データサイエンス　課題 3　変換とダミー変数

締め切り: 6 月 26 日 10:25am

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。

vote2.csv という選挙運動での支出に関するデータセットを開いてください。これは、以下の変数を含むデータセットです。

1. State – 州の郵便番号
2. Con_Dist – 議会地区
3. PCT_Vote_A – A 党候補者への投票率(単位: %)
4. EXP_A – A 党候補者の選挙運動費用(単位: 1000 ドル)
5. EXP_B – B 党候補者の選挙運動費用(単位: 1000 ドル)

問 1

A 党候補者の支出 とB 党候補者の支出が、A 党候補者の総得票数に占める割合(投票率)に与える影響を調べるために重回帰分析を実行しなさい。切片と両説明変数の係数(傾き)の推定値を解釈しなさい。

問 2

A 党候補者による選挙運動費用の二乗を含む、別の重回帰分析を実行し、モデルを推定します。A 党候補者による支出の効果が逓減することを示す証拠はありますか? その場合、A 党候補者による支出が$500,000 の時の傾きは何でしょう？$700,000 の時はどうでしょうか？

(ヒント: 回帰分析の結果をオブジェクト rg に保存した場合、rg$coefficients[3]で 3 つの推定パラメータを呼び出せます)

次に、データセット sleep2.csv を使用します。このデータセットには多くの変数が含まれていますが、ここでは以下のものだけを使用します。

1. sleep – 個人が 1 週間に眠る総時間数（分）
2. totwrk – 個人が 1 週間に働く総時間数（分）
3. male – 男性の場合 1 になるダミー変数
4. marr – 既婚の場合に 1 となるダミー変数
5. educ – 学校教育年数
6. yngkid – 3 歳以下の子供がいる場合に 1 となるダミー変数

問 3

男性(male)、既婚（marr）、週当たりの労働時間（totwrk）、学校教育年数（educ）に基づいて、個人の総睡眠時間（sleep）を予測するモデル（講義資料を参照）を書き出しなさい。モデルとデータセットでの変数の定義を考えると、基本グループは何でしょうか？ 4 つのグループのそれぞれについて、条件付き期待値を書き出しなさい。

（ヒント：R を使う必要はありません。Word の数式エディタを使うと綺麗に書けます。）

問 4

R で、問 3 のモデルを推定します。どの説明変数の係数が統計的に有意ですか？有意な説明変数の係数それぞれについて、係数の推定値が何を意味しているかを解釈してください。