Лабораторна робота №1
з комп'ютерної статистики

Горбунов Даніел Денисович
1 курс магістратури
група "Прикладна та теоретична статистика"

24 вересня 2021 р.

1 Вступ.

У даній роботі були побудовані моделі лінійної регресії різних специфікацій, в залежності від початкових даних, за якими робиться підгонка. Конкретніше, задача полягала у прогнозуванні ціни закриття обраної компанії у наступній біржовій сесії на основі цін інших дев'яти за попередні сесії. Щодо підгонки моделі, розглянуто два варіанти: з урахуванням усіх сесій, що доступні для опрацювання, та за останніми п'ятдесятьма сесіями. Діагностика моделей проводилася стаднартними методами, зокрема графічними. Вперше використовується бульбашкова діаграма залишків для виявлення впливових спостережень. В кінці буде показано наскільки хорошим є прогноз для цих моделей на двадцяти майбутніх сесіях (майбутніх в тому сенсі, що йдуть після сесій, за даними яких робилася підгонка).

2 Хід роботи.

2.1 Збір даних.

Серед компаній, що входять до індексу $S\&P\,500$, були обрані десять таких, порядковий номер яких в алфавітному порядку за зростанням за кодовою назвою лежав в межах від 95 до 104.

Зауваження. Номери обчислювалися за попередньою формулою:

$$100 + N - 9, \ldots, 100 + N, \text{ де } N = 4,$$

В якості змінної для прогнозування були обрані ціни закриття компанії "Colgate-Palmolive Company" (або, за кодовою назвою, *cl*) на одну сесію вперед. Покажемо отриману таблицю:

```
> head(x.clo.lagged, n=2)
       cl     clf     clx     cma   cmcsa     cme   cmg     cmi     cms     cnp
1 23.4417 23.5836 46.6629 45.1660 16.8496 68.3886 43.80 21.1160 11.8006 8.99742
2 23.1904 23.6216 47.0381 45.1820 16.9941 67.3787 42.17 22.3403 11.7450 9.11007
```

Рис. 1: Таблиця з цінами закриття десяти компаній на фондових біржах. Для змінної *cl* ціни зафіксовані з одиничним зсувом по часу вперед.

Загальна кількість рядків у таблиці: $N = 1897 - 20 = 1877$. Кількість регресорів $d = 9$.

2.2 Підгонка регресійної моделі.

Для обох ситуацій ми будемо відштовхуватися від гауссової регресійної моделі з урахуванням усіх змінних, що наявні в таблиці, тобто:

$$Y^{cl} = X\beta + \varepsilon,\ \varepsilon \sim \mathcal{N}(\vec{0}, \sigma^2 I_N),$$

де $X = (X_1, \ldots, X_N)^T$, $X_j = (1, X_j^{clf}, \ldots, X_j^{cnp})$, $\beta \in \mathbb{R}^{d+1}$, ε - вектор похибок, σ^2 - невідома дисперсія похибок, I_N - одинична матриця розмірності N.

Зміни, які будуть стосуватися специфікації регресійної моделі, будуть базуватися на результатах графічної діагностики та тестів Стьюдента, Фішера для загальної лінійної гіпотези. Надалі зафіксуємо стандартний рівень значущості для усіх тестів рівним $\alpha = 0.05$.

2.2.1 Підгонка за всіма сесіями.

У даному підпункті звітується спроба підгонки моделі за всіма сесіями. Переходячи до суті, маємо:

```
Call:
lm(formula = cl ~ ., data = x.clo.lagged)

Residuals:
    Min      1Q  Median      3Q     Max
-4.7950 -1.4826 -0.2205  1.3384  5.8725

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.539834   0.731029   4.842 1.39e-06 ***
clf          0.005567   0.005028   1.107    0.268
clx          0.457615   0.020289  22.555  < 2e-16 ***
cma         -0.328800   0.009136 -35.990  < 2e-16 ***
cmcsa       -0.015184   0.020034  -0.758    0.449
cme          0.055832   0.004491  12.433  < 2e-16 ***
cmg         -0.011008   0.001456  -7.561 6.23e-14 ***
cmi          0.031828   0.007523   4.231 2.44e-05 ***
cms          0.855926   0.070893  12.074  < 2e-16 ***
cnp          0.105331   0.066226   1.590    0.112
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.113 on 1867 degrees of freedom
Multiple R-squared:  0.9492,Adjusted R-squared:  0.949
F-statistic:  3877 on 9 and 1867 DF,  p-value: < 2.2e-16
```

На рівні значущості α, тест Стьюдента для змінних clf, $cmcsa$ та cnp приймає основну гіпотезу про незначущу відмінність відповідних коефіцієнтів від нуля. Тому має місце вилучення цих змінних з специфікації моделі. Звіт про результати підгонки моделі після вилучення невпливових об'єктів наводиться на наступній сторінці.

```
Call:
lm(formula = cl ~ . - clf - cmcsa - cnp, data = x.clo.lagged)

Residuals:
    Min      1Q  Median      3Q     Max
-5.0307 -1.4154 -0.2331  1.3170  5.8851

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.433487   0.596766   7.429 1.65e-13 ***
clx          0.448858   0.018505  24.256  < 2e-16 ***
cma         -0.334592   0.008390 -39.878  < 2e-16 ***
cme          0.058149   0.004268  13.625  < 2e-16 ***
cmg         -0.010417   0.001352  -7.706 2.09e-14 ***
cmi          0.040524   0.004221   9.599  < 2e-16 ***
cms          0.881371   0.046508  18.951  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.116 on 1870 degrees of freedom
Multiple R-squared:  0.949,Adjusted R-squared:  0.9488
F-statistic:  5798 on 6 and 1870 DF,  p-value: < 2.2e-16
```

Отримали підозріло хорошу модель: розкид оцінок коефіцієнтів менший, значущу відмінність коефіцієнтів від нуля приймається тестом Стьюдента, а тест Фішера приймає гіпотезу про наявність залежності ціни закриття cl від взятих регресорів. Коефіцієнт детермінації виявився високим. Проблеми виникають під час діагностики моделі: емпіричний розподіл є зкошеним вліво, а емпіричні квантилі на кінцях суттєво відрізняються від теоретичних. Припущення про нормальну розподіленість порушується.

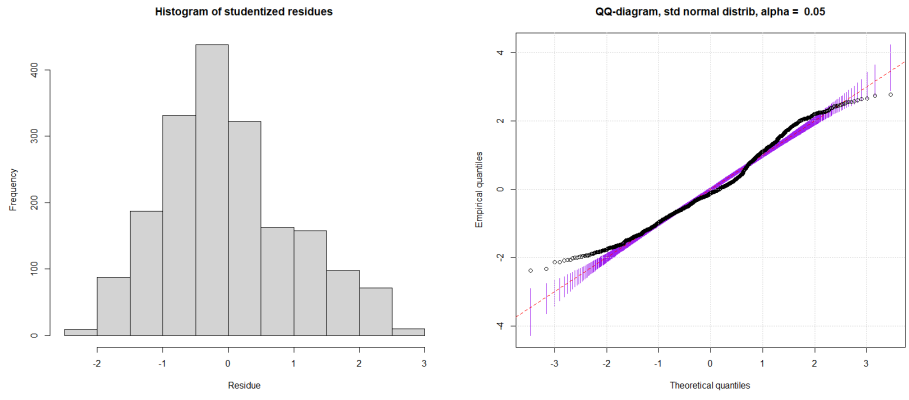

Рис. 2: Гістограма відносних частот зліва та QQ-діаграма (з прогнозними інтервала вірогідного рівня $1 - \alpha$) справа для стьюдентизованих залишків.

Подивимося на діаграми ”прогноз-відгук” та ”прогноз-залишки” для можливого виявлення інших аномалій.

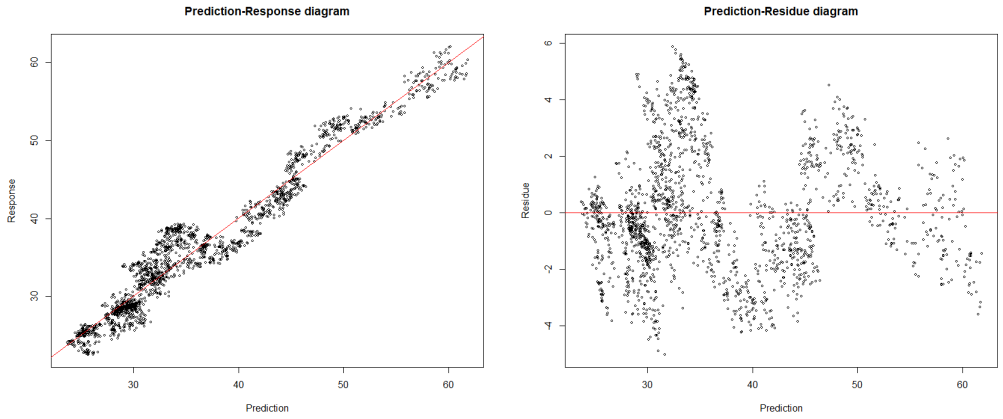

Рис. 3: Зліва – діаграма ”прогноз-відгук”. Справа – діаграма ”прогноз-залишки”.

На першій діаграмі видно, що точки розміщуються вздовж бісектриси першого координатного кута, що свідчить про відносно непоганий прогноз. А друга діаграма ”прогноз-залишки” зовсім не схожа на діаграму розкиду, більше на ”Дев’ятий вал” Айвазовського. Точніше кажучи, можна зауважити, що розкид залишків нерівномірний, тому в моделі наявна гетерескедастичність похибок, що знову порушує класичні припущення гауссової моделі. Також можна розгледіти певну нелінійну (у певному сенсі хаотичну) залежність. Останні факти свідчать про те, що з моделі витягнуто недостатньо інформації з регресорів під час розробки моделі.

На цьому моменті постає питання щодо адекватності прогнозу, що може бути отриманий за цією моделлю. Тут обмежилися лише базовими техніками, щось нове спробуємо в наступному підпункті.

2.2.2 Підгонка за "свіжими" даними.

Тепер звітується результат підгонки за останніми сесіями, що наявні у таблиці, для побудови регресійної моделі. Тут $N := 50$.

```
Call:
lm(formula = cl ~ ., data = x.recent)

Residuals:
     Min       1Q   Median       3Q      Max
-1.32442 -0.37800  0.05435  0.34444  1.10140

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.447753   7.545098   1.385 0.173818
clf         -0.034423   0.111552  -0.309 0.759238
clx          0.541754   0.127523   4.248 0.000125 ***
cma          0.165082   0.109934   1.502 0.141041
cmcsa       -0.317211   0.155519  -2.040 0.048024 *
cme         -0.205070   0.039886  -5.141 7.54e-06 ***
cmg          0.003399   0.019654   0.173 0.863571
cmi         -0.043592   0.035077  -1.243 0.221192
cms         -1.048459   0.405778  -2.584 0.013530 *
cnp          2.408660   0.702782   3.427 0.001424 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5925 on 40 degrees of freedom
Multiple R-squared:  0.8946,Adjusted R-squared:  0.8709
F-statistic: 37.71 on 9 and 40 DF,  p-value: < 2.2e-16
```

За тестом Стьюдента, лише коефіцієнти при змінних $clx, cmcsa, cme, cms$ та cnp значущо відрізняються від нуля при заданому α. Враховуючи попередні результати тесту Стьюдента та незначну кількість спостережень, врахуємо лише найбільш значущі змінні, а саме clx, cme та cnp. Коефіцієнт зсуву у цій версії не вводиться. Тому $d = 3$. Результат підгонки можна побачити далі:

```
Call:
lm(formula = cl ~ . - 1, data = x.remained)

Residuals:
     Min       1Q   Median       3Q      Max
-1.98895 -0.36633  0.09251  0.44590  1.08135

Coefficients:
    Estimate Std. Error t value Pr(>|t|)
cnp  0.84365    0.31278   2.697  0.00968 **
clx  0.53042    0.08763   6.053 2.24e-07 ***
cme -0.08182    0.01264  -6.473 5.16e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6431 on 47 degrees of freedom
Multiple R-squared:  0.9999,Adjusted R-squared:  0.9999
F-statistic: 1.398e+05 on 3 and 47 DF,  p-value: < 2.2e-16
```

Коефіцієнт детермінації зависокий, цьому не хочеться довіряти. Оскільки спостережень відносно мало, тому для перевірки нормальності лише скористємося квантильною діаграмою: дещо дивує, що в даному разі все гаразд, точки лежать у межах прогнозних інтервалів (того самого рівня, що і в попередньому випадку).

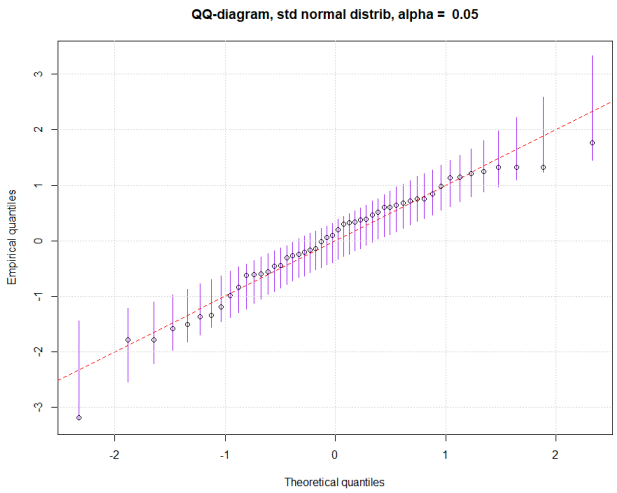

Рис. 4: Квантильна діаграма залишків моделі, з прогнозними інтервалами.

Далі побудуємо діаграми ”прогноз-відгук” та ”прогноз-залишки”.

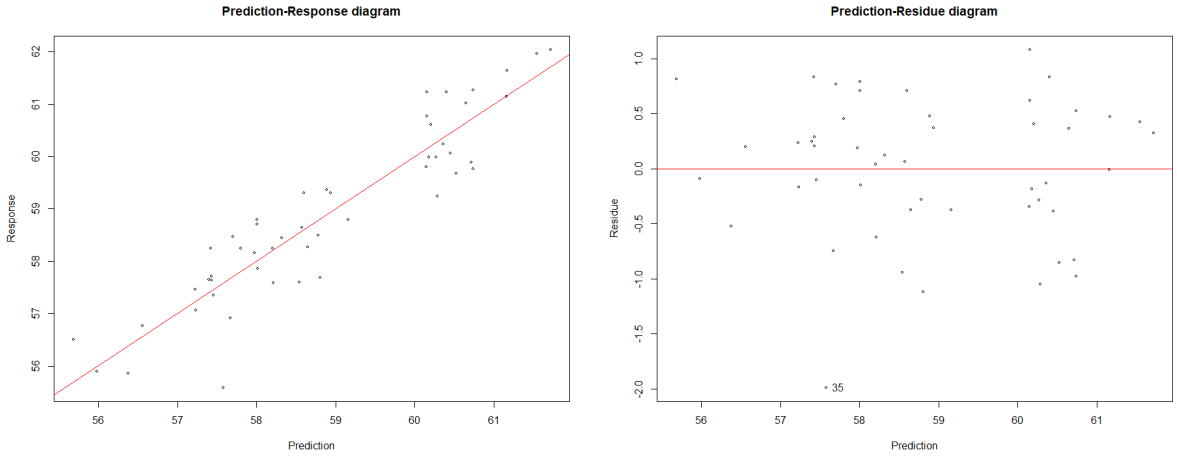

Рис. 5: Зліва – діаграма "прогноз-відгук". Справа – діаграма "прогноз-залишки".

На першій діаграмі точки відносно прямої вишукуються нормально. Порівняно з попереднім випадком, закономірності з другої діаграми виявити важче. Крім того, у першому наближенні, розкид похибок можна вважати однорідним. Хоча є одне підозріле спостереження під номером 35, залишок для якої дещо більший, порівняно з іншими значеннями залишків. Спробуємо виявити впливові спостереження за допомогою бульбашкової діаграми. Маємо таку картину:

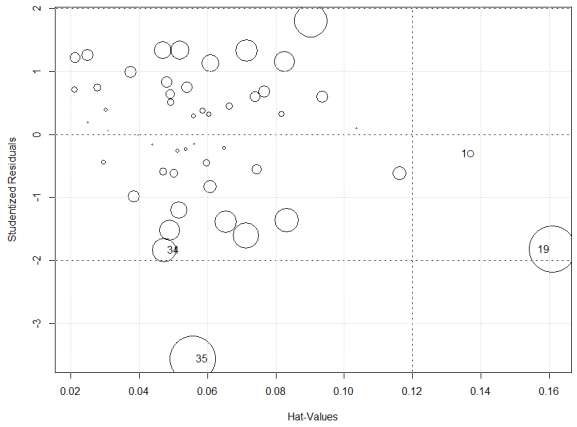

Рис. 6: Бульбашкова діаграма для стьюдентизованих залишків поточної моделі.

	StudRes	Hat	CookD
1	-0.3057095	0.13697789	0.00504178
19	-1.8241825	0.16096395	0.20275464
34	-1.8315150	0.04722150	0.05277389
35	-3.5552029	0.05570188	0.19919393

На діаграмі відмічені так звані підозрілі спостереження, які мають велике значення важеля h_j (j-ий елемент $H = X(X^TX)^{-1}X^T$) або відстані Кука D_j. У даній роботі ми скористаємося правилом для виявлення впливових спостережень за важелем [**?**]: якщо $h_j > 2\frac{\mathrm{tr}H}{N}$, то j-те спостереження вважаємо впливовим. Обчисливши середнє арифметичне діагональних

елементів ортопроектора H, маємо, що спостереження 1 та 19 є впливовими:

$$2 \cdot \frac{\mathrm{tr}H}{N} = 2 \cdot 0.06 = 0.12 < h_{j_0},\ j_0 \in \{1, 19\}$$

Для спостережень з таблиці, що залишилися, бачимо високе значення відстані Кука (тим не менш, і для 19-го спостереження). У попередньо загаданій роботі [?], підхід до вилучення підозрілих точок наступний: у гауссовій лінійній регресійній моделі, знаючи оцінки методу найменших квадратів $\hat{\beta}$, можна було побудувати довірчий еліпсоїд для вектора коефіцієнтів моделі вірогідного рівня $1 - \alpha$:

$$\mathcal{E}_\alpha = \left\{ \beta \in \mathbb{R}^p \mid (\beta - \hat{\beta})^T A (\beta - \hat{\beta}) \leq d\hat{\sigma}_0^2 Q^{F(d,N-d)}(\alpha) \right\},$$

де $\hat{\sigma}_0^2 = \dfrac{||U||^2}{N-d}$. Пропонується вилучати такі j-ті спостереження, для яких МНК-оцінки без урахування даного спостереження $\hat{\beta}_{-j}$ не належать еліпсоїду $\mathcal{E}_\alpha$. Тобто, у термінах відстані Кука, остання умова перезапишеться у такій формі:

$$\hat{\beta}_{-j} \notin \mathcal{E}_\alpha \Leftrightarrow D_j = \frac{(\hat{\beta}_{-j} - \hat{\beta})^T A (\hat{\beta}_{-j} - \hat{\beta})}{d\hat{\sigma}_0^2} > Q^{F(d,N-d)}(\alpha)$$

Нагадаємо, що $\alpha = 0.05$, тому $Q^{F(d,N-d)}(\alpha) = 0.01496638$. Для спостережень $19, 34, 35$, їхні відстані Кука перевищують заданий поріг. Узагальнюючи, ми спробуємо підігнати модель без вказаних чотирьох спостережень, на основі останньої специфікації. Маємо результат:

```
Call:
lm(formula = cl ~ . - 1, data = x.removed)

Residuals:
     Min       1Q   Median       3Q      Max
-1.20178 -0.30810  0.04566  0.37904  0.81135

Coefficients:
    Estimate Std. Error t value Pr(>|t|)
cnp  0.39052    0.29259   1.335    0.189
clx  0.65225    0.08148   8.005 4.67e-10 ***
cme -0.07425    0.01090  -6.809 2.43e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5267 on 43 degrees of freedom
Multiple R-squared:  0.9999,Adjusted R-squared:  0.9999
F-statistic: 1.922e+05 on 3 and 43 DF,  p-value: < 2.2e-16
```

Вилучаємо з моделі змінну *cnp*. Спробуємо ще раз:

```
Call:
lm(formula = cl ~ . - cnp - 1, data = x.removed)

Residuals:
     Min       1Q   Median       3Q      Max
-1.34369 -0.32160 -0.02919  0.41042  0.75205

Coefficients:
     Estimate Std. Error t value Pr(>|t|)
clx  0.760340   0.009076  83.773  < 2e-16 ***
cme -0.072650   0.010934  -6.644 3.81e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5314 on 44 degrees of freedom
Multiple R-squared:  0.9999,Adjusted R-squared:  0.9999
F-statistic: 2.832e+05 on 2 and 44 DF,  p-value: < 2.2e-16
```

Наявність виского коефіцієнта детермінації все ще залишається проблемою. Детальніше про це у висновках, але зараз переходимо до прогнозування.

2.3 Прогнозування.

Прогноз цін закриття *cl* будемо робити на основі трьох моделей: перша базується на повних даних з вилученими за Стьюдентом об'єктами, друга та третя базуються на скорочену запасі даних, відмінність між ними полягає у вилучених спостережень після аналізу впливу. Спостерігається досить сумна ситуація із залишками прогнозу для різних моделей. Серед найгірших, краще за все показали моделі, підігнані за невеликою кількістю данних. Залишки першої моделі виходять неадекватними. Це можна пояснити тим, що умови, на якій ця модель була підігнана, порушувалися. З іншого боку, є підозра, що такий невадлий результат є наслідком перепідгонки моделі. Залишки останньої моделі обмежуються двійкою, що не можна сказати для другої моделі.

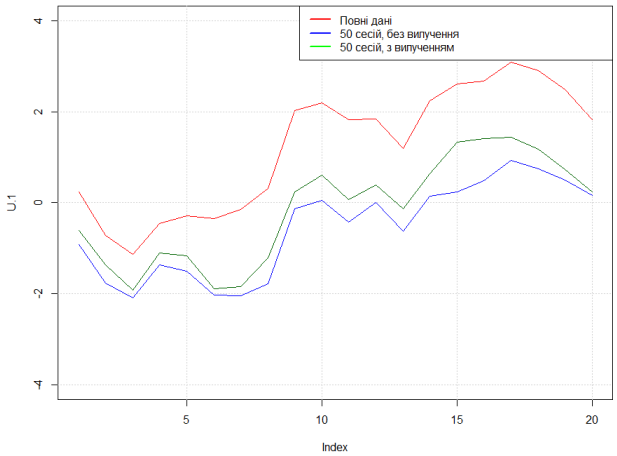

Рис. 7: Значення залишків прогнозу для трьох різних моделей.

3 Висновки.

Підхід до прогнозування цін закриття компаній на фондовій біржі за допомогою класичної лінійної регресії виявився невдалим. З усього аналізу стало зрозуміло, що урахування всіх даних привело до схожого ефекту як перепідгонка: тобто якщо прогнозування було нормальним на тренувальних даних, а на тестових виходили набагато гірші результати. Моделі з підозріло високим коефіцієнтом детермінації проявили себе непогано у прогнозуванні цін на короткий термін часу.

Загалом, побудовані моделі не підійдуть для отримання хороших прогнозів у даній задачі навіть на невеликі часові проміжки. Результати діагностики різних моделей вказували на те, що інформацію про підгонку витягнуто не повністю. Які наступні кроки будуть зроблені для розв'язання поставленої задачі - про це вже в наступній лабораторній роботі.

Література

[1] George A. F. Seber, Alan J. Lee (auth.) - Linear Regression Analysis, Second Edition (2003)