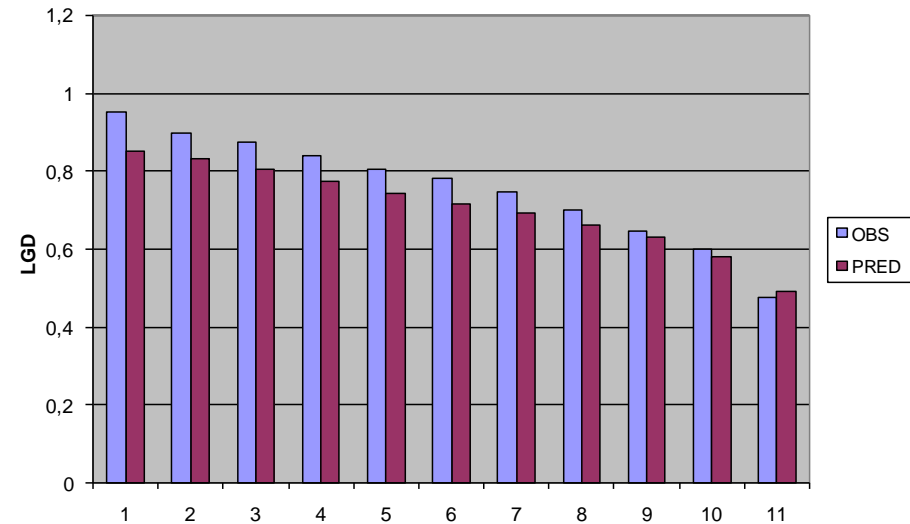
Tobit regression

Model	Restriction	MSE	MSE _{Segment}	Gini	Remark
Tobit	left & right	0,106	0,037	0,37	left censored at 0; right censored at 1
	left	0,103	0,011	0,37	left censored at 0

Comparison: observed vs. predicted LGDs
left censored

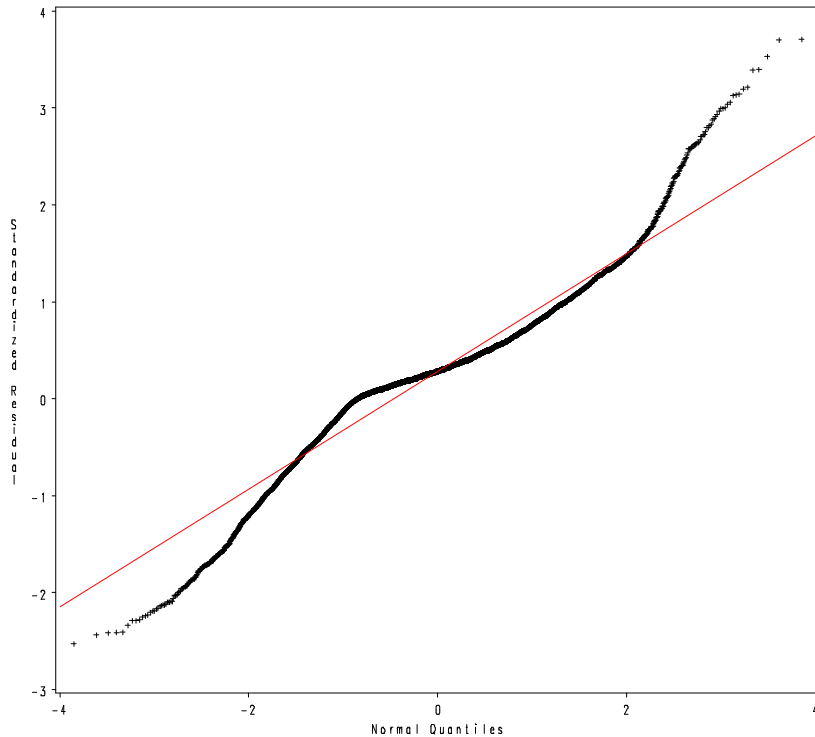


Comparison: observed vs. predicted LGDs
left & right censored

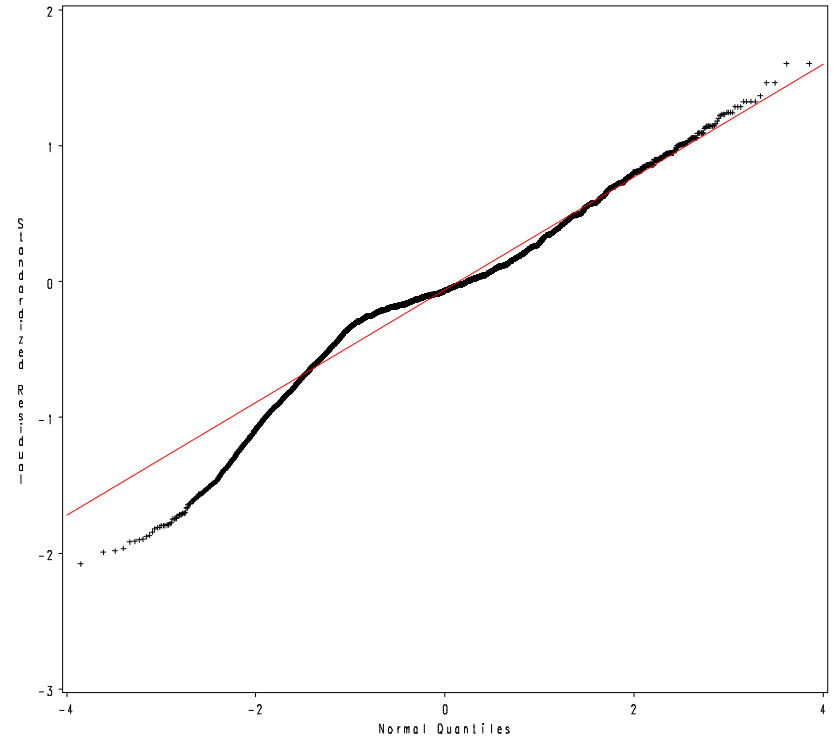


Tobit regression

QQ-Plot - left censored



QQ-Plot - left & right censored



Main Assumption:

- 1) normally distributed residuals.

The QQ-Plot shows a violation of this assumption (but less than the linear model).

In literature there are several warnings: Tobit regression is sensitive to the violation of assumptions

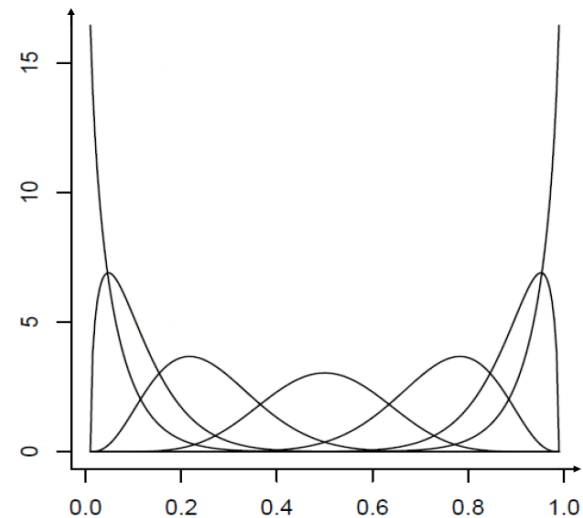
Beta regression

- For datasets where the variable of interest is beta-distributed and restricted to the interval (0,1)
- Beta-distribution is very flexible and therefore interesting for regression-modelling

$$\pi(y | p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1} (1-y)^{q-1}$$

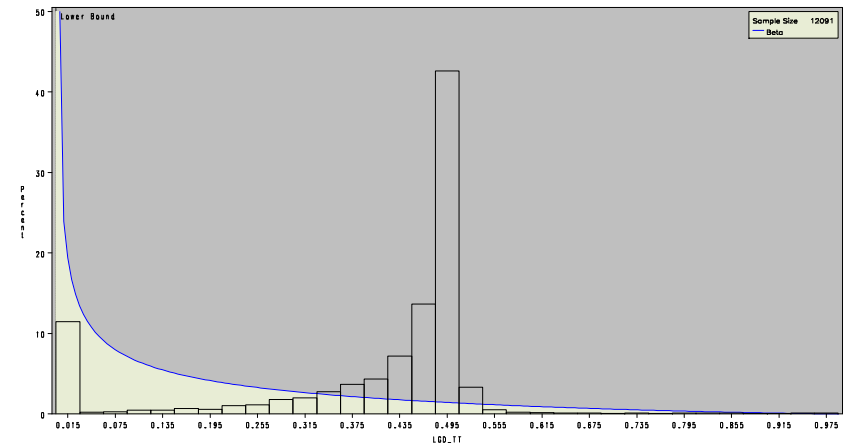
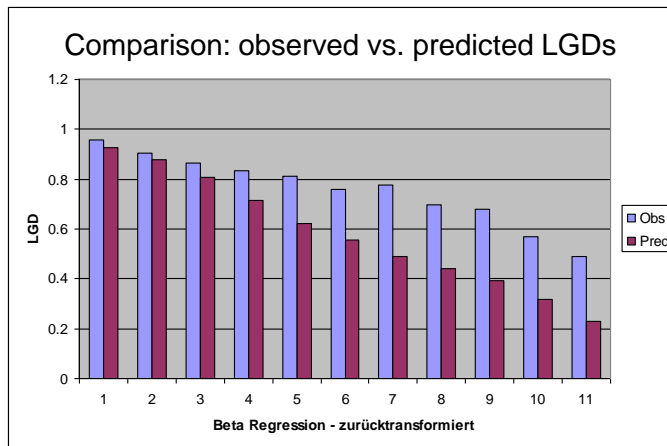
$y \in (0,1)$ $p, q > 0$ $\Gamma(\cdot)$ = Gamma - function

Possible Beta-density-functions:



Beta regression

Model	Restriction	MSE	MSE _{Segment}	Gini	Remark
Beta	-	0,130	0,459	0,36	-



Big difference between observed and predicted values.

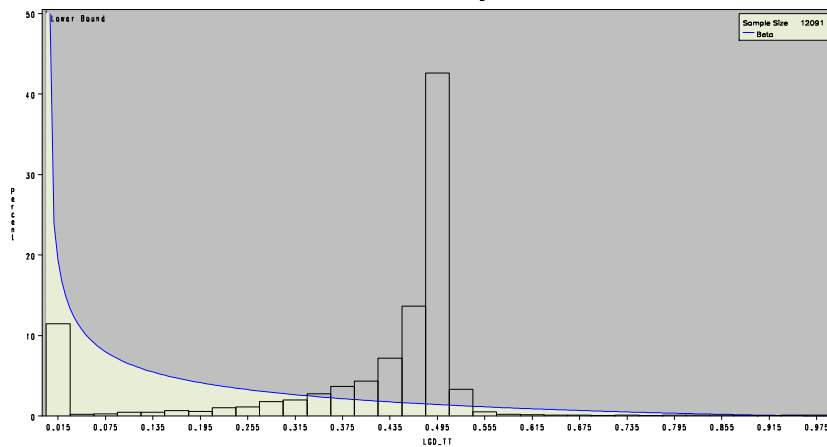
Drawback of Beta-models: bimodal distributions are only assumed, if peaks are at the margins of distribution. The model does not fit the present dataset.

Beta regression

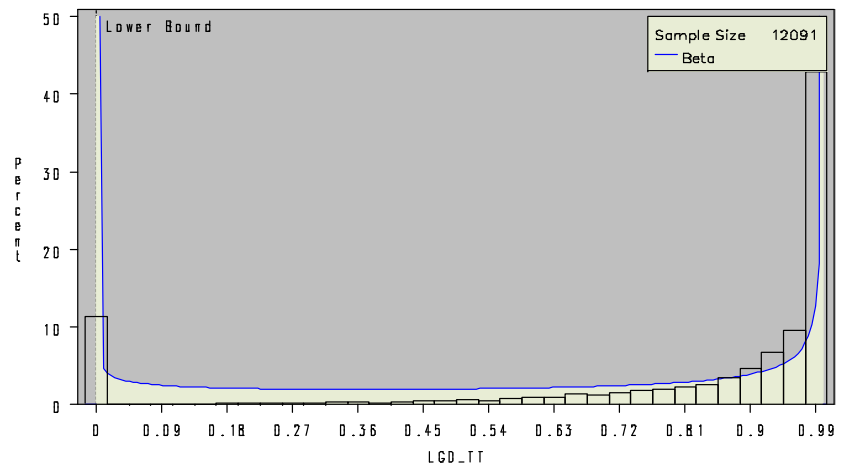
Censoring as a way to eliminate the illustrated drawbacks?

Model	Restriction	MSE	MSE _{Segment}	Gini	Remark
Beta	-	0,130	0,459	0,36	-
	truncating	0,118	0,158	0,33	LGD ≤ 1
	censored	0,111	0,152	0,36	if LGD > 1 set LGD = 1

Beta model with presented drawback



Censored (Truncated) Beta model



Improvement by censoring (or truncating)

LGD at right margin of distribution

Overview

Model	Restriction	MSE	MSE _{Segment}	Gini	Remark
Linear	-	0,103	0,005	0,37	
Logistic	-	0,102	0,002	0,38	
Beta	-	0,130	0,459	0,36	
	truncated	0,118	0,158	0,33	LGD<=1
	censored	0,111	0,152	0,36	if LGD >1 set LGD = 1
Tobit	left & right	0,107	0,037	0,37	left censored at 0; right censored at 1
	left	0,103	0,011	0,37	left censored at 0

Results:

Logistic regression is state of the art (smallest MSE & highest Gini)

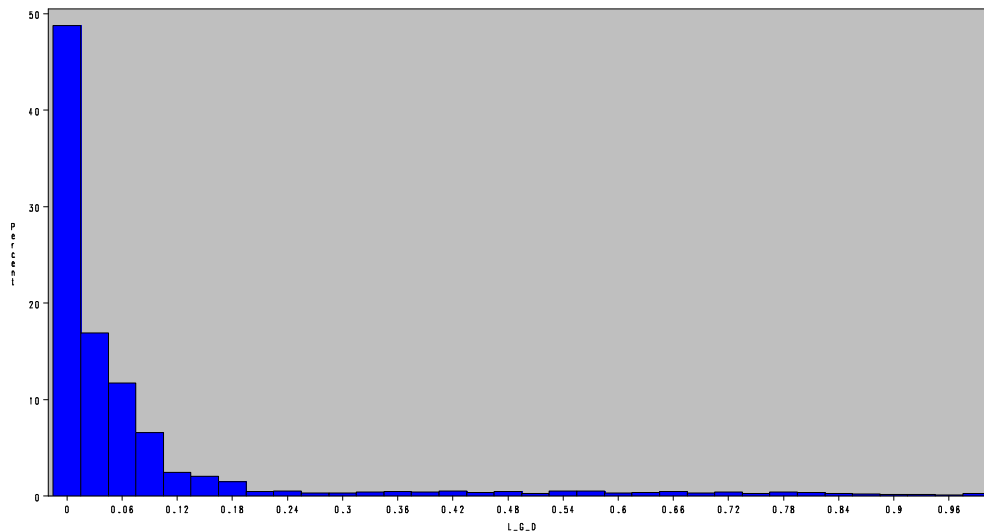
Good predictions of linear regression model- in spite of violated assumptions

In case of violated assumptions tobit regression is damageable

Beta regression is not as flexible as supposed/expected

Results on a different portfolio

- Unimodale LGD-distribution (in portfoliosection: construction financing)
- Only one Peak at LGD = 0
- Still best model quality with logistic regression, but beta regression is second choice



Model	Gini	Remark
Tobit	0,30	left-censored at LGD=0
Linear	0,29	
Logistic	0,33	
Beta	0,32	

Thank you for your attention

Jan-Hendrik Stosshoff
Risikocontrolling
Völklinger Str. 4
40219 Düsseldorf

Phone: +49 211 4955-2234

Email: jan-hendrik.stosshoff@readybank.de