3
Metodologia

Neste capítulo será mostrado inicialmente o tipo de pesquisa utilizada e a seleção da amostra seguida pela coleta dos dados. Na sequência, uma breve explanação dos modelos de regressão múltipla e *probit* ordenado serão apresentadas a fim de facilitar o entendimento do capítulo seguinte, a análise dos resultados.

3.1.
Tipo de pesquisa

Considerando-se a taxionomia proposta por Vergara (2006), este estudo pode ser classificado quanto aos fins e quanto aos meios. Quanto aos fins, esta pesquisa apresenta caráter exploratório e explicativo. Isto se deve ao fato de haver pouco conhecimento acumulado e sistematizado nesta área de pesquisa, objeto deste estudo. Quanto aos meios, pode ser considerada como experimental e bibliográfica já que utiliza referencial teórico desenvolvido a partir de material publicado em livros, jornais e revistas. Posteriormente, manipula os dados retirados de fonte secundária para fazer a verificação empírica.

Quanto à estrutura dos dados, eles são classificados como corte transversal (*cross-sectional)*. Segundo Brooks (2008), os dados de corte transversal são aqueles que se obtêm a partir de uma ou mais variáveis coletadas em único período do tempo.

O estudo desenvolvido nesta dissertação é quantitativo experimental, pois procurou avaliar, por meio de técnicas estatísticas, a influência de variáveis econômico-financeiras sobre o *rating* de emissões de dívida, a variável dependente.

3.2.
Seleção da amostra e coleta dos dados

A população dessa pesquisa consiste no universo de emissões primárias de bônus corporativos emitidos por empresas da América Latina, de países selecionados previamente, a saber: Brasil, Argentina, Colômbia, México, Panamá e Venezuela. As emissões foram realizadas no período de 2001 a 2008.

A amostra é não-probabilística, definida pelo critério de acessibilidade e as emissões foram coletadas a partir da base de dados *CapitalIQ* (divisão da *Standard&Poor's)* .

Os critérios de escolha utilizados na base do *CapitalIQ* para selecionar a amostra foram:

- ✓ Emissões com *rating* de crédito em moeda estrangeira e de longo prazo maiores do que R.
- ✓ Emissões em dólar. A escolha de uma moeda única possibilitou a comparação entre o volume de emissão de diversos países.
- ✓ Emissões com cupom fixo. Esta restrição se deu em razão de não ser possível o acesso aos dados da variação do cupom até o vencimento, o que não permitiria uma comparação adequada.

A amostra inicial obtida a partir deste primeiro critério de escolha continha 209 emissões de empresas e governos soberanos. O primeiro corte realizado na amostra buscou retirar as emissões dos governos soberanos, uma vez que os *ratings* atribuídos a estas emissões são calculados de maneira distinta dos *ratings* de emissões corporativas e não fazem parte do objeto desta pesquisa. Com isso, sobraram empresas dos mais variados setores, dentre eles, empresas do setor financeiro. Entretanto, em virtude das particularidades de análise desse setor, também se optou por excluir todas as empresas com esse perfil. Ao baixar os dados para esta pesquisa na base do CapitalIQ a partir dos critérios de escolha previamente estabelecidos, observou-se que algumas emissões vinham duplicadas. Portanto, ao preparar a base de dados para esta pesquisa também foram eliminadas todas as duplicidades de emissões encontradas.

A última restrição refere-se aos países que compõem o EMBI+. Com isso, restaram 100 emissões na amostra, pois aquelas que pertenciam às empresas do Chile tiveram que ser retiradas, já que o Chile não compõe a carteira do EMBI+ (IPEA)[1].

3.3. Modelagem do estudo

Esta pesquisa utilizará como métodos de análise o modelo de regressão múltipla pelo método dos mínimos quadrados ordinários (MQO) e o modelo *probit* ordenado.

A equação geral de ambos os modelos é a seguinte:

$$R_i = \beta_0 + \beta_1 CPN_i + \beta_2 EMBI_i + \beta_3 GAR_i + \beta_4 MAT_i + \beta_5 PRE_i + \beta_6 VOL_i + u_i \quad (1)$$

Onde R_i é o *rating* da emissão i e as variáveis independentes escolhidas foram: cupom (CPN), risco país medido pelo EMBI+ (EMBI), presença de garantia (GAR), maturidade do título (MAT), preço de emissão (PRE) e volume de emissão (VOL).

Ao início deste trabalho pensou-se em fazer um estudo comparando o desempenho do modelo (1) com um segundo modelo que incluísse além destas variáveis, indicadores financeiros. As variáveis que seriam consideradas eram: índice de cobertura de juros, retorno sobre o ativo, dívida total e ativo total.

Todavia, ao levantar os dados, deparou-se com muitos casos faltantes. Isso ocorreu porque muitas dessas empresas não disponibilizam estes dados publicamente.

Na tentativa de eliminar as empresas que não apresentavam estas informações, a base ficou muito reduzida. E isto iria prejudicar a aplicação de métodos estatísticos paramétricos.

[1] É importante ressaltar que poderíamos ter usado o índice EMBI Global para o Chile. Contudo, por este índice apresentar metodologia de cálculo diferente da metodologia do índice EMBI+, optou-se por excluir o Chile da amostra.

Assim sendo, optou-se por focar somente no objetivo principal deste trabalho. Conforme já descrito anteriormente no capítulo 1, trata-se de avaliar o impacto dos termos contratuais e do risco país no *rating* da emissão.

3.3.1. O modelo de regressão múltipla

De acordo com Wooldridge (2003), a maior parte das análises econométricas aplicadas começa com a seguinte premissa: *y* e *x* são duas variáveis representando uma população e o interesse da análise é verificar como *y* pode ser explicada pelas variações em *x*.

A regressão múltipla é considerada como um método de análise apropriado quando um problema de pesquisa envolve uma única variável dependente métrica que se relaciona com duas ou mais variáveis independentes também métricas (Hair *et al.*, 2005).

Wooldridge (2003) define o modelo de regressão múltipla com *k* variáveis independentes como:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k + u_i \quad (2)$$

Onde:

β_0 = coeficiente linear;

β_1 = parâmetro associado com x_1;

β_2 = parâmetro associado com x_2;

β_3 = parâmetro associado com x_3;

β_k = parâmetro associado com x_k;

u_i = termo do erro aleatório que não é observado

Independente de quantas variáveis independentes o modelo possa incluir, sempre existirão fatores que não serão possíveis incluir no modelo. Todos esses outros fatores estão incluídos no termo do erro, u_i.

O método de estimação mais comum para os modelos de regressão múltipla é o método dos mínimos quadrados ordinários (MQO) (Brooks, 2008). Por esse método, a melhor equação representativa da relação entre a variável dependente (y_i) e as variáveis independentes (x_i) será obtida minimizando-se a soma dos quadrados dos erros, ou seja:

$$\text{Mín}\left[\sum_{i=1}^{n} \hat{u}_i^2\right] = \text{Mín}\left[\sum_{i=1}^{n} (y_i - \hat{y}_i)^2\right] \quad (3)$$

Onde:

y_i = valor real observado de cada observação da variável dependente;

$\hat{y}_i$ = valor previsto pela reta de regressão;

$\hat{u}_i$ = erro ou resíduo, que é a diferença entre o valor real observado e o valor previsto pela reta de regressão

n = número de observações na amostra

A estimação pelo método MQO deve satisfazer às seguintes premissas: (i) que a amostra seja aleatória (ii); a linearidade entre os parâmetros; (iii) que os resíduos possuam uma distribuição normal: $u_t \sim N(0, \sigma^2)$; (iv) que os resíduos sejam estatisticamente independentes entre si, ou seja, não correlacionados: $\text{Cov}(u_i, u_j) = 0$; (v) a homocedasticidade dos resíduos: $\text{Var}(u_t) = \sigma^2 < \infty$ e (vi) que não haja multicolinearidade entre as variáveis independentes (Brooks, 2008).

Para testar as premissas, utilizam-se os seguintes testes, conforme orientação de Brooks (2008): para testar a normalidade dos resíduos utiliza-se o Teste de Jarque-Bera; para testar a homocedasticidade dos resíduos utiliza-se o Teste de White; para verificar se há autocorrelação entre os resíduos utiliza-se o teste de Durbin-Watson. E para verificar se há multicolinearidade entre as variáveis independentes analisa-se a matriz de correlação das variáveis.

Para testar a qualidade de previsão do modelo obtido, usa-se o coeficiente de determinação, o R^2. Segundo Brooks (2008), este coeficiente busca verificar a variação de variável dependente y_t em torno de sua média, $\overline{y}$. Em outras palavras, ele pode ser entendido como o coeficiente entre a soma dos quadrados explicada pela regressão (SSE) e a soma total dos quadrados (SST). Algebricamente falando:

$$R^2 = \frac{SSE}{SST} \quad (4)$$

Onde:

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

$$SST \equiv \sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$SSR \equiv \sum_{i=1}^{n} \hat{u}_i^2$$

Como $SST = SSE + SSR$, o coeficiente R^2 também pode ser escrito em termos da soma dos quadrados dos resíduos, SSR.

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \quad (5)$$

Dessa maneira, um modelo com alta qualidade de previsão deverá ter um coeficiente R^2 próximo de 1, indicando que o valor previsto é muito próximo do valor observado.

Uma importante característica deste coeficiente é que ele nunca diminui. Na verdade, ele geralmente aumenta quando mais variáveis independentes são adicionadas ao modelo (Wooldridge, 2003). Essa característica o torna uma ferramenta fraca ao ter que decidir se mais variáveis devem ser adicionadas ao modelo ou não.

Para contornar este fato, deve-se calcular o coeficiente R^2 ajustado que é calculado como:

$$\bar{R}^2 = 1 - \frac{(1 - R^2)(n-1)}{(n-k-1)} \quad (6)$$

Onde:

n = número de observações na amostra

k = número de variáveis independentes

Para testar a significância geral do modelo, utiliza-se a estatística F. Esta estatística verifica a hipótese nula de que todos os parâmetros do modelo (com exceção do coeficiente linear) são nulos simultaneamente. A hipótese alternativa é de que pelo menos um dos parâmetros é diferente de zero. Se esta hipótese for rejeitada, ou seja, se a probabilidade da estatística (*p-value*) for próxima de zero, pode-se dizer então que o modelo é estatisticamente significante (Wooldridge, 2003). A estatística F pode ser calculada em função do R^2, como a seguir:

$$F = \frac{R^2/k}{(1 - R^2)/(n-k-1)} \quad (7)$$

3.3.2. O modelo probit ordenado

De acordo com Brooks (2008), um dos usos mais importantes em finanças do modelo *probit* ordenado (*ordered probit*) é a modelagem dos fatores determinantes de *ratings* de crédito.

Uma das primeiras contribuições acadêmicas sobre o modelo *probit* ordenado surgiu em 1957 com o estudo de Aitchison e Silvey (1957). Entretanto, ele foi proposto em sua forma moderna por McKelvey e Zavoina (1975) apenas em 1975.

A principal característica de um modelo ordenado é a existência de uma variável dependente ordinal discreta (Bone, 2004). Segundo Hair *et al.* (2005), variáveis medidas em escala ordinal, como é o caso do *rating*, não apresentam uma medida da sua magnitude real em termos absolutos. O que se pode inferir é apenas a ordem entre os valores, mas não a diferença entre eles. Dessa maneira, um *rating* AAA, que receba um valor de 16 numa escala numérica, não pode ser considerado como duas vezes melhor que um *rating* BBB, cujo valor na escala numérica seja 8. Ainda nesta linha de raciocínio para dados ordinais, a diferença entre os valores 15 e 16 não pode ser assumida como equivalente à diferença entre os valores 8 e 9, por exemplo. Brooks (2008) afirma ainda que o máximo que se pode dizer é que se o *rating* aumenta pela escala numérica, existe uma relação monotônica de crescimento na qualidade do crédito.

Em modelos de variáveis dependentes ordenadas, o valor observado de y denota uma resposta cujo valor representa uma categoria ordenada. Este é o caso do modelo *probit* ordenado, que pode ser derivado de um modelo de variável latente[2]. Então, é possível modelar a variável resposta y_i considerando uma variável latente numérica y_i^* que depende linearmente das variáveis independentes x_i:

$$y_i^* = x_i'\beta + u_i \quad (8)$$

Onde $i = 1, 2, 3, \ldots, n$;

Onde x_i é um vetor de variáveis independentes; o termo de erro u_i assume uma distribuição normal e é independente de x_i; β é um vetor de coeficientes e n é o tamanho da amostra. A variável resposta y_i observada pode ser obtida a partir de y_i^* conforme a regra:

$$y_i = \begin{cases} 1 \quad \text{se } -\infty < y_i^* \leq \gamma_1 \\ 2 \quad \text{se } \gamma_1 < y_i^* \leq \gamma_2 \\ \vdots \\ M \quad \text{se } \gamma_M < y_i^* \leq \infty \end{cases} \quad (9)$$

[2] O modelo de variável latente é aquele onde a variável dependente observada é função de uma variável latente ou oculta (Wooldridge, 2003).

É importante ressaltar que os valores escolhidos para representar as categorias de y são completamente arbitrárias. Entretanto, quaisquer que sejam os valores escolhidos, para se preservar a ordenação faz-se necessário garantir que $y_i^* < y_j^*$ o que implica em $y_i < y_j$ (onde $i < j$) (BONE, 2004).

Segue então que a probabilidade de se observar cada valor de y_i é dada por:

$$\begin{aligned} P(y=0|x_i,\beta,\gamma) &= F(\gamma_1 - x_i'\beta) \\ P(y=1|x_i,\beta,\gamma) &= F(\gamma_2 - x_i'\beta) - F(\gamma_1 - x_i'\beta) \\ P(y=2|x_i,\beta,\gamma) &= F(\gamma_3 - x_i'\beta) - F(\gamma_2 - x_i'\beta) \\ &\vdots \\ P(y=M|x_i,\beta,\gamma) &= 1 - F(\gamma_M - x_i'\beta) \end{aligned} \qquad (10)$$

Onde F é a função distribuição normal acumulada.

Os valores de γ e β são estimados através da maximização do logaritmo da função de máxima-verossimilhança:

$$\text{LogL} = \sum_{i=1}^{n}\sum_{j=0}^{M} ln\left(P(y=j|x_i,\beta,\gamma)\right).m_{ij} \qquad (11)$$

Onde $m_{ij} = 1$ se $y_i = j$; senão 0.

Segundo Daykin e Moffatt (2002), a ausência de um termo constante (coeficiente linear) no modelo especificado para a variável latente é consequência dos M pontos limítrofes serem parâmetros livres[3].

[3] Para maiores informações ver estudo de Daykin e Moffatt (2002).

Como medida de "*goodness-of-fit*" para o modelo *probit*, utiliza-se a estatística denominada *McFadden's* R^2, também chamada de pseudo-R^2, definida a seguir (Brooks, 2008)[4]:

$$\text{pseudo-}R^2 = 1 - \frac{LLF}{LLF_0} \qquad (12)$$

Onde LLF é o valor maximizado do logaritmo da função de máxima verossimilhança para o modelo *probit*. LLF_0 é o valor do logaritmo de uma função de máxima verossimilhança para um modelo reduzido onde todos os parâmetros β são zerados (o modelo contem apenas um intercepto).

Em relação à interpretação dos parâmetros β, o impacto marginal da variação de uma unidade da variável independente, x_{ki}, onde β_k é o parâmetro associado à variável x_{ki}, será dado por $\beta_1(k)$ "F(" $z_1 i)$. Vale lembrar que no caso do modelo *probit*, a função F é dada por:

$$\text{"F(" } z_1 i) = \mathbf{1}/(\sigma\sqrt{2\pi})\, e^{\dagger}(-1/2\ ((z_1 i^{\dagger} 2)/\sigma)\ )\ \text{"} \qquad (13)\text{"}$$

Contudo, Daykin e Moffatt (2002) relatam que em última análise, os parâmetros β serão interpretados da mesma forma como os parâmetros da regressão múltipla.

3.4. Descrição das variáveis

Para utilizar a informação do *rating* em regressões diversos autores transformaram as classificações em valores numéricos. Segundo Hair *et al.* (2005), ao fazer esse procedimento obtém-se uma variável ordinal, um tipo de escala não-métrica, que pode ser ordenada. No entanto, observa-se uma divergência na literatura no que tange a ordenação dos valores numéricos atribuídos as classificações. Alguns autores utilizaram uma escala numérica decrescente em relação à variável latente, o *rating*. Ao melhor nível de *rating* era atribuído o maior valor da escala de conversão numérica e ao pior nível, o menor valor. (AFONSO, 2003; AMIRA, 2004; BHOJRAJ e SENGUPTA, 2003; BONE,

[4] Para maiores informações sobre esta medida ver Brooks (2008).

2004; CANTOR e PACKER, 1996; HORRIGAN, 1966; KAPLAN e URWITZ, 1979; KIM e GU, 2004; POON, 2003; POON e FIRTH, 2005; SHENG e SAITO, 2005; ZIEBART e REITER, 1992). Por outro lado, alguns autores ao invés de usarem a escala decrescente, optaram por usar uma escala crescente. Neste caso, ao melhor nível de *rating* eles atribuíram o menor valor da escala numérica e ao pior nível de *rating*, o maior valor da escala. (BUTLER e RODGERS, 2003; CALBO *et al.*, 2008; CANTOR e PACKER, 1997; CHEN *et al.*, 2007; DAMASCENO *et al.*, 2008; GABBI e SIRONI, 2005).

Seguindo a teoria do modelo *probit* ordenado apresentada acima, neste trabalho adotar-se-á a escala decrescente para a transformação da variável dependente de acordo com Amira (2004) e Bone (2004).

Tabela 1 - Conversão Numérica dos Ratings

Escala de Rating S&P	Conversão Numérica	Escala de Rating S&P	Conversão Numérica
AAA	22	BB	11
AA +	21	BB-	10
AA	20	B+	9
AA-	19	B	8
A+	18	B-	7
A	17	CCC+	6
A-	16	CCC	5
BBB+	15	CCC-	4
BBB	14	CC	3
BBB-	13	C	2
BB+	12	SD	1

Fonte: Amira (2004) e Bone (2004)

A variável cupom (CPN) foi escolhida como uma *proxy* para o prêmio de um título. De acordo com Ziebart e Reiter (1992), quanto maior o risco oferecido pelo título maior o prêmio pago aos investidores. Entretanto, para se ter uma base de comparação entre as diferentes emissões, optou-se por aquelas que ofereciam cupom fixo.

A variável EMBI + (EMBI) foi escolhida como uma *proxy* para o risco-país. De acordo com o Banco Central do Brasil (BANCO CENTRAL DO BRASIL, 2005), quanto maior o risco mais baixo, a *priori*, a atratividade de um país para o capital estrangeiro. Como conseqüência, quanto maior o prêmio, maior o retorno que os instrumentos de dívida devem oferecer para compensar o risco assumido pelos investidores.

A presença de garantia atrelada a um título de dívida oferece uma diminuição do risco assumido pelos investidores. A garantia pode ser real ou fidejussória. De acordo com Borba (2003), por garantia real entende-se toda e qualquer garantia que envolva um princípio de alienação. Como exemplos têm-se a hipoteca, o penhor, a alienação fiduciária, a caução de títulos ou a caução de direitos creditórios (FORTUNA, 2005). Como exemplo de garantia fidejussória tem-se o aval e a fiança (FORTUNA, 2005). A variável garantia (GAR) é uma variável *dummy* onde o valor 1 indica a presença de garantia para o título e o valor 0 indica a ausência.

A variável maturidade (MAT) indica o número de anos existentes até o vencimento do papel. Essa variável foi escolhida por entender-se a que a exposição ao risco da taxa de juros é maior em títulos de dívida com vencimentos longos do que em títulos que vencem num futuro próximo (BRIGHAM *et al.*, 2001; ROSS *et al.*, 2002).

A variável preço (PRE) é medida como um percentual do valor de face do título. Esta variável foi escolhida por entender-se que títulos emitidos com deságio podem representar um risco maior aos investidores.

Por fim, a última variável independente do modelo é o volume (VOL). Esta variável foi escolhida por representar uma importante característica da emissão. Porém, este trabalho adota a hipótese de que o volume não afeta a liquidez do título (CHEN *et al.* 2007; CRABBE E TURNER, 1995; GABBI e SIRONI, 2005).

A Tabela 4 abaixo apresenta a definição de cada variável independente usada bem como sua sigla utilizada no software Eviews 6.0.

Tabela 2 - Variáveis Independentes

Sigla	Nome	Definição
CPN	Cupon	Juros cotados pagos em cada título.
EMBI	EMBI +	Índice criado pelo JPMorgan para servir como benchmark dos títulos de dívida dos países emergentes.
GAR	Garantia	Variável dummy. Se 1, o título tem garantia real. Se 0, o título não tem garantia real.
MAT	Maturidade	Número de anos até o vencimento.
PRE	Preço	% do valor de face do título.
VOL	Volume	Quantia emitida em US$ Milhões de dólares.

Fonte: Elaboração da autora