データサイエンス　課題 2　単回帰分析

締め切り: 6 月 19 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。

データファイル internet.csv を読み込んでください。このデータは CIA 2010 World Factbook のもので、212 カ国の一人あたりの GDP（1 千ドル、Gdp）とインターネット利用者の人口比率（Int）に関する情報を含んでいます。ここで、GDP は物価水準の国間格差を考慮し、購買力平価に基づいています。この 2 つの変数に線形的な関連があるかどうかを調査します。特に、Gdp を用いた Int の予測がどの程度有効であるかを調べたい。

問 1

(i) インターネット利用率が最も高い国・最も低い国はどこか？

(ii) インターネット利用率の平均は何か？

(iii) どの国が最も平均値に近いか？[各国のインターネット利用率を見るのではなく、何かコードを書いて調べよ]。

(iv) Gdp と Int の関係を表す散布図を作成せよ。

問 2

(i) Gdp から Int を予測する単回帰分析を実行せよ。

(ii) Gdp の係数の推定値を問題の文脈で説明せよ。

(iii) Int のばらつき（総平方和）が Gdp によって説明される割合は？

(iv) 一人当たり GDP が 20,000 米ドルの場合、その国のインターネット利用者の割合を予測せよ。

問 3

(i) 負の残差が最も大きい国はどこか？

(ii) 正の残差が最も大きい国はどこか？

(iii) この問題の文脈において、これらの大きな正の残差と大きな負の残差は何を意味するのか述べよ。

(iv) 問 2 のモデルについて、残差と予測値（若しくは、説明変数）をプロットせよ。誤差項の仮定について、問題があるか述べよ。