

回帰モデルにおける補助変数を活用した推定精度の向上

前田 篤刀¹, 伊森 晋平^{2,3}, 下平 英寿^{1,3}

¹ 京都大学大学院 情報学研究科

² 広島大学大学院 理学研究科

³ 理化学研究所 革新知能統合研究センター

説明変数 X から目的変数 Y を予測する問題を考える．このとき， X に加えて別の変数 A が訓練データ内に存在する状況を考える．ただし，この変数 A は X に比べて観測のコストが大きいなどの理由で Y の予測時には観測できないとする．このような変数は補助変数と呼ばれるもので， Y の予測精度の向上に利用できると考えられる [2]．例えば，ソーシャルネットワークから収集した画像に付随するメタデータを補助変数として活用することで画像分類タスクの予測精度が向上したことが報告されている [3]．しかし，全ての補助変数が予測精度向上に有用とは限らない．補助変数を利用するのかどうか，また補助変数が複数ある場合にどの補助変数を選ぶべきかという問題がある．このような問題はモデル選択の枠組みとして捉えることができ，モデル選択規準の代表的なものとして赤池情報量規準 (Akaike information criterion; AIC) がある [1]．しかし，AIC では補助変数のように予測時に存在しない変数を想定して導出されておらず，補助変数を含むモデル選択を行う場合 AIC を利用することは理論的に妥当とはいえない．そこで，補助変数の利用を想定した情報量規準が提案されている [4]．

本発表ではこの情報量規準を紹介し，補助変数が実際に有用となる数値例を示す．例えば (Y, X, A) が次のモデルに従う状況を考える．

$$\begin{aligned} Y &= C^T X + \epsilon_1, & A &= f(CX) + \epsilon_2, \\ \epsilon_1 &\sim N(0, \sigma_1^2 I_m), & \epsilon_2 &\sim N(0, \sigma_2^2 I_l). \end{aligned}$$

ここで， f は $f: \mathbb{R}^m \rightarrow \mathbb{R}^l$ となる写像とする．この場合に補助変数を利用した予測モデルを立てると予測精度が向上することが確認された．また，補助変数の利用を想定した情報量規準を利用し，有用な補助変数が一部含まれているような状況においてモデル選択を行う．結果等の詳細については当日報告する．

参考文献

- [1] H. Akaike, “A new look at the statistical model identification”, IEEE transactions on automatic control, Vol. 19, pp. 716-723, 1974.
- [2] R. Tibshirani and G. Hinton, “Coaching variables for regression and classification”, Statistics and Computing, Vol. 8, pp. 25-33, 1998.
- [3] M. Guillaumin, J. Verbeek and C. Schmid, “Multimodal semi-supervised learning for image classification,” 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, pp. 902-909, 2010.
- [4] 伊森 晋平, 下平 英寿, “An information criterion for auxiliary variable selection”, 2016 年度 統計関連学会連合大会, 石川県 (金沢大学), 2016 年 9 月 6 日.