



MODELO DE REGRESSÃO QUASE-BETA MULTIVARIADO

Ricardo Rasmussen Petterle; Wagner Hugo Bonat; Cassius Tadeu Scarpin

Universidade Federal do Paraná - Programa de Pós-Graduação em Engenharia de Produção



Em diversas áreas de pesquisa é comum a análise de dados com variáveis respostas limitadas ao intervalo unitário. Tais variáveis geralmente se apresentam na forma de taxas, proporções, índices e porcentagens, sendo portanto limitadas ao intervalo $(0, 1)$. Quando a variável resposta é multivariada, em geral, cada variável resposta é analisada separadamente, o que não permite investigar possíveis correlações entre elas. O presente trabalho propõe um novo modelo de regressão para análise de variáveis respostas limitadas multivariada. O modelo é especificado usando apenas suposições de primeiro e segundo momentos. A abordagem usada para estimação dos parâmetros combina as funções de estimação quase-escore e Pearson para estimação dos parâmetros de regressão e dispersão, respectivamente. No decorrer do trabalho foram delineados três estudos de simulação. O primeiro foi conduzido para investigar o comportamento do algoritmo NORTA (*NORmal To Anything*) na simulação de variáveis aleatórias beta correlacionadas. O segundo visou explorar a flexibilidade dos estimadores para lidar com dados limitados em estudos longitudinais. E o terceiro foi delineado para checar propriedades dos estimadores, tais como vies, consistência e taxa de cobertura em estudos com múltiplas respostas correlacionadas. O modelo foi motivado por dois conjuntos de dados que não são facilmente manipulados pelos métodos estatísticos convencionais. O primeiro se refere ao índice de qualidade da água de reservatórios de usinas hidrelétricas operadas pela COPEL no Estado do Paraná. E o segundo corresponde ao percentual de gordura corporal, que foi medido em cinco regiões do corpo e representam as variáveis respostas. Além disso, foram adaptadas técnicas de diagnóstico para o modelo proposto, tais como DFFITS, DFBETAS, distância de Cook e o gráfico de probabilidade meio-normal com envelope simulado, para detecção de pontos influentes e *outliers*. Portanto, as principais contribuições do modelo de regressão proposto nesta dissertação estão na análise de dados limitados em estudos longitudinais, além de dados limitados em estudos com múltiplas respostas correlacionadas.

INTRODUÇÃO

Modelos de regressão são usados em diversas áreas de pesquisa para estudar a relação entre uma variável resposta (variável dependente) e possíveis variáveis explicativas. Sua aplicação é ampla, abrangendo diversas áreas do conhecimento como medicina, engenharias, agronomia, ciências sociais dentre outras. Nesse contexto, um dos principais modelos de regressão e sem dúvida um dos mais utilizados por usuários de estatística aplicada é o clássico modelo de regressão linear (Gaussiano). No entanto, para uso desse modelo alguns pressupostos devem ser atendidos, tais como erros independentes e identicamente distribuídos segundo a distribuição normal com média zero e variância constante. Quando a variável respostas pertence ao intervalo $(0, 1)$, diversos modelos foram propostos. Exemplos incluem os modelos de regressão beta e simplex. Embora os modelos supracitados possam ser usados em inúmeras aplicações, eles são limitados à análise de apenas uma variável resposta. Para o caso de duas ou mais variáveis respostas pode-se pensar em um modelo de regressão multivariado, o qual apresenta vantagens na análise dos dados e têm ganhado destaque na literatura. Por exemplo, ao se realizar exames laboratoriais ou medições de qualidade na indústria, tem-se uma grande quantidade de variáveis, que em grande parte dos casos é de difícil análise e interpretação por meio dos métodos estatísticos convencionais. Nesse sentido, há a necessidade de implementação de novos métodos para análise de dados com múltiplas respostas, que permitam investigar possíveis correlações entre as variáveis respostas dada a presença de covariáveis no modelo.

OBJETIVOS

Propor um modelo de regressão para análise de variáveis respostas limitadas multivariada, chamado por modelo de regressão quase-beta multivariado.

MATERIAIS E MÉTODOS

- Conjunto de dados I: índice de qualidade da água;
- Conjunto de dados II: percentual de gordura corporal;
- Funções de estimação quase-escore e Pearson;
- Estudos de simulação;
 - Comportamento do algoritmo NORTA;
 - Estudos longitudinais;
 - Estudos com múltiplas respostas;

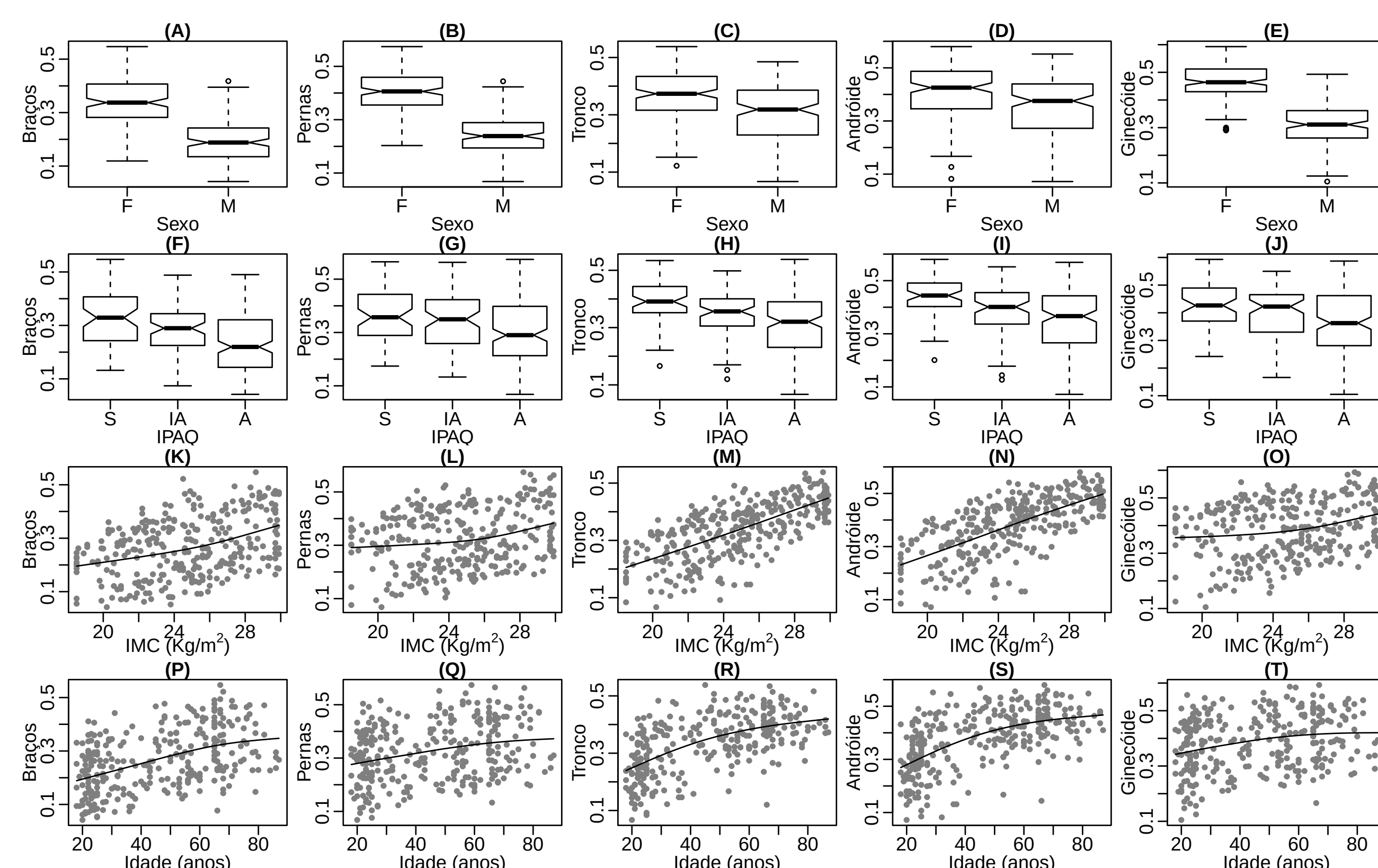


Figura : Análise descritiva e exploratória para o conjunto de dados do percentual de gordura corporal.

RESULTADOS E DISCUSSÕES

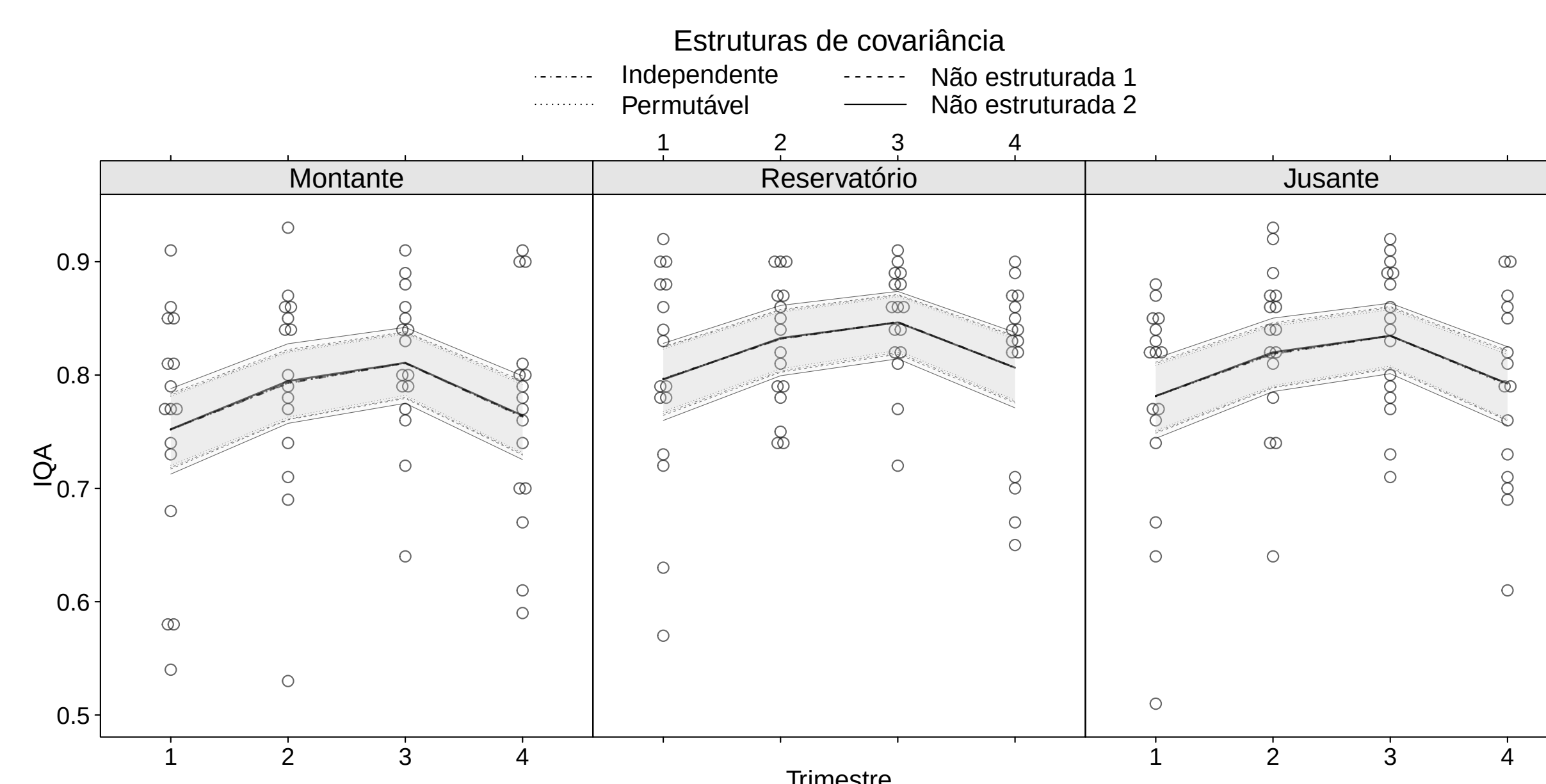


Figura : Curvas de predição com bandas de confiança (95%) para a média do IQA por local e trimestre ajustado pelo modelo de regressão quase-beta multivariado usando diferentes estruturas de covariância

CONCLUSÃO

O objetivo geral desta dissertação foi propor um novo modelo de regressão para análise de variáveis respostas limitadas multivariada. O modelo foi especificado usando apenas suposições de primeiro e segundo momentos. Para estimação dos parâmetros, adotou-se uma abordagem que combina as funções de estimação quase-escore e Pearson para estimação dos parâmetros de regressão e dispersão, respectivamente. Portanto, o modelo de regressão proposto nesta dissertação permite lidar com dados limitados em estudos longitudinais, além de dados limitados em estudos com múltiplas respostas correlacionadas. Além disso, pode-se acomodar facilmente dados no intervalo $[0, 1]$, incluindo zeros e uns.