統計学・データ分析勉強会／第４回宿題解答例

第 4 回宿題の解答例を以下に示します。今回の問題は、時系列データの分析をする際に陥りがちな「見せかけの回帰」について知っておくことを目的としています。

問 1

まず corp_stats という名前でデータを読み込みます。

```
setwd('/Users/stanaka/Rstats')
corp_stats <- read.csv(' 演習データ（時系列）.csv')
```

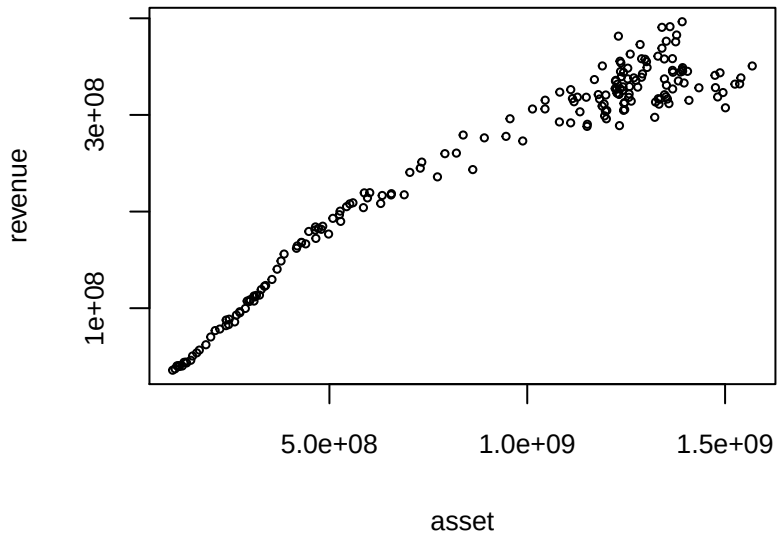
中身を確認しましょう。corp_stats は 1 列目に期間（四半期）、2 列目に総資産、3 列目に売上高が入ったデータフレームです。

```
head(corp_stats, 5)
```

```
##              期間     総資産    売上高
## 1 1970年1 - 3 月 103761211 35826142
## 2 1970年4 - 6 月 109870599 37248437
## 3 1970年7 - 9 月 114662139 40312069
## 4  1970年10-12月 118759543 40617531
## 5 1971年1 - 3 月 120910968 39666742
```

最初に、総資産（asset）を説明変数、売上高（revenue）を被説明変数として回帰分析を試みましょう。corp_stats から両データを抽出し、散布図と相関係数を表示し、回帰式を求めます。

```
asset <- corp_stats[,2]
revenue <- corp_stats[,3]
plot(asset, revenue, cex=0.5, cex.axis=0.8, cex.lab=0.8)
```



```r
cor(asset, revenue)
```

```
## [1] 0.9668794
```

```r
reg <- lm(revenue~asset)
summary(reg)
```

```
##
## Call:
## lm(formula = revenue ~ asset)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -73558700 -19102528   2233536  21738975  59773623
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.244e+07  4.318e+06   12.14   <2e-16 ***
## asset       2.189e-01  4.225e-03   51.80   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26630000 on 187 degrees of freedom
## Multiple R-squared:  0.9349, Adjusted R-squared:  0.9345
## F-statistic:  2684 on 1 and 187 DF,  p-value: < 2.2e-16
```

散布図を見ると売上高と総資産はきれいな直線に乗っており、相関係数も約 0.97 と非常に高い値を示しています。当然、回帰分析は検定をパスしており（p $<$ 2e-16 ）、決定係数も約 0.93 と申し分ありません。この結果から、ある四半期における企業の売上高は、その四半期における総資産の値によって高い精度で説明できると結論してよいでしょうか？

結論から言うと、**この分析は誤りです**。問２でその理由を考えます。

問２

今回扱っているデータセット corp_stats は、$1970 \sim 2017$ 年まで四半期ごとに企業の売上高と総資産を集計した時系列のデータでした。データを明示的に時系列データとして扱おうとするときは、ts() 関数を使って時系列オブジェクトに変換しておくと便利です。

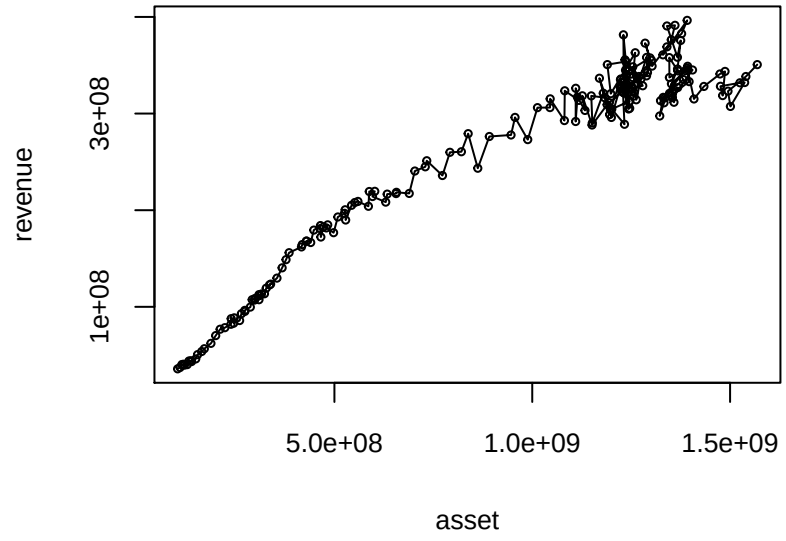
```r
asset_ts <- ts(asset, frequency = 4, start=c(1970, 1))
revenue_ts <- ts(revenue, frequency = 4, start=c(1970, 1))
```

この段階で、総資産と売上高をそれぞれグラフにしてみましょう。

```r
par(mfcol=c(1,2))
plot(asset_ts, cex=0.8, cex.axis=0.8, cex.lab=0.8)
plot(revenue_ts, cex=0.8, cex.axis=0.8, cex.lab=0.8)
```

総資産、売上高とも年（四半期）を追うごとに増加しています。さらに、先ほど描いた散布図において、各データが登場する順番に線で結んでみましょう。`plot()` 関数のオプションとして `type='o'` と指定すれば、データを時系列順に結んだ散布図を作成できます。

```
plot(revenue~asset, type='o', cex=0.5, cex.axis=0.8, cex.lab=0.8)
```



これを見ると、散布図上で互いに近接して現れるデータ点は、時期（四半期）が近い（多くの場合は隣り合っている）データであることが分かります。散布図の左下ほど昔のデータ点、右上ほど最近のデータ点を表しています。従って、売上高や総資産を説明する要因としては、時間の影響を第一に疑うべきという示唆が得られます。

勉強会の第 1 ～ 3 回では、「時間に依存せず、同一の分布から独立に抽出したデータ」を前提として、仮説検定や回帰分析などの手法を学んできました。一方、今回のような時系列データを分析するうえでも、重要な前提が存在します。それは**「データが定常性をもつ（ざっくり言うと、データの平均や分散が時間によらず常に一定）」**という条件を満たすことです。**定常性を持たない 2 つの時系列データ同士を単純に回帰分析すると、それらが互いに全く無関係の系列でも検定が有意になったり、決定係数が高くなったりといった「見せかけの回帰」という現象が生じる**ことが知られています。

上のグラフや散布図から分かるように、**今回扱っている総資産や売上高のデータは四半期を追うごとに値が増加しており、明らかに定常**

性の条件を満たしません。このため、2 つの変量を単純に回帰分析することはできません。

なお、時系列データが回帰分析や相関分析に適しているかを厳密に判定するには、データ系列に**単位根が存在するかどうか**を、**「単位根検定」**と呼ぶ方法で調べます。数学的な詳細は省きますが、「単位根が存在するデータ系列」というのは、「元の系列自体は非定常だが、一時点前との差分をとった系列は定常になるもの」を指します。前の時点のデータに今の時点でのランダムな値を加える、いわゆる「ランダムウオーク」とよばれる過程も単位根をもちます。経済や金融の分野で扱うデータはこの種の系列であることが多いため、特に注意が必要です。

単位根検定では「データ系列に単位根が存在する」という帰無仮説のもとで p 値を計算し、帰無仮説を棄却できれば定常（回帰分析などに適している）、棄却できなければ非定常（回帰分析などに適していない）という結論になります。ここでは R で特別なパッケージをインストールせずに使える `PP.test()`（フィリップス・ペロン検定）関数で、`asset_ts` と `revenue_ts` を単位根検定にかけてみましょう。ただし、データの桁数が大きくそのまま検定にかけると処理エラーが出るため、ここでは `asset_ts` と `revenue_ts` をそれぞれ 1000 で割っています。

```
PP.test(asset_ts/1000)
```

```
##
##  Phillips-Perron Unit Root Test
##
## data:  asset_ts/1000
## Dickey-Fuller = -0.93042, Truncation lag parameter = 4, p-value =
## 0.947
```

```
PP.test(revenue_ts/1000)
```

```
##
##  Phillips-Perron Unit Root Test
##
## data:  revenue_ts/1000
## Dickey-Fuller = -1.9715, Truncation lag parameter = 4, p-value =
## 0.5881
```

総資産のデータでは p 値が 0.947、売上高のデータでは p 値が 0.588 となり、いずれも有意水準 5% で帰無仮説を棄却できません。従って両系列は定常であるとは見なせず、そのまま分析すると見せかけの回帰を引き起こす恐れがあります。

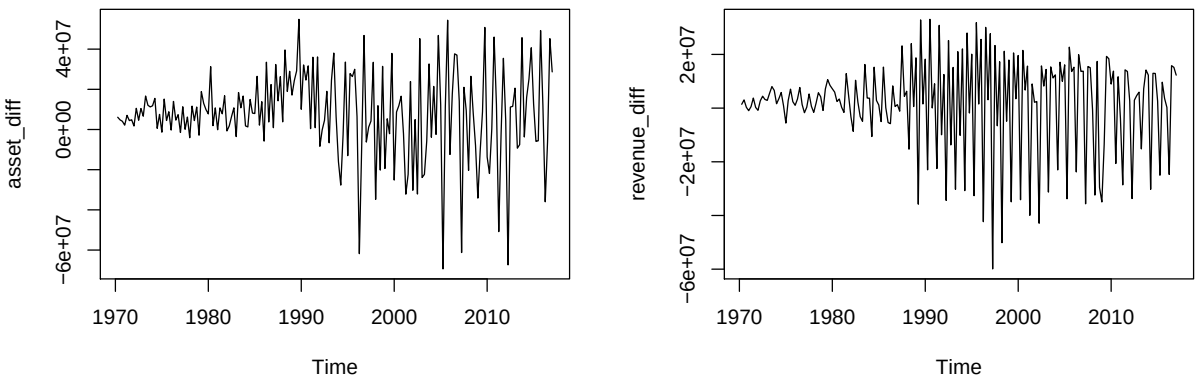
問３

見せかけの回帰の影響を回避して正しく分析を進める方法の一つは、差分系列（時系列データにおいて、１時点前との差分を取った系列）を使うことです。`diff()` 関数を使って、総資産と売上高それぞれのデータについて、１四半期前との差分を計算します。なお、時系列データの分析をもとに実際に記事を書く際は、差分の絶対値よりも前の時点からの変化率などの方が便利なケースも多いと考えられますが、ここでは最もシンプルな分析として単純な差分を使います。

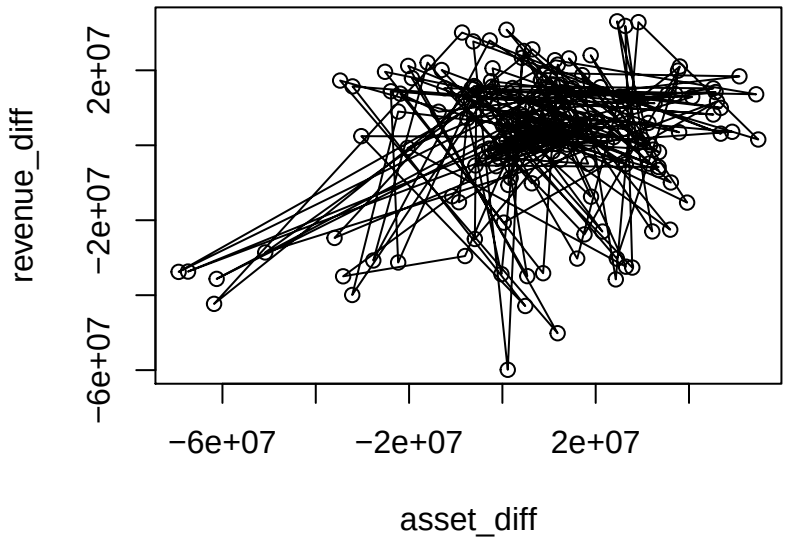
```
asset_diff = diff(asset_ts)
revenue_diff = diff(revenue_ts)
```

先ほどと同じように、それぞれの系列のグラフと散布図を作成し、相関係数を計算しましょう。

```
par(mfcol=c(1,2))
plot(asset_diff, cex=0.8, cex.axis=0.8, cex.lab=0.8)
plot(revenue_diff, cex=0.8, cex.axis=0.8, cex.lab=0.8)
```




```
plot(asset_diff, revenue_diff, type='o', cex=0.8, cex.axis=0.8, cex.lab=0.8)
```




```
cor(asset_diff, revenue_diff)
```

```
## [1] 0.2335809
```

まだ完全に平均や分散が一定とは言えませんが、先ほどと比べてかなり定常なデータと言えそうです。散布図上のデータの現れ方も、時間に強く依存せずランダムに近づきました。

`asset_diff` と `revenue_diff` は、以下のように単位根検定を有意水準 5% でパス（帰無仮説を棄却、つまり定常といえる）します。

```
PP.test(asset_diff/1000)
```

```
##
##  Phillips-Perron Unit Root Test
##
```

```
## data:  asset_diff/1000
## Dickey-Fuller = -14.442, Truncation lag parameter = 4, p-value =
## 0.01
```

```
PP.test(revenue_diff/1000)
```

```
##
##  Phillips-Perron Unit Root Test
##
## data:  revenue_diff/1000
## Dickey-Fuller = -24.825, Truncation lag parameter = 4, p-value =
## 0.01
```

一方で、相関係数をみると売上高と総資産の間の直線関係はだいぶ弱くなってしまいました。この状態でも「売上高を総資産で説明できる」と言えるでしょうか。改めて回帰分析を実行しみてましょう。

問4

差分系列で回帰分析を再度実行します。

```
reg_new <- lm(revenue_diff~asset_diff)
summary(reg_new)
```

```
##
## Call:
## lm(formula = revenue_diff ~ asset_diff)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -60331775  -5894921   1131411  11325090  31506354
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.065e+05  1.323e+06   0.156  0.87610
## asset_diff  1.884e-01  5.751e-02   3.276  0.00125 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17060000 on 186 degrees of freedom
## Multiple R-squared:  0.05456,    Adjusted R-squared:  0.04948
## F-statistic: 10.73 on 1 and 186 DF,  p-value: 0.001255
```

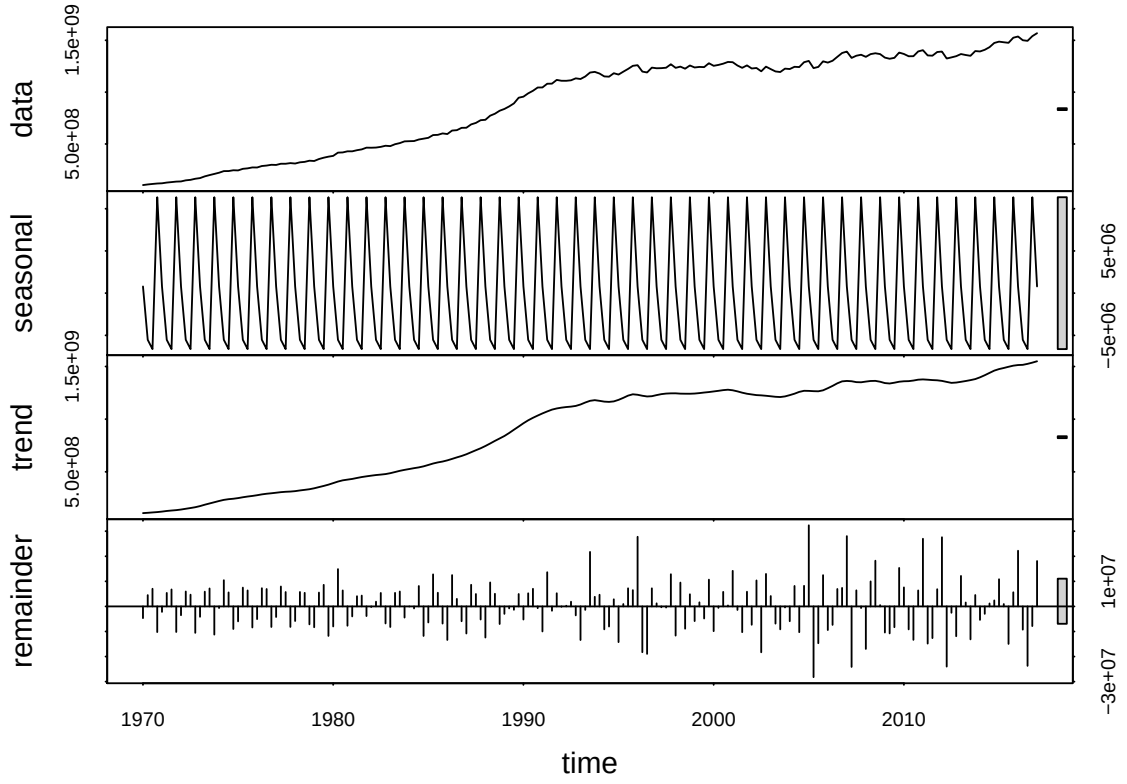
傾きの検定は有意水準 5% でパスしていますが、決定係数は 0.05 程度と非常に低い値になっています。従って、売上高の変動を総資産で説明するのは無理があるという結論になりました。残念ながら問 1 で現れた 2 変量の強い関係性は「見せかけ」であり、他のデータセットや分析手法を試す必要があります。

(補足)

上で見たように、時間の経過による増加・減少といった強いトレンドが含まれる時系列データをそのまま分析すると、見せかけの回帰のような誤った結論を導いてしまう恐れがあります。一般に時系列データには周期性（季節成分)、上昇や下落といったトレンド、それ以外の残差の 3 成分が混じっています。分析の際はそれらをまとめて扱うのではなく、元のデータを各成分に分解したうえで、自分の関心のある成分に着目する必要があります。

時系列データを 3 成分に分解する手法として、`stl()` 関数があります。試しに元々の総資産のデータ系列 `asset_ts` を分解したうえで、プロットしてみます。

```
asset_stl <- stl(asset_ts, s.window="periodic")
plot(asset_stl, cex=0.8, cex.axis=0.8, cex.lab=0.8)
```



このプロットは、上から `data`（元の系列）、`seasonal`（季節成分）、`trend`（トレンド）、`remainder`（残差）を示しています。このうち特定の成分に注目した分析をしたければ、`asset_stl` オブジェクトから以下のようにしてデータを抽出できます（ここでは残差を抽出しています）。

```
remainder <- asset_stl$time.series[,3]
```