Comparison of regression models for LGD estimation
A. Arsova, M. Haralampieva, T. Tsvetanova

Credit Scoring and Credit Control XII 2011, Edinburgh
Agenda

- Challenges in modelling LGD
- Data Description
- Regression Models review
- Summary of results
- Conclusions
Distribution of LGD

- LGD is a fraction - values lie usually in [0, 1]
- Non-normal distribution:
  - Often bi-modal with modes in end-point values 0 and 1
  - Uni-modal distributions with mode in either 0 or 1 can also be observed
Data Description

- All models are tested on three different data sets:
  - Fixed-term unsecured portfolio of a large international bank in the CEE region with 5700 observations of post-default LGD
  - Credit cards portfolio of a large international bank in the CEE region with 10000 observations of post-default LGD
  - Portfolio of auto loans of a leasing company in the CEE region with 2000 observations of LGD as of point of application
- Specific recoveries (from sale to debt collection agencies, insurance, etc.) and collection costs not included in calculation of LGD
- LGD capped between 0 and 100%
Data Preparation – Fine and Coarse classing of predictors

- All continuous predictors are split into intervals holding 2-5% of the population.
- Intervals with similar average LGD (difference < 0.03) merged together to smooth out the trend of mean LGD.
- Each coarse interval is modelled as a dummy variable.
LGD Models Overview

LGD Models

Single-stage Models

Direct Estimation

- OLS Regression
- Beta Regression
- Fractional Regression

Indirect Estimation

- Binary Transformation of LGD with Uniform Random Number
- Binary Transformation of LGD with weights

Multi-stage models

2-stage model: Ordinal logistic regression with nested linear regression

“Decision Tree Approach”: 3-stage model with nested logistic and linear regressions
Direct Estimation: Linear Regression Models

- Simple Linear Regression
  - All dummy variables have been considered for the first iteration of the model
  - Risk of over-fitting – standard errors of estimates may be biased due to non-normality of residuals

- Staged Linear Regression with stepwise selection
  - 2 or 3 stages; at each subsequent stage the residual from the previous stage is modelled
  - Predictors for each stage are selected by business logic
  - For each stage Stepwise linear regression is used
Direct Estimation: Beta Regression

- Assumes LGD has Beta distribution
- Beta distribution is defined over $(0,1)$ → small constant $(10^{-5})$ is added to / subtracted from end-point values 0 and 1
- Different parameterization of the Beta density function of $\xi \in \text{Beta}(\alpha, \beta)$ suggested by Ferrari and Cribari - Neto (2004) allows modelling the mean LGD

\[
\mu = \frac{\alpha}{\alpha + \beta}, \quad \phi = \alpha + \beta \quad \rightarrow \quad \mathbb{E}(\xi) = \mu, \quad \text{Var}(\xi) = \frac{\mu(1-\mu)}{1+\phi}
\]

\[
f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma((1-\mu)\phi)}y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1}, \quad y \in (0, 1)
\]

- Logit link function $g(.)$ is used for mapping $(0,1)$ into $\mathbb{R}$: $g(\mu_t) = \sum_{i=1}^{k} x_{ii}\beta_i = \eta_t$

- Log-likelihood function:

\[
l_t(\mu_t, \phi_t) = \log \Gamma(\phi) - \log \Gamma(\mu_t\phi) - \log \Gamma((1-\mu_t)\phi) + (\mu_t\phi - 1) \log y_t + ((1-\mu_t)\phi - 1) \log(1-y_t)
\]
Direct Estimation: Fractional Regression

- This model assumes that \( E(y_i|x_i) = G(x_i\beta) \) (1), where \( 0 < G(z) < 1 \) for all \( z \) in \( \mathbb{R} \).

- Logit link function: 
  \[
  G(z) = \frac{e^z}{1 + e^z}
  \]

- Quasi-maximum likelihood method using the Bernoulli likelihood proposed by Papke and Wooldridge (1996) is employed:
  \[
  l_i(b) \equiv y_i \log[G(x_i b)] + \log(1 - y_i)\log[1 - G(x_i b)]
  \]

- The QMLE estimator of \( \beta \) obtained by maximizing 
  \[
  \max_b \sum_{i=1}^{N} l_i(b)
  \]

  is consistent and asymptotically normal regardless of the true distribution of \( y_i \) conditional on \( x_i \) provided that (1) holds.
Indirect Estimation: Logistic regression with binary transformation of LGD by random number

- Continuous LGD transformed to binary outcome by using uniformly distributed random number in [0,1] as threshold:

  If LGD >= Random Number, then LGD Binary = 1
  else LGD Binary = 0

- Examples:

<table>
<thead>
<tr>
<th>LGD</th>
<th>Random Number</th>
<th>Binary LGD</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.024092</td>
<td>0.888345209</td>
<td>0</td>
</tr>
<tr>
<td>0.024092</td>
<td>0.007237462</td>
<td>1</td>
</tr>
<tr>
<td>0.935647</td>
<td>0.272748621</td>
<td>1</td>
</tr>
<tr>
<td>0.678236</td>
<td>0.560836250</td>
<td>1</td>
</tr>
</tbody>
</table>
Indirect Estimation: Logistic regression with binary transformation of LGD by random number

- Transformation does not affect end-point values 0 and 1
- \( \text{Prob}(\text{LGD Binary} = 1 | \text{LGD} = c) = c \to 0 \) as \( c \to 0 \)
- Average transformation bias is 0:
  - For each of a large number of transformations (\( n \geq 1000 \))
    transformation bias is calculated as
    \( \text{Bias}_i = \text{Mean LGD}_i - \text{Mean Binary LGD}_i \)
    - T-test \( H_0: \text{Mean Bias} = 0 \) against \( H_1: \text{Mean Bias} <> 0 \) has p-value of 0.83
- Transformed Binary LGD is modelled with Stepwise Logistic Regression
- Score to realized LGD calibration performed prior to model assessment
Indirect Estimation: Logistic regression with binary transformation of LGD by weights

- Continuous LGD transformed to binary outcome by duplicating records with 0<LGD<1 and weighting them:

<table>
<thead>
<tr>
<th>LGD</th>
<th>LGD Binary</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>c, 0&lt;c&lt;1</td>
<td>1</td>
<td>c</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1-c</td>
</tr>
</tbody>
</table>

- Transformed Binary LGD is modelled with Stepwise Logistic Regression
- Estimated LGD is the same for duplicated observations
- Score to realized LGD calibration performed prior to model assessment
Multi-stage models: A “Decision Tree Approach”

```
Logistic Regression 1
No Recovery vs. Some Recovery
(LGD = 1 vs. LGD < 1)

No Recovery
P₁ = Prob (LGD = 1)

Recovery

Logistic Regression 2
Full Recovery vs. Partial Recovery
(LGD = 0 vs. 0 < LGD < 1)

Partial Recovery

Linear Regression
0 < L < 1

Full Recovery
P₀ = Prob (LGD = 0)
```
Multi-stage models: A “Decision Tree Approach”

- Assumption:
  Factors may have different influence upon “decisions” between (LGD = 1) vs. (LGD < 1) and (LGD = 0) vs. 0<LGD<1

- This is reflected in the different discriminatory power of the two logistic regressions

- Estimated LGD = P_1 + (1-P_1)(1-P_0)*Est. LGD, where
  P_1 = Prob (LGD = 1) from Logistic Regression 1,
  P_0 = Prob (LGD = 0) from Logistic Regression 2 and
  Est. LGD is estimate from Linear Regression for 0<LGD<0
Multi-stage models: Ordinal Logistic Regression with nested Linear Regression

Estimated LGD = $P_1 + P_{01} \times \text{Est. LGD}$
## Models Assessment – Fixed Term loans  
Performance Metrics on 80% Development Sample

<table>
<thead>
<tr>
<th>Regression Model</th>
<th>Pearson’s Correlation</th>
<th>MSE / Area over the REC curve</th>
<th>R-squared</th>
<th>MAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Linear Regression</td>
<td>0.570</td>
<td>0.130</td>
<td>0.324</td>
<td>0.303</td>
</tr>
<tr>
<td>Staged linear model with stepwise selection</td>
<td>0.555</td>
<td>0.134</td>
<td>0.305</td>
<td>0.304</td>
</tr>
<tr>
<td>Beta Regression</td>
<td>0.558</td>
<td>0.139</td>
<td>0.277</td>
<td>0.338</td>
</tr>
<tr>
<td><strong>Fractional Logit (QMLE Estimation)</strong></td>
<td><strong>0.570</strong></td>
<td><strong>0.130</strong></td>
<td><strong>0.325</strong></td>
<td><strong>0.301</strong></td>
</tr>
<tr>
<td>Binary Logistic Regression with weighting transformation</td>
<td>0.564</td>
<td>0.131</td>
<td>0.319</td>
<td>0.305</td>
</tr>
<tr>
<td>Logistic Regression with Random Number transformation</td>
<td>0.562</td>
<td>0.132</td>
<td>0.316</td>
<td>0.307</td>
</tr>
<tr>
<td>&quot;A Decision Tree Approach&quot;</td>
<td><strong>0.569</strong></td>
<td><strong>0.130</strong></td>
<td><strong>0.324</strong></td>
<td><strong>0.303</strong></td>
</tr>
<tr>
<td>Ordinal Logistic Regression with Nested Linear Regression</td>
<td>0.565</td>
<td>0.133</td>
<td>0.318</td>
<td>0.307</td>
</tr>
</tbody>
</table>
# Models Assessment – Fixed Term Loans

Performance Metrics on 20% Validation Sample

<table>
<thead>
<tr>
<th>Regression Model</th>
<th>Pearson’s Correlation</th>
<th>MSE / Area over the REC curve</th>
<th>R-squared</th>
<th>MAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Linear Regression</td>
<td>0.556</td>
<td>0.131</td>
<td>0.308</td>
<td>0.302</td>
</tr>
<tr>
<td>Staged linear model with stepwise selection</td>
<td>0.527</td>
<td>0.137</td>
<td>0.273</td>
<td>0.306</td>
</tr>
<tr>
<td>Beta Regression</td>
<td>0.538</td>
<td>0.139</td>
<td>0.263</td>
<td>0.335</td>
</tr>
<tr>
<td><strong>Fractional Logit (QMLE Estimation)</strong></td>
<td><strong>0.556</strong></td>
<td><strong>0.131</strong></td>
<td><strong>0.308</strong></td>
<td><strong>0.300</strong></td>
</tr>
<tr>
<td>Binary Logistic Regression with weighting transformation</td>
<td>0.542</td>
<td>0.134</td>
<td>0.293</td>
<td>0.306</td>
</tr>
<tr>
<td>Logistic Regression with Random Number transformation</td>
<td>0.542</td>
<td>0.133</td>
<td>0.293</td>
<td>0.306</td>
</tr>
<tr>
<td>&quot;A Decision Tree Approach&quot;</td>
<td>0.533</td>
<td>0.135</td>
<td>0.284</td>
<td>0.307</td>
</tr>
<tr>
<td>Ordinal Logistic Regression with Nested Linear Regression</td>
<td>0.542</td>
<td>0.133</td>
<td>0.294</td>
<td>0.306</td>
</tr>
</tbody>
</table>
# Models Assessment – Credit Cards
Performance Metrics on 80% Development Sample

<table>
<thead>
<tr>
<th>Regression Model</th>
<th>Pearson’s Correlation</th>
<th>MSE / Area over the REC curve</th>
<th>R-squared</th>
<th>MAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Linear Regression</td>
<td>0.412</td>
<td>0.101</td>
<td>0.169</td>
<td>0.232</td>
</tr>
<tr>
<td>Staged linear model with stepwise selection</td>
<td>0.407</td>
<td>0.102</td>
<td>0.164</td>
<td>0.232</td>
</tr>
<tr>
<td>Beta Regression</td>
<td>0.403</td>
<td>0.106</td>
<td>0.130</td>
<td>0.268</td>
</tr>
<tr>
<td><strong>Fractional Logit (QMLE Estimation)</strong></td>
<td><strong>0.417</strong></td>
<td><strong>0.100</strong></td>
<td><strong>0.174</strong></td>
<td>0.231</td>
</tr>
<tr>
<td>Binary Logistic Regression with weighting transformation</td>
<td>0.417</td>
<td>0.101</td>
<td>0.173</td>
<td>0.231</td>
</tr>
<tr>
<td>Logistic Regression with Random Number transformation</td>
<td>0.413</td>
<td>0.101</td>
<td>0.171</td>
<td>0.231</td>
</tr>
<tr>
<td>&quot;A Decision Tree Approach&quot;</td>
<td>0.418</td>
<td>0.100</td>
<td>0.174</td>
<td>0.228</td>
</tr>
<tr>
<td>Ordinal Logistic Regression with Nested Linear Regression</td>
<td>0.418</td>
<td>0.100</td>
<td>0.174</td>
<td>0.228</td>
</tr>
</tbody>
</table>
## Models Assessment – Credit Cards
### Performance Metrics on 20% Validation Sample

<table>
<thead>
<tr>
<th>Regression Model</th>
<th>Pearson’s Correlation</th>
<th>MSE / Area over the REC curve</th>
<th>R-squared</th>
<th>MAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Linear Regression</td>
<td>0.382</td>
<td>0.098</td>
<td>0.141</td>
<td>0.226</td>
</tr>
<tr>
<td>Staged linear model with stepwise selection</td>
<td>0.379</td>
<td>0.098</td>
<td>0.138</td>
<td>0.227</td>
</tr>
<tr>
<td>Beta Regression</td>
<td>0.373</td>
<td>0.102</td>
<td>0.108</td>
<td>0.262</td>
</tr>
<tr>
<td><strong>Fractional Logit (QMLE Estimation)</strong></td>
<td><strong>0.387</strong></td>
<td><strong>0.097</strong></td>
<td><strong>0.148</strong></td>
<td><strong>0.225</strong></td>
</tr>
<tr>
<td>Binary Logistic Regression with weighting transformation</td>
<td><strong>0.387</strong></td>
<td><strong>0.097</strong></td>
<td><strong>0.148</strong></td>
<td><strong>0.225</strong></td>
</tr>
<tr>
<td>Logistic Regression with Random Number transformation</td>
<td>0.384</td>
<td>0.097</td>
<td>0.145</td>
<td>0.226</td>
</tr>
<tr>
<td>&quot;A Decision Tree Approach&quot;</td>
<td>0.379</td>
<td>0.098</td>
<td>0.139</td>
<td>0.224</td>
</tr>
<tr>
<td>Ordinal Logistic Regression with Nested Linear Regression</td>
<td>0.380</td>
<td>0.098</td>
<td>0.141</td>
<td><strong>0.223</strong></td>
</tr>
</tbody>
</table>
# Models Assessment – Auto Loans

## Performance Metrics on Total Sample

<table>
<thead>
<tr>
<th>Regression Model</th>
<th>Pearson’s Correlation</th>
<th>MSE / Area over the REC curve</th>
<th>R-squared</th>
<th>MAD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple Linear Regression</td>
<td>0.238</td>
<td>0.170</td>
<td>0.0564</td>
<td>0.366</td>
</tr>
<tr>
<td>Beta Regression</td>
<td>0.234</td>
<td>0.182</td>
<td>&lt;0</td>
<td>0.408</td>
</tr>
<tr>
<td><strong>Fractional Logit (QMLE Estimation)</strong></td>
<td><strong>0.245</strong></td>
<td><strong>0.169</strong></td>
<td><strong>0.0599</strong></td>
<td><strong>0.365</strong></td>
</tr>
<tr>
<td>Binary Logistic Regression with weighting transformation</td>
<td>0.232</td>
<td>0.170</td>
<td>0.0536</td>
<td>0.367</td>
</tr>
<tr>
<td>Logistic Regression with Random Number transformation</td>
<td>0.235</td>
<td>0.170</td>
<td>0.0552</td>
<td>0.366</td>
</tr>
<tr>
<td>&quot;A Decision Tree Approach&quot;</td>
<td>0.227</td>
<td>0.175</td>
<td>0.0300</td>
<td>0.349</td>
</tr>
<tr>
<td>Ordinal Logistic Regression with Nested Linear Regression</td>
<td>0.228</td>
<td>0.171</td>
<td>0.0500</td>
<td>0.373</td>
</tr>
</tbody>
</table>
Conclusions and Further Development

- **Conclusions:**
  - No model performs significantly better than the rest on all data sets
    The above statement is also supported by some of the academic research articles listed in the References

- **Further development**
  - Other model development algorithms
  - Further refinements of the tested modelling techniques
  - Inclusion of macroeconomic factors
References

Thank you!

Credit Scoring and Credit Control XII 2011, Edinburgh