A mortgage scoring model with spatial contagion

Raffaella Calabrese
University of Edinburgh Business School
raffaella.calabrese@ed.ac.uk

joint work with Meagan McCollum and Kelley Pace

RSS Edinburgh
14 February 2017 (Happy St. Valentine's Day)

Outline

Binary Choice Models

$$Y_i = \begin{cases} 1, & Y_i^* = U_{id} - U_{ip} > 0 \\ 0, & \text{otherwise.} \end{cases}$$

$$U_{ij} = \mathbf{x}_i \beta_j + \epsilon_{ij} \quad \text{for } j = p,d.$$

- The borrower's default probability $P\{Y_i = 1\} = P\{U_{id} > U_{ip}\}$ is given by the probability that the decision-maker i chooses the alternative d.
- McFadden (1978) assumed that ϵ_{ij} are distributed as a type I extreme value (Gumbel) and deriving the logit closed-form under this assumption.
- As $V(\epsilon_{ij}) = \sigma^2(\pi^2/6)$ in McFadden's model, therefore the logit choice model implies homoscedasticity across choice alternatives.

Binary Choice Models

$$Y_i = \begin{cases} 1, & Y_i^* = U_{id} - U_{ip} > 0 \\ 0, & \text{otherwise.} \end{cases}$$

$$U_{ij} = \mathbf{x}_i \beta_j + \epsilon_{ij} \quad \text{for } j = p,d.$$

- The borrower's default probability $P\{Y_i = 1\} = P\{U_{id} > U_{ip}\}$ is given by the probability that the decision-maker i chooses the alternative d.
- McFadden (1978) assumed that ϵ_{ij} are distributed as a type I extreme value (Gumbel) and deriving the logit closed-form under this assumption.
- As $V(\epsilon_{ij}) = \sigma^2(\pi^2/6)$ in McFadden's model, therefore the logit choice model implies homoscedasticity across choice alternatives.

Spatial Binary Choice Models

To take into account omitted variables in mortgage default decisions that are spatially dependent, we introduce spatial interdependence in the error terms ϵ_j

$$\epsilon_j = A\mathbf{v}_j. \tag{1}$$

$A = (I - \rho W)^{-1}$ Spatial error model (SEM)

$A = (I - \rho W)^{-1/2}$ Conditional Autoregressive model (CAR)

$A = (I + \rho W)$ Model Averaging (MA)

- ρ is the spatial autocorrelation parameter
- W is the spatial weight matrix (exogenously given)

$$w_{ij} = \begin{cases} d_{ij} & \text{if the } i\text{-th and } j\text{-th observations are contiguous;} \\ 0 & \text{otherwise} \end{cases}$$

Asymmetric error distributions

- If there is an omitted variable asymmetrically distributed, the error term $\epsilon_d - \epsilon_p$ has a skewed distribution.
- If there is misspecification of the independent variables X, such as using a model with the explanatory variables in levels when they are in log-form, the error term $\epsilon_d - \epsilon_p$ might have a skewed distribution.
- If ϵ_d and ϵ_p have symmetric distributions, but unequal variances, the disturbances $\epsilon_d - \epsilon_p$ will have an asymmetric distribution.

FBSGEV model

We suggest the model

$$\boldsymbol{U}_d - \boldsymbol{U}_p = X(\beta_d - \beta_p) + A(\mathbf{v}_d - \mathbf{v}_p)$$

where the error component is GEV distributed with a cumulative distribution function

$$F_{GEV}(v_{ij}) = \begin{cases} \exp\left\{-\left[1+\tau\left(\frac{v_{ij}-\mu}{\sigma}\right)\right]_+^{-\frac{1}{\tau}}\right\} & \tau \neq 0 \\ \exp\left[-\left(\frac{v_{ij}-\mu}{\sigma}\right)\right] & \tau = 0 \end{cases} \quad (2)$$

where τ is the shape parameter, $\mu \in R$ is the location parameter, $\sigma \in R^+$ is the scale parameter and $x_+ = max(x,0)$. For simplicity, we consider $\mu = 0$ and $\sigma = 1$.

FBSGEV model

- The homoscedasticity assumption across alternatives could be violated by mortgage default choice, as the decision of default may have a higher level of uncertainty than the choice of repayment.
- We remove the homoscedasticity assumption across alternatives and we assume that $var(v_{id})/var(v_{ip}) \approx 0$.
- Under this assumption,

$$\mathbf{v}_d - \mathbf{v}_p \sim GEV_n(\boldsymbol{\mu} = \mathbf{0}, I_n, \tau) \tag{3}$$

where I_n is the identity matrix.

We define this model the Fast Binary Spatial GEV (FBSGEV) choice model.

FBSGEV model

- The GEV distribution is very flexible with the shape parameter τ controlling the tail behavior.
- The GEV distribution is:
  1. negatively skewed for $\tau < log2 - 1$
  2. positively skewed for $\tau > log2 - 1$
- Three groups of distributions are defined based on the value of the parameter τ:
  1. If $\tau > 0$, the Fréchet-type distribution.
  2. If $\tau = 0$, the Gumbel class (cloglog model).
  3. If $\tau < 0$, the Weibull distribution.

Estimation procedure

To estimate the FBSGEV model, we have to compute

$$F_{n,GEV}(\mathbf{b}) = \int_{-\infty}^{b_n} \int_{-\infty}^{b_{n-1}} \dots \int_{-\infty}^{b_1} f_{n,GEV}(v_1, v_2, \dots, v_n) dv_1 dv_2 \dots dv_n$$

where $v_i = v_{id} - v_{ip}$.

- The Geweke-Hajivassiliou-Keene (GHK) simulator (Geweke, 1991; Hajivassiliou and MacFannen, 1990; Keane, 1994) reduces the integral of a truncated multivariate normal to a recursive sequence of n univariate integrals.
- Beron and Vijverberg (2004) used the GHK method to estimate the parameters of a spatial probit using the Cholesky decomposition. The Recursive Importance Sampling (RIS) requires $O(n^2)$ operations to compute the multivariate integral.

Estimation procedure

- In spatial econometrics an observation depends only on a low number of nearby observations. The spatial weight matrix W may contain a large proportion of zeros, so the matrix is defined as being **sparse**.
- The variance-covariance matrix Σ can be sparse too.
- Even if the variance-covariance matrix Σ is not sparse in a CAR model, the inverse of the variance-covariance matrix $\Psi = \Sigma^{-1}$, known as a precision matrix, is sparse.
- We suggest to apply the GHK algorithm to the precision matrix.

Estimation procedure

- Let **v** be a vector of i.i.d GEV r.vs $v_i \sim GEV(\tau)$.
- The aim is to multiply **v** by a matrix to obtain a vector ϵ of correlated GEV rvs whose variance-covariance matrix is Σ.
- We can write the multivariate integral as follows

$$\int_{-\infty}^{b_n} \dots \int_{-\infty}^{b_1} f_{GEV}(v_n) f_{GEV}(v_{n-1}|v_{t>n-1}) \dots f_{GEV}(v_1|v_{t>1}) dv_1 \dots dv_n$$

- We apply the Cholesky decomposition to the precision matrix $\Psi = LQ = \Sigma^{-1}$, where L is a lower triangular matrix, Q is an upper triangular matrix and Q is equal to the transpose of L ($Q = L'$).
- $\Sigma = (LQ)^{-1} = Q^{-1}L^{-1}$
- $\epsilon = Q^{-1}\mathbf{v} = L\mathbf{v}$
- $E(\epsilon\epsilon') = E(Q^{-1}\mathbf{v}\mathbf{v}'L^{-1}) = \Sigma$.

Estimation procedure

- $Q\epsilon = \mathbf{v}$ s.t. $\epsilon_i < b_i$ for $i = 1,2,...,n$.
- The procedure begins with the last observation n, which does not depend on any other observation, and works towards the first observation.
-

$$a_n = b_n Q_{nn}$$
$$\overline{P}_n = F_{GEV}[v_n < a_n]$$
$$v_n^* \sim TGEV(a_n)$$
$$\epsilon_n^* = \frac{v_n^*}{Q_{nn}}$$

where $TGEV$ is a truncated GEV random variable.

Estimation procedure

- For the general i-th observation, the ϵ^*_{i+1}, ..., ϵ^*_n calculated in the previous steps are used as follows

$$a^*_i = b_i Q_{ii} + \sum_{t=i+1}^{n} Q_{it}\epsilon^*_t$$

$$\overline{P}_i = F_{GEV}[v_i < a^*_i]$$

$$v^*_i \sim TGEV(a^*_i) \tag{4}$$

$$\epsilon^*_i = \frac{v^*_i - \sum_{t=i+1}^{n} Q_{it}\epsilon^*_n}{Q_{it}}. \tag{5}$$

- Repeating this procedure R times, the joint probability $\overline{P}$ is

$$\overline{P} = \sum_{i=1}^{n} ln\left(\frac{\sum_{d=1}^{R} \overline{P}_i(d)}{R}\right).$$

Data

We selected Clark County, in the US state of Nevada, because

- the Metropolitan Statistical Area (MSA) lies entirely in a single county,
- Las Vegas, the most populous city in Clark County, has the largest concentration of subprime mortgage originations in the US.

Data include information on property transactions for every single family property in Clark County in 2009-2010. We only include property sales records from individuals. The final data set included 282,366 observations.

We coded the dependent variable as zeros if the property received a notice of default, otherwise as ones. In 2009 the default rate is 2.7%, in 2009-2010 it increases to 5.54%.

We estimate the following model

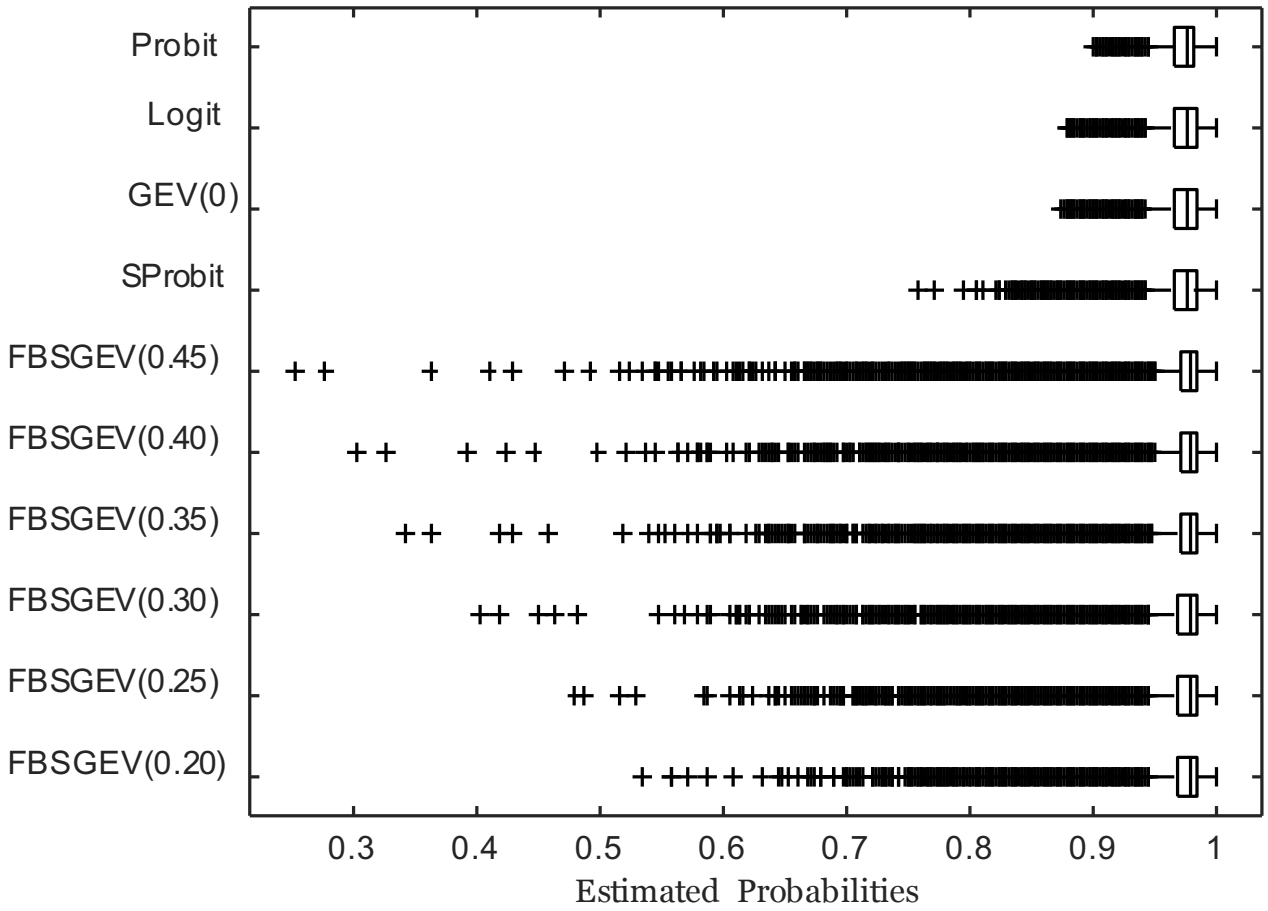
$$U_{ip} - U_{id} = \beta_1 + \ln(L/T) \cdot \beta_2 + \text{FR} \cdot \beta_3 + A(\rho)(\epsilon_{ip} - \epsilon_{id})$$
$$U_{ip} - U_{id} > 0 \rightarrow Y_i = 1$$
$$U_{ip} - U_{id} < 0 \rightarrow Y_i = 0$$
$$\epsilon_{ip} - \epsilon_{id} \sim GEV(\tau,0,1)$$
$$A(\rho) = (I_n - \rho W)^{-1/2}$$

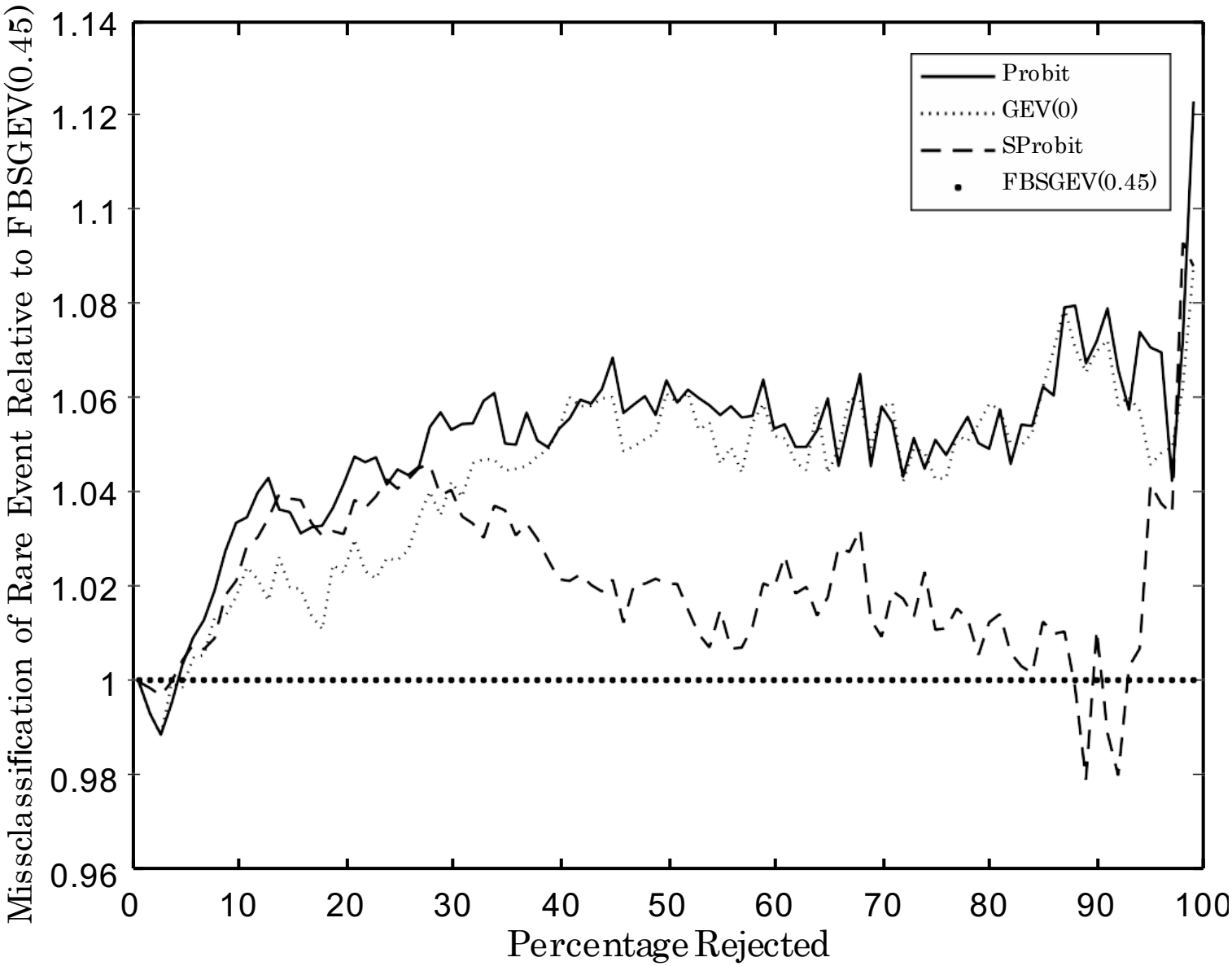
Estimates

The $t-$statistics of the fixed rate dummy t_F, log of loan-to-value $t_{L/V}$, the spatial parameter ρ and its associated t value t_ρ.

2009	t_F	$t_{L/V}$	ρ	t_ρ	ΔL	Time
Probit	20.829	−33.450			412.465	0.011
Logit	20.993	−36.340			296.008	0.004
GEV(0)	21.088	−36.719			286.747	0.005
SProbit	22.712	−54.164	0.484	20.719	283.658	27.743
FBSGEV(0.40)	21.911	−49.482	0.405	28.620	0.000	11.295

2009-2010	t_F	$t_{L/V}$	ρ	t_ρ	ΔL	Time
Probit	25.771	−48.846			685.940	0.008
Logit	25.679	−53.005			438.106	0.003
GEV(0)	25.840	−54.103			401.460	0.004
SProbit	28.056	−78.443	0.437	25.784	501.833	18.790
FBSGEV(0.35)	26.175	−73.045	0.384	34.732	0.000	7.346


Probit
Logit
GEV(0)
SProbit
FBSGEV(0.45)
FBSGEV(0.40)
FBSGEV(0.35)
FBSGEV(0.30)
FBSGEV(0.25)
FBSGEV(0.20)
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Estimated Probabilities


Misclassification of Rare Event Relative to FBSGEV(0.45)
1.14
1.12
1.1
1.08
1.06
1.04
1.02
1
0.98
0.96
0
10
20
30
40
50
60
70
80
90
100
Percentage Rejected
Probit
GEV(0)
SProbit
FBSGEV(0.45)

Value at Risk

Confidence level	0.95	0.99	0.999
Models		$LTV > 2$	
Probit	13,282.077	22,041.712	36,653.456
Logit	13,918.716	23,979.318	41,194.576
GEV(0)	13,983.279	24,224.738	42,096.184
SProbit	13,640.904	22,492.317	41,389.561
FBSGEV(0.45)	13,266.947	25,386.768	54,613.227
FBSGEV(0.40)	13,525.560	25,947.909	55,302.474
FBSGEV(0.35)	13,833.506	26,335.914	55,723.940
FBSGEV(0.30)	14,014.923	26,296.526	55,375.011
FBSGEV(0.25)	14,145.043	26,269.438	54,623.861
FBSGEV(0.20)	14,239.471	26,393.733	55,058.068

- We introduced a spatial choice model that is accurate in classifying binary rare events and can handle large sample sizes, obtaining a number of computations almost linearly with the sample size ($O(n)$).
- Main advantages:
  1. Superior performance of the FBSGEV model in classifying defaulted mortgages for different default rates in the sample.
  2. The FBSGEV model provides more reliable estimates of the probabilities of repayment compared to classic alternatives.
- The spatial dependence had an important impact on model fit, with the t statistic for the spatial dependence parameter that exceeded the t statistic associated with the fixed rate dummy. The conventional model that ignores neighbors could overestimate the probability of repayment in mortgage application or renegotiation.