PREÇO DOS CARROS VENDIDOS NA ÍNDIA

Lucas Beltrame, Malcolm dos Reis e Tiago Mattos
Instituto de Matemática, Estatística e Computação Científica - Unicamp - Inferência Bayesiana
l236193@unicamp.br, m187642@unicamp.br e t236314@unicamp.br

Introdução

O comércio de automóveis movimenta bilhões de dólares anualmente, e por isso é de interesse saber quanto cada carro deve custar a depender de seus atributos. Nesse caminho, será feito um modelo de regressão linear múltipla para tentar descrever o comportamento do preço dos carros utilizando técnicas estatísticas.

Materiais e métodos

Dados

Os dados são de um site de venda de veículos na Índia e estão dispostos como mostrado na Tabela 1

Tabela 1: Banco de dados

Variável	Descrição	Formato
name	Modelo do carro	caractere
year	Ano de lançamento do carro	numérico
selling_price	Preço de venda do carro	numérico
km_driven	Quilômetros rodados	numérico
fuel	Tipo de combustível	categórico
seller_type	Tipo de vendedor	categórico
transmission	Tipo de câmbio do carro	categórico
owner	Quantidade de donos anteriores	categórico

Análise Exploratória

Para fim da implementação de um método bayesiano aplicamos a função logaritmica no preço de venda para que assumissimos uma distribuição Normal da nossa variável resposta.

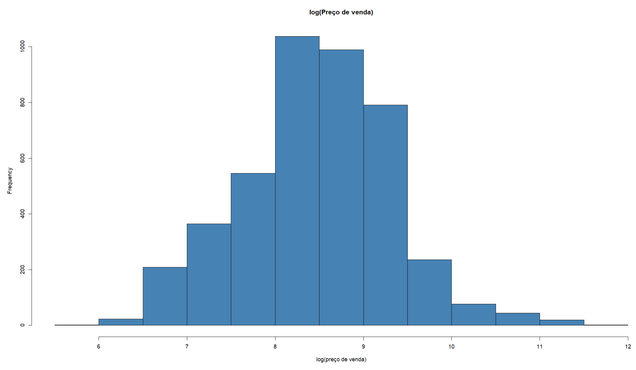

Figura 1: Histograma do log do preço de venda

O modelo será definido da seguinte forma:

$$\log(selling\_price) = \beta_0 + \beta_1 Year_i + \beta_2 KmRodados_1 + \beta_3 Transmissão\_manual_i + \epsilon_i$$

Amostrador de Gibbs

Para estimação dos betas iremos utilizar o método MCMC chamado "*gibbs sampling*"o qual utiliza uma cadeia de Markov para criar amostras de uma distribuição passada. Foi aplicada a função logarítmica no preço de venda para que fosse assumido uma distribuição normal com média desconhecida e precisão desconhecida para o preço de venda. Assim ficamos com:

$$log(y_i)|\mu, \tau \sim N(\mu, \tau)$$

para as distribuições conjugadas temos:

$$\mu \sim N(\mu_0, \tau_0)$$
$$\tau \sim Gamma(a, b)$$

Onde as distribuições posterioris dos parâmetros são:

$$\mu|\tau, y \sim N(\frac{n\bar{y}\tau + \mu_0\tau_0}{n\tau + \tau_0}, n\tau + \tau_0)$$

$$\tau|\mu, y \sim Gamma(a + \frac{n}{2}, b + \frac{1}{2}\sum_i (y_i - \mu)^2)$$

Com isso, sendo definidos as distribuições a priori dos coeficientes, esse amostrador de Gibbs parte dessa priori e a cada iteração da cadeia de Markov, vai se criando uma distribuição posteriori.

Implementação no R

Primeiro, obteve-se a matriz de design dos dados, fazendo com que a variável categórica *Transmissão* se expandisse para uma variável dummy (assumindo valores de 0 e 1). Em seguida, foram realizadas duas amostragens de Gibbs utilizando a função *MCMCregress* do pacote *MCMCpack*. Essas amostragens utilizam a distribuição a priori dos coeficientes para alcançar, por meio de iterações de uma cadeia de Markov, a distribuição a posteriori. Foram feitas duas amostragens com 100 mil iterações cada, iniciando com parâmetros diferentes distribuições nos betas para permitir que as cadeias comecem de pontos distintos, facilitando a verificação da convergência das cadeias de Markov posteriormente. Definiram-se dois vetores de médias diferentes para os coeficientes e a mesma matriz de covariância para ambas as amostragens, sendo as distribuições prioris de cada coeficiente da primeira cadeia dado por:

$$B_{01} \sim N(\mu_1, 0.0001), B_{11} \sim N(\mu_1, 0.01), B_{21} \sim N(\mu_1, 0.01), B_{31} \sim N(\mu_1, 0.01)$$

E da segunda cadeia por:

$$B_{02} \sim N(\mu_2, 0.0001),\ B_{12} \sim N(\mu_2, 0.01),\ B_{22} \sim N(\mu_2, 0.01),\ B_{32} \sim N(\mu_2, 0.01)$$

Sendo as médias definidas:

$$\mu_1 = -40, \mu_2 = 40$$

Após obter os coeficientes de cada iteração, realizou-se um burn-in de metade das iterações (50 mil) e aplicou-se um espaçamento de 100 iterações. Desse modo, utilizando a função *gelman.diag* do pacote *coda*, foi possível observar o diagnóstico de convergência de Gelman-Rubin ($\hat{R}$), que indica se as duas cadeias para cada beta convergiram entre si (no caso de convergir, é desejável que esse $\hat{R}$ seja próximo de 1).

Resultados

Através das duas cadeias produzidas com 100 mil iterações, foi possível atingir a convergência das cadeias de markov, tendo para cada coeficiente o diagnóstico de convergência estimado:

$$\hat{R_0} = 1, \hat{R_1} = 1, \hat{R_2} = 1, \hat{R_3} = 1, \hat{R_s} = 0.999$$

Com esses $\hat{R}$ podemos afirmar que para todos os coeficientes, as duas cadeias de Markov convergiram. Dessa forma, como esperado, as cadeias após o burn-in e o espaçamento trouxeram as distribuições posterioris dos coeficientes, sendo possível serem observadas através da Figura 2 e através das estatísticas sumarizadas na Tabela 2.

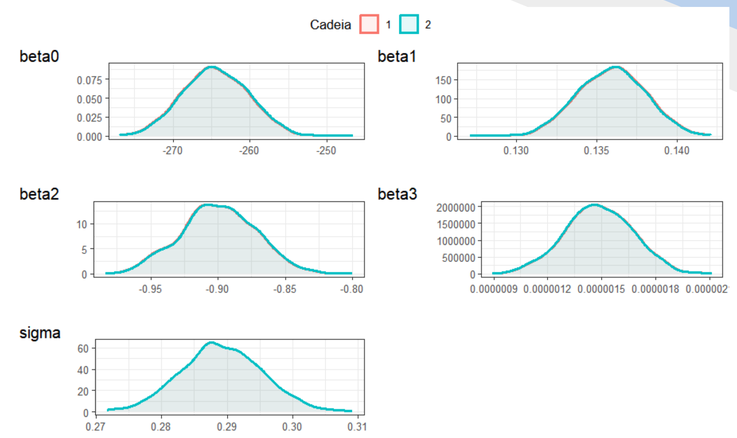

Figura 2: Gráfico das densidades dos coeficientes

	2.5%	25%	50%	75%	97.5%	Média
Intercepto	-272.559	-267.445	-264.673	-261.574	-256.043	-264.510
Ano	0.131	0.134	0.136	0.137	0.139	0.135
Transmissão Manual	-0.953	-0.918	-0.900	-0.880	-0.845	-0.899
Km Rodado	$1{,}107 \cdot 10^{-6}$	$1{,}363 \cdot 10^{-6}$	$1{,}483 \cdot 10^{-6}$	$1{,}616 \cdot 10^{-6}$	$1{,}85 \cdot 10^{-6}$	$1{,}487 \cdot 10^{-6}$
Sigma	0.277	0.285	0.289	0.293	0.301	0.289

Tabela 2: Intervalos quantílicos das distribuições e médias obtidas na cadeia

Pode-se observar que todos os coeficientes convergiram para uma distribuição aproximadamente Gaussiana.

Utilizando as médias apresentadas, podemos definir o modelo obtido da seguinte forma:

$$log(Preco) = -264.510 + 0.135 \cdot Ano + 1,487 \cdot 10^{-6} \cdot KmRodados - 0.899 \cdot TransmissaoManual$$

Diagnóstico do modelo

Sendo possível, com o modelo definido, fazer uma análise dos resíduos, utilizando a Figura 3 pode-se conjecturar que os resíduos não apresentam associação linear e são homocedásticos (variância constante). Além disso, através da sua densidade, pode-se sugerir que os resíduos seguem uma distribuição bimodal com as duas modas muito próximas de zero.

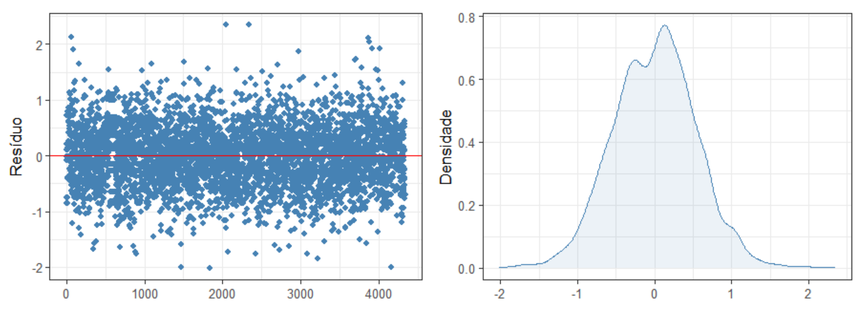

Figura 3: Diagrama de pontos e densidade dos resíduos

Sabendo, através dos diagnósticos de convergência de Gelman e Rubin, que as duas cadeias de Markov de cada coeficiente convergiram para a mesma distribuição e que, pela análise dos resíduos, estes são independentes, homocedásticos e sua distribuição é próxima de uma $N(0, \sigma^2)$, temos um modelo que explica muito bem a variabilidade do logaritmo do preço dos carros vendidos na Índia.

Desse modo, sabendo, também que nenhum coeficiente possui uma distribuição o qual o zero está dentro do intervalo de confiAssim, podemos, através do modelo, apontar que a cada aumento de 1 ano do ano de fabricação do carro (quanto maior o ano do carro, mais novo ele é), aumenta-se 0.1374 dólares no logarítmo do preço. Além disso, se o carro tiver transmissão manual, o logarítmo do preço cai 0.8997 dólares, ou seja, se o carro for automático, ele é mais caro. E por último, a cada 1 km a mais que o carro rodou, o logaritmo do seu preço será $1.4876 \cdot 10^{-6}$ dólares maior.

Considerações finais

Utilizando uma abordagem bayesiana para inferir os betas dados nossas observações, conseguimos através de um método computacional conhecido como "Amostrador de Gibbs"chegar às distribuições desses betas e analisamos a estabilidade do método partindo de pontos distintos em cadeias diferentes que levaram à uma mesma distribuição após um *burn-in* e espaçamento para diminuir a correlação de amostras subsequentes. Vale ressaltar que diferentemente do método frequentista, temos aqui um intervalo de credibilidade para cada beta que é calculado a partir de sua distribuição *a posteriori*. Com esses resultados em mão é feito o diagnóstico do modelo observando o comportamento dos resíduos e se eles seguem as premissas impostas.

Referências

- MCMCregress: Markov Chain Monte Carlo for Gaussian Linear Regression - `https://www.rdocumentation.org/packages/MCMCpack/versions/1.7-0/topics/MCMCregress`
- Notas de Aula : Inferência Bayesiana 1s2024