Міністерство освіти і науки України

Черкаський державний технологічний університет

Кафедра програмного забезпечення автоматизованих систем

**ЗВІТ**

з лабораторної роботи №3

з предмету «Аналіз великих даних»

|  |  |
| --- | --- |
| Перевірив:  к. т. н., д. каф. ПЗАС  Рідкокаша А.А.  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  «\_\_\_\_» \_\_\_\_\_\_\_\_\_\_ 2019 р. | Виконав:  студент 1-го курсу  групи МПЗ-1904  Гаврилюк В. Є. |

Черкаси 2019

**Лабораторна робота №3**

**Тема:** Статистичний аналіз даних.

**Мета:** Провести статистичний аналіз згідно завдання.

**Завдання:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **1** | LinearRegression | **8** | LinearRegression | **15** | SMOreg |
| **2** | SimpleLinear | **9** | GaussianProcesses | **16** | LinearRegression |
| **3** | GaussianProcesses | **10** | SMOreg | **17** | SimpleLinear Regression |
| **4** | LinearRegression | **11** | GaussianProcesses | **18** | GaussianProcesses |
| **5** | SMOreg | **12** | SimpleLinear Regression | **19** | LinearRegression |
| **6** | GaussianProcesses | **13** | LinearRegression | **20** | SMOreg |
| **7** | SimpleLinear Regression | **14** | GaussianProcesses | **21** | GaussianProcesses Regression |

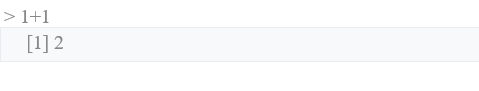
**Короткі теоретичні відомості**

R – мова програмування і програмне середовище для статистичних обчислень, аналізу та зображення даних в графічному вигляді. Розробка R відбувалась під істотним впливом двох наявних мов програмування: мови програмування S з семантикою успадкованою від Scheme. R названа за першою літерою імен її засновників Роса Іхаки (Ross Ihaka) та Роберта Джентлмена (Robert Gentleman) працівників Оклендського Університету в Новій Зеландії. Незважаючи на деякі принципові відмінності, більшість програм, написаних мовою програмування S запускаються в середовищі R.

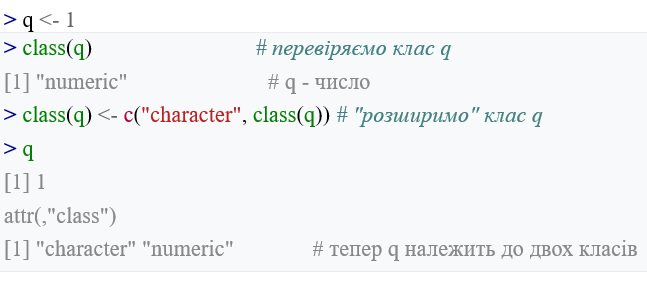
R поширюєтся безкоштовно за ліцензією GNU General Public License у вигляді вільнодоступого вихідного коду або відкомпільованих бінарних версій більшості операційних систем: Linux, FreeBSD, Microsoft Windows, Mac OS X, Solaris. R використовує текстовий інтерфейс, однак існують різні графічні інтерфейси користувача (див. Графічні Редактори Скриптів та IDE).

R має значні можливості для здійснення статистичних аналізів, включаючи лінійну і нелінійну регресію, класичні статистичні тести, аналіз часових рядів (серій), кластерний аналіз і багато іншого. R легко розбудовується завдяки використанню додаткових функцій і пакетів доступних на сайті Comprehensive R Archive Network (CRAN). Більша частина стандартних функцій R, написана мовою R, однак існує можливість підключати код написаний C, C++, або Фортраном. Також за допомогою програмного коду на C або Java можна безпосередньо маніпулювати R об'єктами.

R належить до інтерпретованих мов програмування і для роботи використовується командний інтерпретатор. Наприклад робота R в терміналі має такий вигляд:



R підтримує концепцію Об'єктно-орієнтованого програмування (ООП) включаючи generic функції, результат виконання якої залежить від аргументів (типу об'єктів), що передаються generic функції. В мові програмування R всі змінні є об'єктами, кожен об'єкт належить до певного класу. При цьому R має дві класові моделі: S3 та S4. Перша була реалізована від початку існування R, друга була додана у версії 1.7.0 з пакетом methods. S3 не є справжньою класовою системою, класи S3-об'єкта визначаються простим атрибутом — вектором символьних рядків:



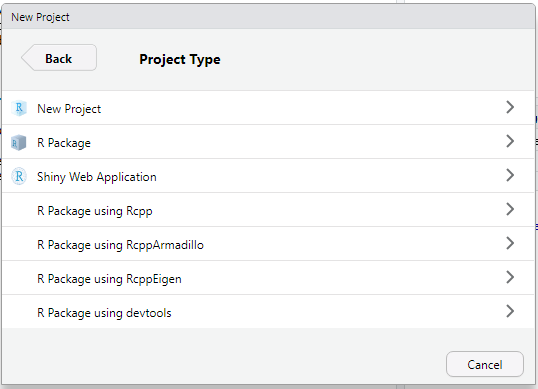
**Хід роботи**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 4.87141 | 0.00 | 18.100 | 0 | 0.6140 | 6.4840 | 93.60 | 2.3053 | 24 | 437.0 | 20.20 | 396.21 | 18.68 | 16.70 |
| 15.02340 | 0.00 | 18.100 | 0 | 0.6140 | 5.3040 | 97.30 | 2.1007 | 24 | 437.0 | 20.20 | 349.48 | 24.91 | 12.00 |
| 10.23300 | 0.00 | 18.100 | 0 | 0.6140 | 6.1850 | 96.70 | 2.1705 | 24 | 437.0 | 20.20 | 379.70 | 18.03 | 14.60 |
| 14.33370 | 0.00 | 18.100 | 0 | 0.6140 | 6.2290 | 88.00 | 1.9512 | 24 | 437.0 | 20.20 | 383.32 | 13.11 | 21.40 |
| 5.82401 | 0.00 | 18.100 | 0 | 0.5320 | 6.2420 | 64.70 | 3.4242 | 24 | 437.0 | 20.20 | 396.90 | 10.74 | 23.00 |
| 5.70818 | 0.00 | 18.100 | 0 | 0.5320 | 6.7500 | 74.90 | 3.3317 | 24 | 437.0 | 20.20 | 393.07 | 7.74 | 23.70 |
| 5.73116 | 0.00 | 18.100 | 0 | 0.5320 | 7.0610 | 77.00 | 3.4106 | 24 | 437.0 | 20.20 | 395.28 | 7.01 | 25.00 |
| 2.81838 | 0.00 | 18.100 | 0 | 0.5320 | 5.7620 | 40.30 | 4.0983 | 24 | 437.0 | 20.20 | 392.92 | 10.42 | 21.80 |
| 2.37857 | 0.00 | 18.100 | 0 | 0.5830 | 5.8710 | 41.90 | 3.7240 | 24 | 437.0 | 20.20 | 370.73 | 13.34 | 20.60 |
| 3.67367 | 0.00 | 18.100 | 0 | 0.5830 | 6.3120 | 51.90 | 3.9917 | 24 | 437.0 | 20.20 | 388.62 | 10.58 | 21.20 |
| 5.69175 | 0.00 | 18.100 | 0 | 0.5830 | 6.1140 | 79.80 | 3.5459 | 24 | 437.0 | 20.20 | 392.68 | 14.98 | 19.10 |
| 4.83567 | 0.00 | 18.100 | 0 | 0.5830 | 5.9050 | 53.20 | 3.1523 | 24 | 437.0 | 20.20 | 388.22 | 11.45 | 20.60 |
| 0.15086 | 0.00 | 27.740 | 0 | 0.6090 | 5.4540 | 92.70 | 1.8209 | 4 | 711.0 | 20.10 | 395.09 | 18.06 | 15.20 |
| 0.18337 | 0.00 | 27.740 | 0 | 0.6090 | 5.4140 | 98.30 | 1.7554 | 4 | 711.0 | 20.10 | 344.05 | 23.97 | 7.00 |
| 0.20746 | 0.00 | 27.740 | 0 | 0.6090 | 5.0930 | 98.00 | 1.8226 | 4 | 711.0 | 20.10 | 318.43 | 29.68 | 8.10 |
| 0.10574 | 0.00 | 27.740 | 0 | 0.6090 | 5.9830 | 98.80 | 1.8681 | 4 | 711.0 | 20.10 | 390.11 | 18.07 | 13.60 |
| 0.11132 | 0.00 | 27.740 | 0 | 0.6090 | 5.9830 | 83.50 | 2.1099 | 4 | 711.0 | 20.10 | 396.90 | 13.35 | 20.10 |
| 0.17331 | 0.00 | 9.690 | 0 | 0.5850 | 5.7070 | 54.00 | 2.3817 | 6 | 391.0 | 19.20 | 396.90 | 12.01 | 21.80 |
| 0.27957 | 0.00 | 9.690 | 0 | 0.5850 | 5.9260 | 42.60 | 2.3817 | 6 | 391.0 | 19.20 | 396.90 | 13.59 | 24.50 |
| 0.17899 | 0.00 | 9.690 | 0 | 0.5850 | 5.6700 | 28.80 | 2.7986 | 6 | 391.0 | 19.20 | 393.29 | 17.60 | 23.10 |
| 0.28960 | 0.00 | 9.690 | 0 | 0.5850 | 5.3900 | 72.90 | 2.7986 | 6 | 391.0 | 19.20 | 396.90 | 21.14 | 19.70 |
| 0.26838 | 0.00 | 9.690 | 0 | 0.5850 | 5.7940 | 70.60 | 2.8927 | 6 | 391.0 | 19.20 | 396.90 | 14.10 | 18.30 |
| 0.23912 | 0.00 | 9.690 | 0 | 0.5850 | 6.0190 | 65.30 | 2.4091 | 6 | 391.0 | 19.20 | 396.90 | 12.92 | 21.20 |
| 0.17783 | 0.00 | 9.690 | 0 | 0.5850 | 5.5690 | 73.50 | 2.3999 | 6 | 391.0 | 19.20 | 395.77 | 15.10 | 17.50 |
| 0.22438 | 0.00 | 9.690 | 0 | 0.5850 | 6.0270 | 79.70 | 2.4982 | 6 | 391.0 | 19.20 | 396.90 | 14.33 | 16.80 |
| 0.06263 | 0.00 | 11.930 | 0 | 0.5730 | 6.5930 | 69.10 | 2.4786 | 1 | 273.0 | 21.00 | 391.99 | 9.67 | 22.40 |
| 0.04527 | 0.00 | 11.930 | 0 | 0.5730 | 6.1200 | 76.70 | 2.2875 | 1 | 273.0 | 21.00 | 396.90 | 9.08 | 20.60 |
| 0.06076 | 0.00 | 11.930 | 0 | 0.5730 | 6.9760 | 91.00 | 2.1675 | 1 | 273.0 | 21.00 | 396.90 | 5.64 | 23.90 |
| 0.10959 | 0.00 | 11.930 | 0 | 0.5730 | 6.7940 | 89.30 | 2.3889 | 1 | 273.0 | 21.00 | 393.45 | 6.48 | 22.00 |
| 0.04741 | 0.00 | 11.930 | 0 | 0.5730 | 6.0300 | 80.80 | 2.5050 | 1 | 273.0 | 21.00 | 396.90 | 7.88 | 11.90 |

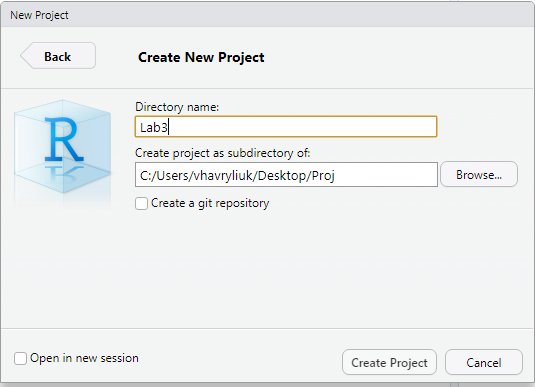
*Таблиця 1 – Набір даних для обробки.*

**Детальний опис процесу використання програмного забезпечення**

1. Першим кроком необхідно перейти за посиланням https://rstudio.com/products/rstudio/download/ та завантажити програму RStudio Desktop. Після її встановлення потрібно завантажити компілятор мови R за посиланням https://cran.rstudio.com/ та встановити його.
2. Після завантаження та встановлення всіх необхідних інструментів потрібно створити новий проект в RStudio Desktop (рис. 1). Для цього потрібно перейти в File > New Project > New Directory > New Project. Далі потрібно вказати назву директорії та власне самого проекту (рис. 2).

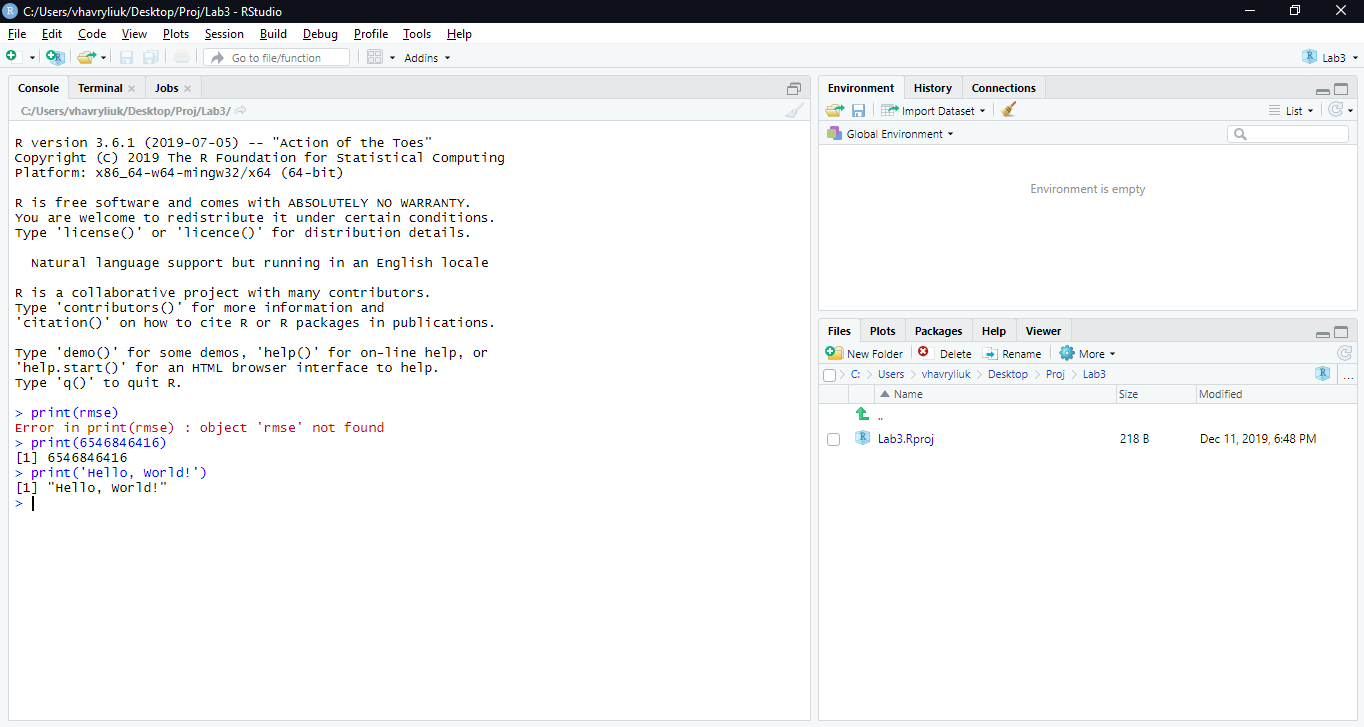


*Рис. 1 – Створення нового проекту в RStudio Desktop.*



*Рис. 2 – Вікно з параметрами нового проекту.*

1. Після створення нового проекту потрібно перейти до консолі, що розташовується зверху зліва (рис. 3).



*Рис. 3 – Консоль та інші об’єкті інтерфейсу RStudio Desktop.*

1. Після створення та налаштування нового проекту можна починати введення нових команд згідно завдання лабораторної роботи.

**Текст програми**

> #load data

> data(statisticB)

> main <- lm(Booked~.,statisticB)

> summary(main)

Call:

lm(formula = Booked ~ ., data = statisticB)

Residuals:

Min 1Q Median 3Q Max

-0.41011 -0.15767 -0.02816 0.10155 0.45539

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.482e+03 8.904e+02 -3.911 0.003560 \*\*

GNP.deflator 1.506e-02 8.492e-02 0.177 0.863141

GNP -3.582e-02 3.349e-02 -1.070 0.312681

Unbooked -2.020e-02 4.884e-03 -4.136 0.002535 \*\*

Armed.Forces -1.033e-02 2.143e-03 -4.822 0.000944 \*\*\*

Population -5.110e-02 2.261e-01 -0.226 0.826212

Year 1.829e+00 4.555e-01 4.016 0.003037 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3049 on 9 degrees of freedom

Multiple R-squared: 0.9955, Adjusted R-squared: 0.9925

F-statistic: 330.3 on 6 and 9 DF, p-value: 4.984e-10

> fit <- step(main)

Start: AIC=-33.22

Booked ~ GNP.deflator + GNP + Unbooked + Armed.Forces + Population +

Year

Df Sum of Sq RSS AIC

- GNP.deflator 1 0.00292 0.83935 -35.163

- Population 1 0.00475 0.84117 -35.129

- GNP 1 0.10631 0.94273 -33.305

<none> 0.83642 -33.219

- Year 1 1.49881 2.33524 -18.792

- Unbooked 1 1.59014 2.42656 -18.178

- Armed.Forces 1 2.16091 2.99733 -14.798

Step: AIC=-35.16

Booked ~ GNP + Unbooked + Armed.Forces + Population + Year

Df Sum of Sq RSS AIC

- Population 1 0.01933 0.8587 -36.799

<none> 0.8393 -35.163

- GNP 1 0.14637 0.9857 -34.592

- Year 1 1.52725 2.3666 -20.578

- Unbooked 1 2.18989 3.0292 -16.628

- Armed.Forces 1 2.39752 3.2369 -15.568

Step: AIC=-36.8

Booked ~ GNP + Unbooked + Armed.Forces + Year

Df Sum of Sq RSS AIC

<none> 0.8587 -36.799

- GNP 1 0.4647 1.3234 -31.879

- Year 1 1.8980 2.7567 -20.137

- Armed.Forces 1 2.3806 3.2393 -17.556

- Unbooked 1 4.0491 4.9077 -10.908

> summary(fit)

Call:

lm(formula = Booked ~ GNP + Unbooked + Armed.Forces + Year,

data = statisticB)

Residuals:

Min 1Q Median 3Q Max

-0.42165 -0.12457 -0.02416 0.08369 0.45268

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -3.599e+03 7.406e+02 -4.859 0.000503 \*\*\*

GNP -4.019e-02 1.647e-02 -2.440 0.032833 \*

Unbooked -2.088e-02 2.900e-03 -7.202 1.75e-05 \*\*\*

Armed.Forces -1.015e-02 1.837e-03 -5.522 0.000180 \*\*\*

Year 1.887e+00 3.828e-01 4.931 0.000449 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2794 on 11 degrees of freedom

Multiple R-squared: 0.9954, Adjusted R-squared: 0.9937

F-statistic: 589.8 on 4 and 11 DF, p-value: 9.5e-13

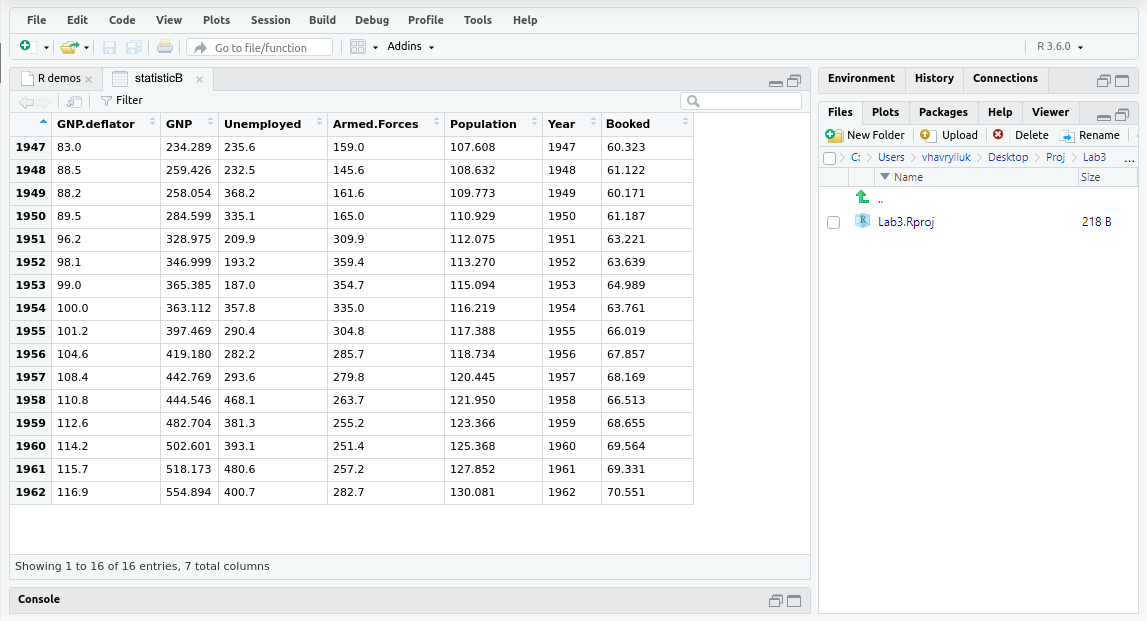
> predictions <- predict(fit, statisticB)

> rmse <- mean((statisticB$Booked - predictions)^2)

> print(rmse)

[1] 0.05366753

**Результати роботи програмного забезпечення**



**Висновок:** Під час виконання даної лабораторної роботи я провів статистичний аналіз згідно завдання. Видначив статистичні показники та побудував полігон розподілу. Ознайомився із мовою програмування R. Здобув навички планування та проведення пошукового експерименту.