Міністерство освіти і науки України

Черкаський державний технологічний університет

Кафедра програмного забезпечення автоматизованих систем

**Звіт з лабораторної роботи №1**

**з дисципліни**

“ **Аналіз великих даних**  ”

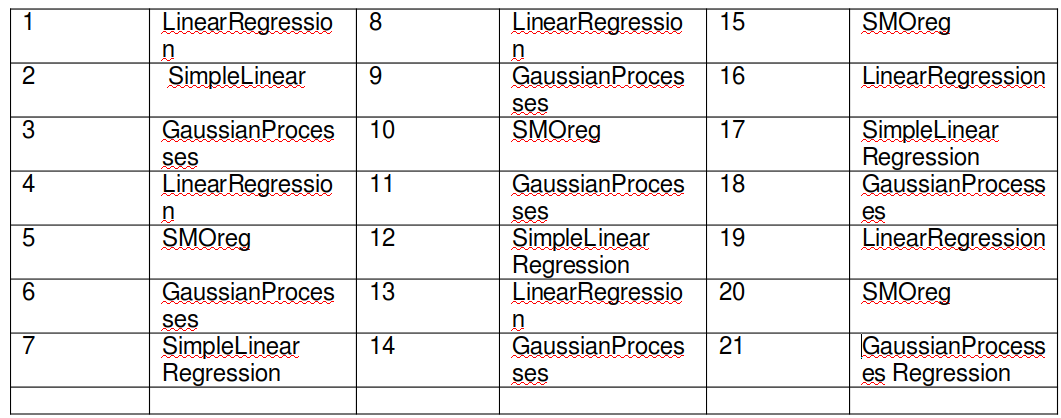
|  |  |
| --- | --- |
| Перевірив роботу: | Виконав роботу: |
| доцент | студент 1-го  курсу гр. МПЗ-1904 |
| Рідкокаша А.А.  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ | Кононенко О. В. |

Черкаси, 2019

**Тема.** Статистичний аналіз даних.

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

**Завдання:**



**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:

|  |
| --- |
| > 1+1  [1] 2 |

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:

|  |
| --- |
| **>** q <- 1  **>** class(q) *# перевіряємо клас q*  [1] "numeric" # q - число  **>** class(q) <- c("character", class(q)) *# "розширимо" клас q*  **>** q  [1] 1  attr(,"class")  [1] "character" "numeric" # тепер q належить до двох класів |

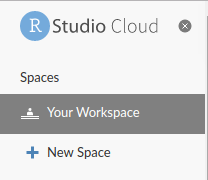
**Набір даних для обробки (якщо він великий – навести фрагмент).**

|  |
| --- |
| 0.00632 18.00 2.310 0 0.5380 6.5750 65.20 4.0900 1 296.0 15.30 396.90 4.98 24.00  0.02731 0.00 7.070 0 0.4690 6.4210 78.90 4.9671 2 242.0 17.80 396.90 9.14 21.60  0.02729 0.00 7.070 0 0.4690 7.1850 61.10 4.9671 2 242.0 17.80 392.83 4.03 34.70  0.03237 0.00 2.180 0 0.4580 6.9980 45.80 6.0622 3 222.0 18.70 394.63 2.94 33.40  0.06905 0.00 2.180 0 0.4580 7.1470 54.20 6.0622 3 222.0 18.70 396.90 5.33 36.20  0.02985 0.00 2.180 0 0.4580 6.4300 58.70 6.0622 3 222.0 18.70 394.12 5.21 28.70  0.08829 12.50 7.870 0 0.5240 6.0120 66.60 5.5605 5 311.0 15.20 395.60 12.43 22.90  0.14455 12.50 7.870 0 0.5240 6.1720 96.10 5.9505 5 311.0 15.20 396.90 19.15 27.10  0.21124 12.50 7.870 0 0.5240 5.6310 100.00 6.0821 5 311.0 15.20 386.63 29.93 16.50  0.17004 12.50 7.870 0 0.5240 6.0040 85.90 6.5921 5 311.0 15.20 386.71 17.10 18.90  0.22489 12.50 7.870 0 0.5240 6.3770 94.30 6.3467 5 311.0 15.20 392.52 20.45 15.00  0.11747 12.50 7.870 0 0.5240 6.0090 82.90 6.2267 5 311.0 15.20 396.90 13.27 18.90  0.09378 12.50 7.870 0 0.5240 5.8890 39.00 5.4509 5 311.0 15.20 390.50 15.71 21.70  0.62976 0.00 8.140 0 0.5380 5.9490 61.80 4.7075 4 307.0 21.00 396.90 8.26 20.40  0.63796 0.00 8.140 0 0.5380 6.0960 84.50 4.4619 4 307.0 21.00 380.02 10.26 18.20  0.62739 0.00 8.140 0 0.5380 5.8340 56.50 4.4986 4 307.0 21.00 395.62 8.47 19.90  1.05393 0.00 8.140 0 0.5380 5.9350 29.30 4.4986 4 307.0 21.00 386.85 6.58 23.10  0.78420 0.00 8.140 0 0.5380 5.9900 81.70 4.2579 4 307.0 21.00 386.75 14.67 17.50  0.80271 0.00 8.140 0 0.5380 5.4560 36.60 3.7965 4 307.0 21.00 288.99 11.69 20.20  0.72580 0.00 8.140 0 0.5380 5.7270 69.50 3.7965 4 307.0 21.00 390.95 11.28 18.20  1.25179 0.00 8.140 0 0.5380 5.5700 98.10 3.7979 4 307.0 21.00 376.57 21.02 13.60 |

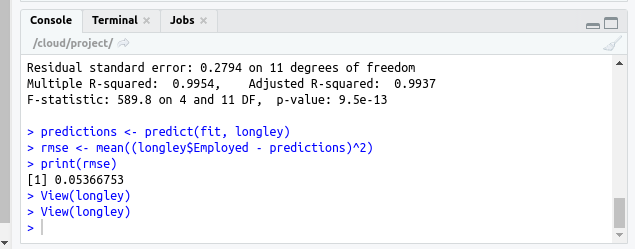
**Детальний опис процесу використання програмного забезпечення**

1) Першим кроком необхідно перейти за посиланням https://rstudio.cloud

2) Створити новий проект.



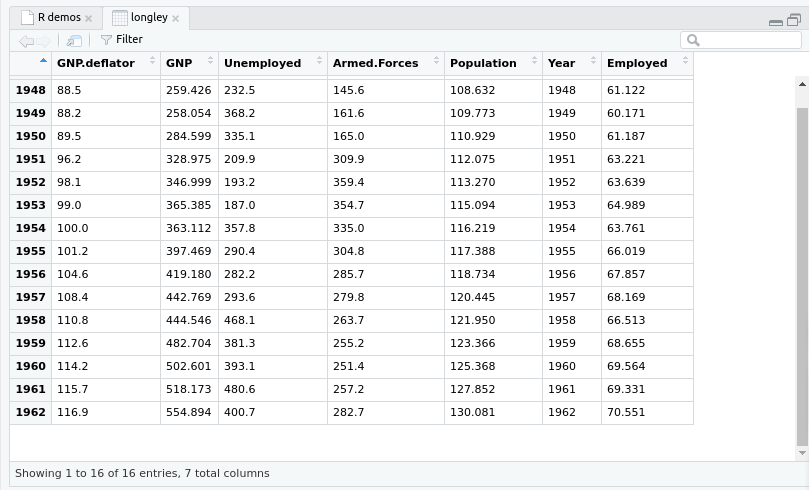
3) Перейти до консолі

4) Почати вводити команди.

**Текст програми**

|  |
| --- |
| R version 3.6.0 (2019-04-26) -- "Planting of a Tree"  Copyright (C) 2019 The R Foundation for Statistical Computing  Platform: x86\_64-pc-linux-gnu (64-bit)  R is free software and comes with ABSOLUTELY NO WARRANTY.  You are welcome to redistribute it under certain conditions.  Type 'license()' or 'licence()' for distribution details.  R is a collaborative project with many contributors.  Type 'contributors()' for more information and  'citation()' on how to cite R or R packages in publications.  Type 'demo()' for some demos, 'help()' for on-line help, or  'help.start()' for an HTML browser interface to help.  Type 'q()' to quit R.  > demo()  > print('asd')  [1] "asd"  > print('asd','asdasd')  Error in print.default("asd", "asdasd") : invalid 'digits' argument  In addition: Warning message:  In print.default("asd", "asdasd") : NAs introduced by coercion  > print('asd,asdasd')  [1] "asd,asdasd"  > #load data  > data(longley)  > base <- lm(employed~.,longley)  Error in eval(predvars, data, env) : object 'employed' not found  > base <- lm(Employed~.,longley)  > summary(base)  Call:  lm(formula = Employed ~ ., data = longley)  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  Unemployed -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(base)  Start: AIC=-33.22  Employed ~ GNP.deflator + GNP + Unemployed + 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  - Unemployed 1 1.59014 2.42656 -18.178  - Armed.Forces 1 2.16091 2.99733 -14.798  Step: AIC=-35.16  Employed ~ GNP + Unemployed + 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  - Unemployed 1 2.18989 3.0292 -16.628  - Armed.Forces 1 2.39752 3.2369 -15.568  Step: AIC=-36.8  Employed ~ GNP + Unemployed + 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  - Unemployed 1 4.0491 4.9077 -10.908  > summary(fit)  Call:  lm(formula = Employed ~ GNP + Unemployed + Armed.Forces + Year,  data = longley)  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 \*  Unemployed -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, longley)  > rmse <- mean((longley$Employed - predictions)^2)  > print(rmse)  [1] 0.05366753  > View(longley)  > View(longley) |

**Результати роботи програмного забезпечення**



**Висновок**: Під час виконання роботи було здобуто навички планування та проведення пошукового експерименту, визначення статистичних показників, побудова полігону розподілу і гістограми.

Висновок: В ході виконання лабораторної роботи було проведено статистичний аналіз згідно завдання. Було здобуто базові навички роботи з мовою R.