data-science 課題2 単回帰分析

2323050 井上祐斗. 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？

```
# インターネット利用率が最も高い国
max_int_country ← internet_data[which.max(internet_data$Int), ]
print(paste("インターネット利用率が最も高い国:", max_int_country$Country,
"(", max_int_country$Int, "%)"))

# インターネット利用率が最も低い国
min_int_country ← internet_data[which.min(internet_data$Int), ]
print(paste("インターネット利用率が最も低い国:", min_int_country$Country,
"(", min_int_country$Int, "%)"))
```

```
[1] "インターネット利用率が最も高い国: Iceland ( 97 %)"
[1] "インターネット利用率が最も低い国: Burma ( 0.2 %)"
```

インターネット利用率が最も高い国はアイスランドであり、もっと低い国はミャンマーである。

(ii) インターネット利用率の平均は何か？

```
# インターネット利用率の平均
mean_int ← mean(internet_data$Int)
print(paste("インターネット利用率の平均:", round(mean_int, 2), "%"))
```

```
[1] "インターネット利用率の平均: 32.56 %"
```

212カ国のインターネット利用率の平均は32.56%である。

(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]。

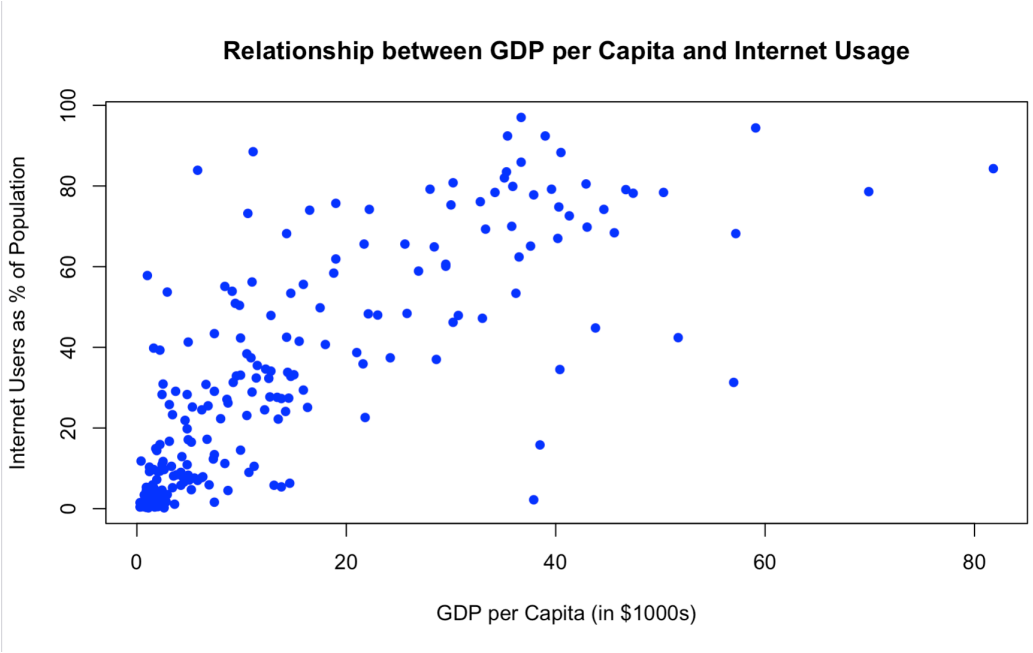
```
# インターネット利用率の平均値に近い国
closest_int_country ← internet_data[which.min(abs(internet_data$Int - mean_int)), ]
print(paste("インターネット利用率が最も平均値に近い国:", closest_int_country$Country,
"(", closest_int_country$Int, "%)"))
```

```
[1] "インターネット利用率が最も平均値に近い国: Costa Rica ( 32.4 %)"
```

最もインターネット利用率が平均値に近い国は、コスタリカである。

(iv) GdpとIntの関係を表す散布図を作成せよ。

```
# 散布図の作成
plot(internet_data$Gdp, internet_data$Int,
    xlab = "GDP per Capita (in $1000s)",
    ylab = "Internet Users as % of Population",
    main = "Relationship between GDP per Capita and Internet Usage",
    pch = 16, col = "blue")
```



x軸を1000$あたりのGDP、y軸をインターネット利用率として散布図をプロットすると上記のようになる。

問 2

(i) Gdp から Int を予測する単回帰分析を実行せよ。

```
# 単回帰分析の実行
model ← lm(Int ~ Gdp, data = internet_data)

# モデルのサマリーを表示
summary(model)
```

```
Call:
lm(formula = Int ~ Gdp, data = internet_data)

Residuals:
   Min     1Q  Median     3Q    Max
-61.742 -11.914  -3.276   9.417  63.644

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.36278    1.71827   7.195 1.09e-11 ***
Gdp          1.36093    0.07975  17.065  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.14 on 210 degrees of freedom
Multiple R-squared:  0.581,   Adjusted R-squared:  0.579
F-statistic: 291.2 on 1 and 210 DF,  p-value: < 2.2e-16
```

線形のモデルでフィッティングできると仮定して単回帰分析を行った結果は上記である。
(ii) Gdpの係数の推定値を問題の文脈で説明せよ。

Gdpの係数の推定値は `1.36093` であり、一人当たりGDPが1000$増加するごとに、その国のインターネット利用率が1.294%増加することを意味する。この値は正であり、GDPとインターネット利用率の間には正の相関関係があることを示す。t値も2以上であり統計的に有意である係数である。

(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？

```
# 決定係数を取得
r_squared ← summary(model)$r.squared
print(paste("Intのばらつきがによって説明される割合 (R-squared):",
round(r_squared * 100, 2), "%"))
```

```
[1] "Intのばらつきがによって説明される割合 (R-squared): 58.1 %"
```

決定係数を求めると`58.1`%であり、Gdpの58.1%をIntによって説明できるということである。半数近くを線形モデルで適合できているということになる。

(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。

```
# 予測用の新しいデータフレームを作成
new_gdp_data ← data.frame(Gdp = 20)

# 予測を実行
predicted_int ← predict(model, newdata = new_gdp_data)
print(paste("一人当たりGDPが20,000米ドルの場合、インターネット利用者の割合予測:",
round(predicted_int, 2), "%"))
```

```
[1] "一人当たりGDPが20,000米ドルの場合、インターネット利用者の割合予測: 39.58 %"
```

求めた予測モデルから、一人当たりGDPが20,000米ドルの場合のインターネット利用者の割合を求めると39.58 %である。

問 3

(i) 負の残差が最も大きい国はどこか？

```
# 残差の計算
residuals ← residuals(model)

# 負の残差が最も大きい国（最も負の値が大きい国）
most_negative_residual_country ← internet_data[which.min(residuals), ]
print(paste("負の残差が最も大きい国:", most_negative_residual_country$Country,
"(残差:", round(min(residuals), 2), ")"))
```

```
[1] "負の残差が最も大きい国: Equatorial Guinea (残差: -61.74 )"
```

残差を計算し、負の残差が最も大きい国を求めると、赤道ギニア共和国である。

(ii) 正の残差が最も大きい国はどこか？

```
# 正の残差が最も大きい国
most_positive_residual_country ← internet_data[which.max(residuals), ]
print(paste("正の残差が最も大きい国:", most_positive_residual_country$Country,
"(残差:", round(max(residuals), 2), ")"))
```

```
[1] "正の残差が最も大きい国: Niue (残差: 63.64 )"
```

正の残差が最も大きい国を計算すると、ニウエである。

(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。

正の残差は、予測されるインターネット利用率よりも実際のインターネット利用率が高いことを表す。つまり、何かしらのGDP以外の要因によって、インターネット普及を促進している可能性を示唆しています。例

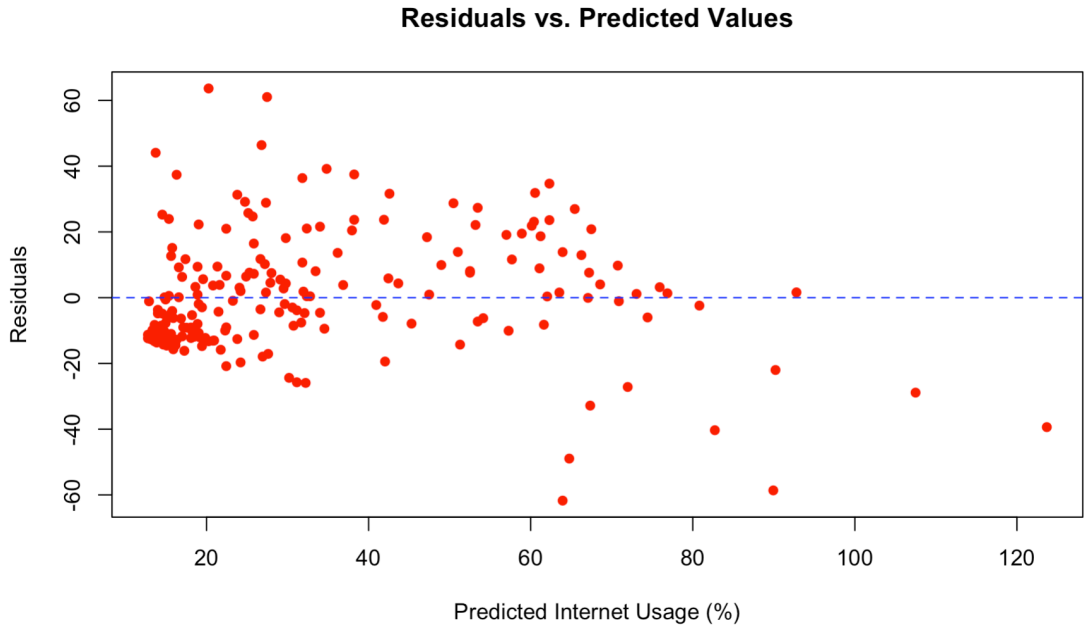
えば、国民の高いITリテラシーが高いこと、政府が積極的に推進していることなどが考えられる。

負の残差、予測されるインターネット利用率よりも実際のインターネット利用率が低いことを表す。つまり、何かしらのGDP以外の要因によって、インターネット普及を妨害している可能性を示唆しています。例えば、社会文化的要因、政府の政策による妨害などが考えらえる。

(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。

```
# 予測値の計算
predicted_values ← predict(model)

# 残差と予測値の散布図
plot(predicted_values, residuals,
    xlab = "Predicted Internet Usage (%)",
    ylab = "Residuals",
    main = "Residuals vs. Predicted Values",
    pch = 16, col = "red")
abline(h = 0, col = "blue", lty = 2)
```



誤差項の条件付き平均がゼロという仮定について、この散布図を見るとネット利用率20%、残差-20あたりで点が集中していることがわかる。これは、ネット利用率20%での誤差の総和はマイナスになることを意味する。また、ネット利用率40-80%では正の誤差の点が多く、誤差の総和は正になることが予想される。つまり、誤差が予測のネット利用率が変わるたびに偏って動いているので、仮定を満たしていない。そのため、線形性モデルで表現できるという仮定には問題がある。