Київський національний університет імені Тараса Шевченка
Кафедра теорії ймовірностей, статистики та актуарної математики

Р. Майборода

Самостійна робота по курсу

"Комп'ютерна статистика"

Для студентів магістратури за напрямом "статистика"

Індивідуальні завдання
та рекомендації по виконанню
Робоча версія від 06.09.2017

Київ — 2017

Вступ

Для виконання завдань потрібно встановити R та RStudio на своєму комп'ютері.

Щоб встановити R для Windows зайдіть на сторінку

`http://cran.r-project.org/bin/windows/base/`

і виберіть `Download R 3.4.1 for Windows` (номер версії, скоріше за все, буде вже іншим). Після цього запустіть програму, яка буде завантажена на ваш комп'ютер і відповідайте на її запити.

Якщо вам потрібна версія R для іншої операційної системи, зайдіть на сторінку

`http://www.r-project.org/`

і виберіть там варіант, який вас влаштовує.

Для того, щоб встановити RStudio, зайдіть на сторінку

`www.rstudio.com`

і виберіть там варіант для завантаження. Встановлювати RStudio треба після того, як буде встановлено R.

Книжку [3], присвячену статистичному аналізу даних за допомогою R, можна отримати за адресою:

`http://probability.univ.kiev.ua/userfiles/mre/compsta.pdf`

Завдання 1.

1. Отримайте файл з даними про котирування на американських фондових біржах акцій компаній, що входять до індексу S&P 500. Файл-каталог можна завантажити з сайту компанії Quantquote:

`quantquote.com/files/quantquote_daily_sp500_83986.zip`

Розпакуйте цей архів у зручний для вас каталог (теку). Дані по кожній компанії містяться в в окремому файлі. Імена файлів містять скорочені назви компаній, наприклад, `table_ibm.csv` — файл з даними про котирування акцій компанії IBM. Список компаній з їх скороченими та повними назвами і сферою їх діяльності можна подивитись тут:

`en.wikipedia.org/wiki/List_of_S%26P_500_companies`

Кожен файл містить таблицю у форматі csv з семи стовпчиків:

— дата біржових торгів (формат ррррммдд) — dat

— індикатор — z,

— ціна відкриття — opn,

— максимальне ціна — mx,

— мінімальна ціна — mn,

— ціна закриття — clo,

— обсяг продаж — vol.

2. Знайдіть дані по компаніях, що відповідають вашому варіанту: це компанії, які нумерації за скороченими назвами в алфавітному порядку мають номери від 10N-9 до 10N де N — номер вашого варіанту. (За бажанням, можна вибрати інші компанії, які вас цікавлять і узгодити список з викладачем). Виділіть відповідні файли у окремий каталог. Надалі ви будете працювати тільки з ними.

3. Для цих компаній підрахуйте логарифмічні норми прибутку з лагом 1 (log-retutns) за змінною clo;

4. Виберіть одну компанію, log-returns якої ви будете прогнозувати і використайте log-returns інших компаній з лагом -1 для прогнозування. Побудуйте прогноз використовуючи лінійну регресійну модель. Розгляньте два варіанти:

— використання всіх наявних даних крім останніх 10 сесій (повні дані);

— використання даних по 50 сесіях, що передують 10-ти останнім (останні данні).

5. Проведіть аналіз залишків та аналіз впливу у ваших моделях. Відобразіть бульбашкову діаграму впливу. При потребі, вилучіть з даних

впливові елементи і повторіть підгонку.

6. Перевірте якість прогнозу на 20-ти останніх даних, порівняйте результати підгонки моделі за повними і останніми даними.

Результати і висновки опишіть у звіті.

Рекомендації по виконанню завдання 1.

Прочитати дані з різних файлів і зібрати стовпчики clo в один файл можна наступним чином:

```
> # y filenames --- повні імена всіх файлів, що лежать у каталозі C:\\rem\\d
> filenames=list.files(path="C:\\rem\\d", full.names=TRUE)
> # читаємо файли, виймаємо 1-ший і 6-й стовпчики і кладемо в окремий фрейм
> datalist = lapply(filenames,
+ function(x){x0<-read.csv(file=x,header=F)[,c(1,6)];
+ colnames(x0)<-c("data",
+ unlist(strsplit(x,"[_.]"))[2]);# назва компанії стає назвою стовчика
+ x0})
> y<-Reduce(function(x,y) {merge(x,y,by="data")}, datalist) # зливаємо фрейми в од
```

У цьому прикладі ми розглядаємо регресію за початковими даними, log-returns не підраховуються. (Як перейти до log-returns розберіться самостійно).

Створюємо файл з даними для підгонки регресійної моделі для ціни компанії adi за даними по всіх інших компаніях:

```
> Data<-y[-nrow(y),-1]
> Data$adi<-y$adi[-1]
```

Підганяємо модель за даними по останніх 50 сесіях, виводимо таблицю результатів:

```
> nn<-nrow(Data)
> model1<-lm(adi~.-adi,data=Data[(nn-50):nn,])
> summary(model1)


Call:
lm(formula = adi ~ . - adi, data = Data[(nn - 50):nn, ])

Residuals:
     Min       1Q   Median       3Q      Max
```

```
-1.66667 -0.36419  0.01634  0.43573  1.02146

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.291750   8.911019   0.145  0.88544
aapl         0.024811   0.008355   2.970  0.00491 **
abbv         0.374864   0.200480   1.870  0.06849 .
abc          0.238507   0.119236   2.000  0.05196 .
abt          0.117824   0.225605   0.522  0.60424
ace          0.016992   0.087651   0.194  0.84722
acn         -0.095824   0.061724  -1.552  0.12806
act         -0.019595   0.044026  -0.445  0.65854
adbe         0.203669   0.147952   1.377  0.17594
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6067 on 42 degrees of freedom
Multiple R-squared:  0.8987,        Adjusted R-squared:  0.8795
F-statistic:  46.6 on 8 and 42 DF,  p-value: < 2.2e-16
```

Для графічного аналізу даних скористайтесь

```
plot(model1)
```

Для виведення бульбашкової діаграми впливу —

```
> library(car)
> influencePlot(model1)

       StudRes       Hat       CookD
111 -3.2251296 0.1120708 0.119190052
117 -2.2132836 0.2810901 0.194738756
123  0.2138565 0.3345383 0.002614004
```

(Результат див. на рис. 1)

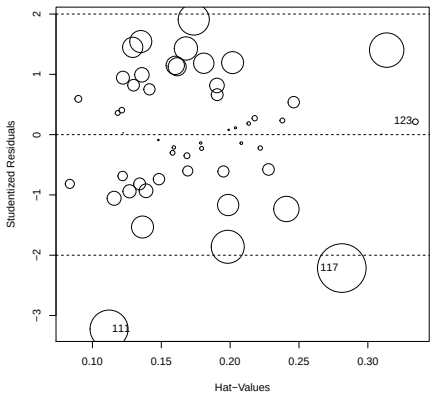

Рис. 1: Діаграма впливу

Литература

[1] Карташов М.В. "Імовірність, процеси, статистика". Київ, Видавничо-поліграфічний центр “Київський університет”, 2007, 494 с.

[2] Майборода Р.Є. Регресія: Лінійні моделі.- К. ВПЦ "Київський університет 2007, 296с.

[3] Майборода Р. Комп’ютерна статистика: професійний старт.— 2017

[4] Майборода Р.Є., Сугакова О.В. "Аналіз даних за допомогою пакета R". , 2015 65 с.

[5] Себер Дж. Линейный регрессионный анализ.— М.: Мир, 1980.— 456с.

[6] Турчин В.М. Теорія ймовірностей і математична статистика.- Дніпропетровськ, ІМА-пресс, 2014 - 566 с.

[7] James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning with Applications in R.— Springer NY 2013.— 440p.