データサイエンス　課題 4　ロジスティック回帰

締め切り: 7 月 17 日 10:30am

各質問に答えてください。また、それぞれの答えについて、関連する R コードと出力を、文書に貼り付けてください。ファイルは PDF で提出してください。

wells.csv (Gelman and Hill 2007 のデータ) を読み込んでください。バングラデシュや南アジア諸国では、飲料水として使用される井戸の多くが天然のヒ素で汚染されており、推定 1 億人が影響を受けていると言われている。ヒ素は蓄積性の毒物であり、曝露によりがんやその他の疾病のリスクが高まり、そのリスクは曝露量に比例すると推定される。このデータセットは、飲料水の水源を井戸から別の井戸に切り替えるかどうかの決定に関わるものである。 ある地域では、隣り合った井戸のヒ素濃度が大きく異なることがある。もし、ある家の井戸のヒ素濃度が高く、安全でない場合、近くに安全な井戸を見つけて水を汲むことが（歩いて行く意思があれば）可能である。分析には重要ではないが、ヒ素のレベルが 0.5（1 リットルあたり数百マイクログラムの単位）以下であれば、井戸は安全であるとみなされる。データセットに含まれる変数は以下の通り。

switch - その世帯が新しい井戸に切り替えた場合は 1、現在の井戸を使い続けた場合は 0

arsenic - その世帯の井戸のヒ素のレベル

dist - 最も近い安全な井戸までの距離（メートル）

assoc - 世帯のメンバーの中に、地域コミュニティのメンバーがいるかどうかを示す。メンバーである場合は 1、メンバーでない場合は 0

educ - 世帯主の教育レベル（年）

<u>問 1</u>

距離に基づいて切り替えの確率を予測するロジスティック回帰を実行せよ。（説明変数は距離のみ）距離 1m 離れた最も近い安全な井戸に切り替えるオッズはどのくらいか？

<u>問 2</u>

ヒ素レベルの説明変数を追加して、問 1 のモデルを更新してください。ヒ素レベルが井戸の切り替えに与える影響は？ ロジットとオッズの両方の観点から結果を解釈せよ。

問 3

更に、社会的指標である assoc と educ を追加し、モデルを更新せよ。それぞれの係数と有意性を解釈せよ 。

我々は、健康関連の意思決定における教育の効果に最も興味があると仮定する。0 年、4 年、8 年、12 年、そして 16 年の 5 つの異なる教育レベルでの切り替えの予測確率を計算し、結果を解釈せよ。その際、他の連続変数（dist と arsenic）を中央値に設定し、assoc を 1 に設定して、結果を表にせよ。

教育レベル (年)	0	4	8	12	16
切り替えの確率					

問 4

2 つの説明変数間の交互作用を提案し、それが適切である理由を説明しなさい。提案した交互作用を問 3 のモデルに追加してください。分析から、交互作用項の係数の符号に基づいて、交互作用項を含める根拠が正しいといえますか？ 交互作用項の有意性はありますか？ また、有意性の有無にかかわらず、モデル中の交互作用項の効果を解釈せよ。