ASSOCIAÇÃO BRASILEIRA DE ESTATÍSTICA
23º SINAPE – Simpósio Nacional de Probabilidade e Estatística
São Pedro, de 23 a 28 de setembro de 2018

Estimação Bayesiana de Modelos Lineares com Aplicações em Economia

Schmidt, A.V.[1]; Pozzobon, F.[2]
[1]Graduanda, Departamento de Ciências Econômicas da Universidade do Estado de Santa Catarina
[2]Professor, Departamento de Ciências Econômicas da Universidade do Estado de Santa Catarina
Contato: aishameriane.schmidt@posgrad.ufsc.br

Introdução e Objetivos

Com o advento da computação, a econometria bayesiana tem cada vez mais espaço nos trabalhos acadêmicos, tanto por sua flexibilidade na incorporação de informações a priori como também pela possibilidade de gerar estimativas confiáveis mesmo na ausência ou escassez de eventos observáveis. Neste sentido, o presente trabalho se propõe a apresentar o modelo normal de regressão linear múltipla comparando os resultados obtidos pela estimação de mínimos quadrados ordinários (MQO) com as estimativas bayesianas utilizando uma priori conjugada normal-gama. Os principais teoremas das estimativas bayesianas para o modelo foram demonstrados analiticamente.

Adicionalmente, realizaram-se simulações onde foram comparados os resultados obtidos pela abordagem clássica e bayesiana, considerando diferentes especificações dos valores de parâmetros da densidade a priori e processo gerador dos dados. Nas simulações, constatou-se os resultados teóricos desenvolvidos, tanto com relação à convergência das estimativas bayesianas para o estimador de MQO com o aumento do tamanho da amostra, como para os efeitos de variação nas informações a priori (estes resultados estão somente no texto do trabalho). Ao final, utilizando o ferramental bayesiano apresentado, adaptou-se o estudo de Bonini e Pozzobon (2016) sobre discriminação de gênero e prêmio salarial no mercado de Tecnologia da Informação, utilizando os dados da RAIS de 2011 para trabalhadores de TI do Estado de Santa Catarina. As estimativas obtidas, utilizando como priori as estimativas reportadas no estudo original, foram consistentes com as de Bonini e Pozzobon (2016) e o desempenho da estimativa bayesiana na previsão de dados para fora da amostra foi superior do que o desempenho das estimativas de MQO.

Motivação

A respeito do uso de métodos bayesianos em econometria, Geweke (2001) e Zellner (1985) citam autores da metade do século XX como precursores no desenvolvimento do assunto, sendo que um dos primeiros livros textos de econometria bayesiana é a obra de Zellner (1971), intitulada "An introduction to bayesian inference in econometrics". Daquela época até os dias atuais, com o advento da computação, métodos bayesianos se tornaram cada vez mais acessíveis e inúmeros trabalhos foram desenvolvidos na área.

Atualmente, existem diversos modelos macroeconométricos onde os métodos bayesianos são amplamente utilizados, como por exemplo os modelos de vetores aleatórios (VAR), desenvolvidos por Christopher Sims no início da década de 80. Doan, Litterman e Sims (1984) estimaram pela primeira vez um VAR bayesiano (BVAR) enquanto que a generalização do modelo considerando parâmetros variando no tempo (TVP-VAR), desenvolvida por Cogley e Sargent em 2001, já foi estimada utilizando o amostrador de Gibbs, um método bayesiano. Mais tarde, em 2005, os mesmos autores incluíram volatilidade estocástica multivariada no TVP-VAR e novamente estimaram o modelo utilizando o amostrador de Gibbs. De acordo com Geweke, Koop e Van Dijk (2011), os modelos dinâmicos estocásticos de equilíbrio geral (DSGE), utilizados amplamente em Bancos Centrais, são predominantemente estimados utilizando inferência bayesiana. De fato, desde 2011 o Banco Central do Brasil (BACEN) utiliza o SAMBA (modelo analítico estocástico com abordagem bayesiana) para auxiliar na condução da política macroeconômica no país. Caldeira, Moura e Santos (2015) fazem uma comparação de modelos, incluindo um TVP-VAR bayesiano com uso de priori de Minnesota para previsão de dados macroeconômicos do Brasil.

Parte 1. Resultados teóricos do MNRL

O modelo normal de regressão linear múltipla pode ser escrito na forma matricial $\boldsymbol{Y} = \boldsymbol{X}\beta + \varepsilon$, em que $\boldsymbol{Y}$ é um vetor $N \times 1$ contendo a variável endógena ou dependente; $\boldsymbol{X}$ é uma matriz $N \times k$ cuja primeira coluna é toda de 1 e as demais colunas representam as variáveis exógenas (ou independentes); β é um vetor com k parâmetros e ε é um vetor de erros com tamanho N. Estamos assumindo que s erros tem distribuição normal multivariada com média zero e matriz de variância e covariância homocedástica, com $Var[\varepsilon_i] = \sigma^2 = h^{-1}$ (o parâmetro h é chamado de *precisão*). Os elementos de X são fixos ou, no caso de serem v.a., são independentes do termo de erro e tem f.d.p. dada por $p(X|\lambda)$, em que λ não depende nem de β nem de h.

Na abordagem bayesiana é necessário definir uma distribuição à priori para os parâmetros do modelo, que neste caso são β e h. Neste trabalho optamos por utilizar uma priori conjugada natural (normal-gamma), que, por definição, leva a uma posteriori da mesma família de distribuições.

Ao utilizarmos uma priori conjugada natural temos o benefício de obter a posteriori em fórmula fechada, o que facilita o processo de atualização à medida que novos dados estão disponíveis. Ao mesmo tempo, é possível analisar a influência de cada componente da priori e da verossimilhança nos resultados do modelo. Estes benefícios não são gratuitos: nesta formulação não é possível ser informativo a respeito β ao mesmo tempo que se é não informativo a respeito de h, uma vez que a distribuição do primeiro é condicional ao segundo.

Como a posteriori é uma densidade conhecida, para fazer inferência neste modelo iremos precisar apenas das quantidades de MQO e dos hiperparâmetros da priori. Com isto podemos calcular momentos, intervalos de credibilidade e fazer comparação entre modelos.

Sob as hipóteses descritas anteriormente, a verossimilhança do modelo será dada por:

$$p(Y|\beta,h) = \frac{h^{N/2}}{(2\pi)^{N/2}} \exp\left[-\frac{h}{2}(\boldsymbol{Y}-\boldsymbol{X}\beta)'(\boldsymbol{Y}-\boldsymbol{X}\beta)\right]. \qquad (1)$$

Podemos combinar (1) com uma priori normal para $\beta|h$ vezes uma densidade gama para h de forma a obter a seguinte expressão a posteriori:

$$p(\beta,h|Y) \propto h^{\frac{\bar{\nu}+k-2}{2}} \exp\left[-\frac{h}{2}\left(\bar{\nu}\bar{s}^2 + (\beta-\bar{\beta})'\bar{V}^{-1}(\beta-\bar{\beta})\right)\right]. \qquad (2)$$

O lado direito de (2) é o núcleo de uma densidade Normal-Gama com parâmetros:

- $\bar{V} = \left(\underline{V}^{-1} + X'X\right)^{-1}$;
- $\bar{\beta} = \bar{V}\left(\underline{V}^{-1}\underline{\beta} + X'X\hat{\beta}\right)$, em que $\hat{\beta}$ é o estimador de MQO, enquanto que $\underline{V}$ e $\underline{\beta}$ são hiperparâmetros da priori. Logo, a expressão nos diz que o parâmetro $\overline{\beta}$ é uma média ponderada da média à priori com o estimador de MQO. Se a precisão $\underline{V}^{-1}$ for baixa, a contribuição da priori é pequena;
- $\bar{\nu} = \underline{\nu} + N$, ou seja, os graus de liberdade da priori mais o tamanho amostral resultam nos g.l. da posteriori;
- $\overline{\nu s}^2 = \underline{\nu s}^2 + \nu s^2 + \left(\hat{\beta}-\underline{\beta}\right)'\left[\underline{V} + (X'X)^{-1}\right]^{-1}\left(\hat{\beta}-\underline{\beta}\right)$, na qual $\underline{\nu s}^2$ é a SQR da priori, νs^2é a SQR da verossimilhança e o último termo serve de penalização para diferenças entre a média à priori e o estimador de MQO.

Parte 2. Aplicação empírica

Na parte empírica do trabalho, atualizamos uma parte do estudo de Bonini e Pozzobon (2016) que investigaram as diferenças salariais entre homens e mulheres trabalhadores de setores específicos dos três estados da região Sul do Brazil no ano de 2011. Neste trabalho optamos por investigar apenas os trabalhadores da área de tecnologia da informação (TI) de Santa Catarina para verificar se as diferenças salariais devidas ao gênero se modificaram quando analisados os dados de 2016. As prioris dos parâmetros são dadas por $\beta|h \sim N\left(\underline{\beta}, h^{-1}\underline{V}\right)$ e $h \sim Gamma\left(\underline{s}^{-2}, \underline{\nu}\right)$. Os resultados de Bonini e Pozzobon (2016) foram utilizados como hiperparâmetros e para acomodar a incerteza temporal, adotamos um valor alto da variância (10^2).

Uma vez que os logaritmos dos salários estão quase totalmente restritos ao interval 3.0-5.5 e por hipótese os erros são normalmente distribuídos, utilizamos $s^2 = 1$, o que implica que a priori de h é igual a 1. Como este valor é bastante arbitrátio, utilizamos um valor baixo para os g.l. da priori: $\underline{\nu} = 44$.

O modelo estimado é dado por:

$$\mathbf{ln(sal) = \beta_0 + \beta_1 \cdot S_1 + \beta_2 \cdot S_2 + \beta_3 \cdot S_3 + \beta_4 \cdot Idade + \beta_5 \cdot Idade^2 + \beta_6 \cdot TE + \beta_7 \cdot Gênero + \varepsilon},$$

em que:

- ln(sal) é o vetor com o logaritmo natural dos salários (em reais);
- β_0 é o intercepto do modelo que representa fatores não observáveis comuns a todos trabalhadores;
- S_1, S_2 e S_3 são variáveis dummy que representam a escolaridade – superior incompleto, completo e pós graduação. Quando as três são iguais a zero, a escolaridade é Ensino médio completo;
- $Idade$ é a idade em anos, que estamos conjecturando ter retornos marginais decrescentes;
- TE é o tempo de emprego, em meses;
- $Gênero$ é uma variável binária que assume valor 0 para mulheres e 1 para homens;
- as hipóteses sobre os erros são as mesmas anteriores.

Aplicação empírica (dados)

Utilizamos dados da RAIS para o estado de Santa Catarina de todos trabalhadores registrados em ocupações na área de TI no ano de 2016, totalizando 10,919 pessoas. As estatísticas descritivas do banco de dados estão nas tabelas e gráficos abaixo.

	ln(sal) (R$)	T. Emprego (meses)	Idade (anos)
Mínimo	2.067	0.00	17.00
1º quartil	4.112	14.80	27.00
Mediana	4.578	36.90	31.00
Média	4.580	55.41	32.64
3º quartil	5.015	76.50	37.00
Máximo	7.707	507.80	69.00
Desvio Padrão	0.656	62.87	8.11

Tabela 1: Descriptives of the quantitative variables entered in the model.

	n	%
Homens	8792	80.59
Ens. Superior Incompl.	1157	10.61
Superior completo	8339	76.44
Pós Graduação	184	1.69

Tabela 2: Frequências das categorias na amostra. N corresponde ao total de indivíduos na categoria. Na escolaridade o valor representa o total de pessoas cuja escolaridade máxima é a linha correspondente.

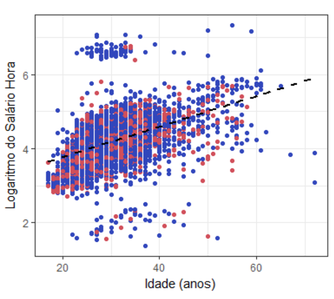

Gráfico 1: Distribuição do salário e idade, por gênero.

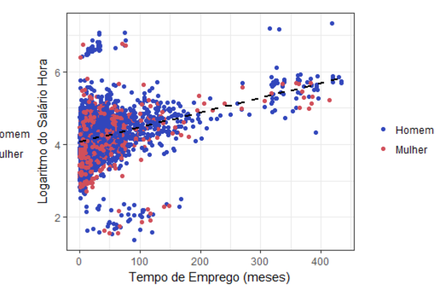

Gráfico 2: Dostribuição dos salários por tempo de emprego e gênero.

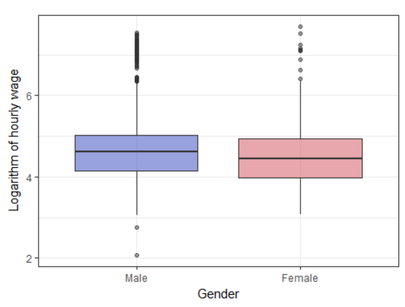

Gráfico 3: Boxplot dos salários por gênero.

Aplicação empírica (resultados)

As distribuições a posteriori dos parâmetros são dadas por $\beta|y \sim t(\overline{\beta}, \overline{s}^2\overline{V}, \overline{\nu})$, $h|y \sim \left(\overline{s}^{-2}, \overline{\nu}\right)$ e sua conjunta é $\beta, h|y \sim NG\left(\overline{\beta}, \overline{V}, \overline{s}^2, \overline{\nu}\right)$, cujos parâmetros podem ser consultados em Koop (2003). As médias a posteriori, desvio padrão, e probabilidades associadas aos β_i foram calculados utilizando a distribuição t.

Parâmetro	Média Priori	Média posteriori	Des. Pad. Posteriori	IC95%	P($\beta_i > 0$)
β_0	2.07	1.8739	0.0647	[1.747;2.001]	1
β_1	0.12	0.0949	0.0198	[0.0561;0.1338]	1
β_2	0.45	0.4943	0.0153	[0.4642;0.5243]	1
β_3	1.07	0.9525	0.0387	[0.8766;1.0284]	1
β_4	0.13	0.1053	0.0037	[0.0981;0.1126]	1
β_5	-0.00143	-0.0011	≈ 0	[-0.0012;-0.001]	0
β_6	0.00271	0.0030	0.0001	[0.0028;0.0032]	1
β_7(Gênero)	**-0.15**	**-0.1334**	**0.0117**	**[-0.1563;-0.1105]**	**0**

Tabela 3: Resultados a posteriori

Conclusões

O MNRL em sua versão bayesiana confere mais flexibilidade aos pesquisadores pois permite a inclusão de informações a priori além de cálculos de probabilidades a posteriori e maior facilidade na atualização dos resultados quando novos dados estão disponíveis. A priori conjugada natural normal-gamma permite que o procedimento de inferência seja feito em fórmula fechada, o que torna o processo computacional mais simples e rápido.

Neste trabalho vimos os resultados teóricos do MNRL bem como os resultados das simulação baseadas em experimentos de Monte Carlo (estes resultados foram omitidos do poster por questões de espaço). Por fim, propusemos um exercício empírico onde o modelo teórico utilizado no trabalho de Ponini e Pozzobon (2015) foi adaptado para uma subamostra com dados mais atuais. Os resultados mostraram concordância entre os dois estudos no sentido de que os salários das mulheres são mais baixos que os salários dos homens na amostra estudada.

Referências

[1] Bonini, P., & Pozzobon, F. (2016). *Discriminação Salarial Feminina e o Prêmio Salarial De TI Na Indústria De Tecnologia Da Região Sul*. Análise Econômica, 34(66).
[2] Koop, G. (2003). *Bayesian econometrics*. Wiley-Interscience, 1st edition.
[3] Schmidt, A. V. (2017). *Estimação bayesiana em modelos lineares com aplicações em economia.* Trabalho de Conclusão de Curso (Bacharelado em Economia). Florianópolis: Universidade do Estado de Santa Catarina, Centro de Ciências da Administração e Socioeconômicas.