Самостійна робота №1
з алгоритмів машинного навчання

Горбунов Даніел Денисович
3 курс бакалаврату
група 'комп'ютерна статистика'

29 березня 2020 р.

Варіант 4

Задача. Необхідно провести регресійний аналіз за даними в таблиці 'NAICExpense.csv'. Обрані регресори за такими колонками: 'RBC', 'STAFFWAGE', 'AGENTWAGE', 'LONGLOSS', 'SHORTLOSS'. Відгук - вектор за колонкою 'EXPENSES'.

Аналіз початкових даних. Перш ніж починати регресійний аналіз, спробуємо розібратися з даними. Покажемо значення статистик середнього положення, квартилі та мінімальні, максимальні елементи для кожного об'єкта.

	EXPENSES	RBC	STAFFWAGE	AGENTWAGE	LONGLOSS	SHORTLOSS
Min.	-0.002038	0.0000	51.73	47.47	-0.0706227	-0.0029507
1st Qu.	0.001689	0.6425	80.06	74.81	0.0000028	0.0002786
Median	0.009024	2.8366	84.36	78.77	0.0020900	0.0044799
Mean	0.044948	22.1085	87.26	80.15	0.0257871	0.0385088
3rd Qu.	0.030137	11.0154	93.82	85.44	0.0120320	0.0223448
Max.	1.236946	838.7967	137.48	126.17	0.8539152	1.1710587

Бачимо широкий розмах даних з колонок 'RBC', 'STAFFWAGE', 'AGENTWAGE', порівняно з іншими. Для 'EXPENSES', 'RBC', 'LONGLOSS' та 'SHORTLOSS' бачимо громіздкі додатні значення коефіцієнта скошеності.

	EXPENSES	RBC	STAFFWAGE	AGENTWAGE	LONGLOSS	SHORTLOSS
CV	2.753945	3.308757	0.1366994	0.1135816	3.300994	3.191366
Range	1.238984	838.7967	85.75407	78.69901	0.9245379	1.174009
Skewness	6.327171	6.993754	0.79443	0.2224716	6.265785	6.404269

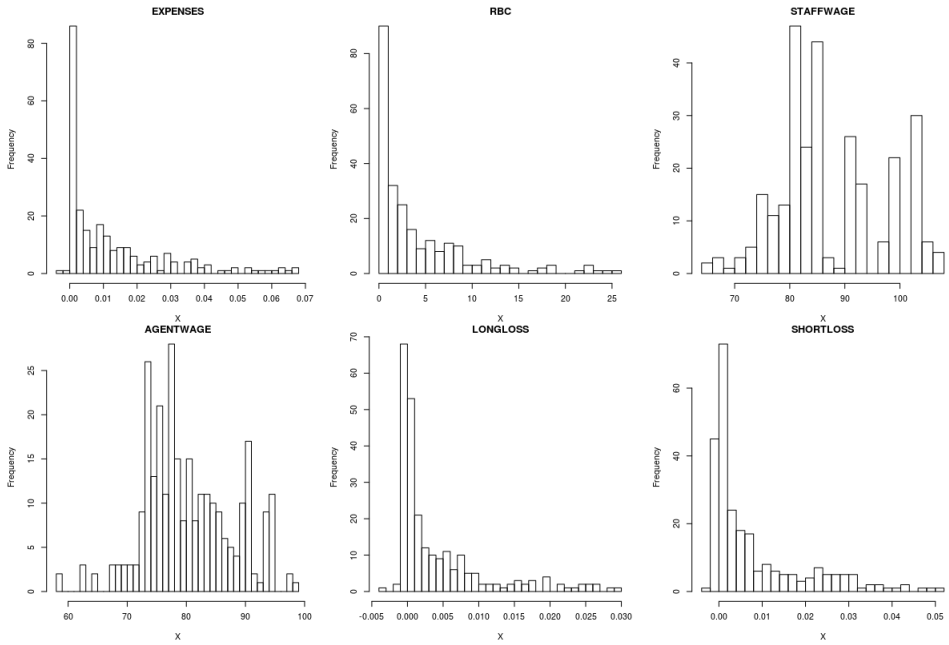

Рис. 1: Гістограми даних регресорів та відгуку.

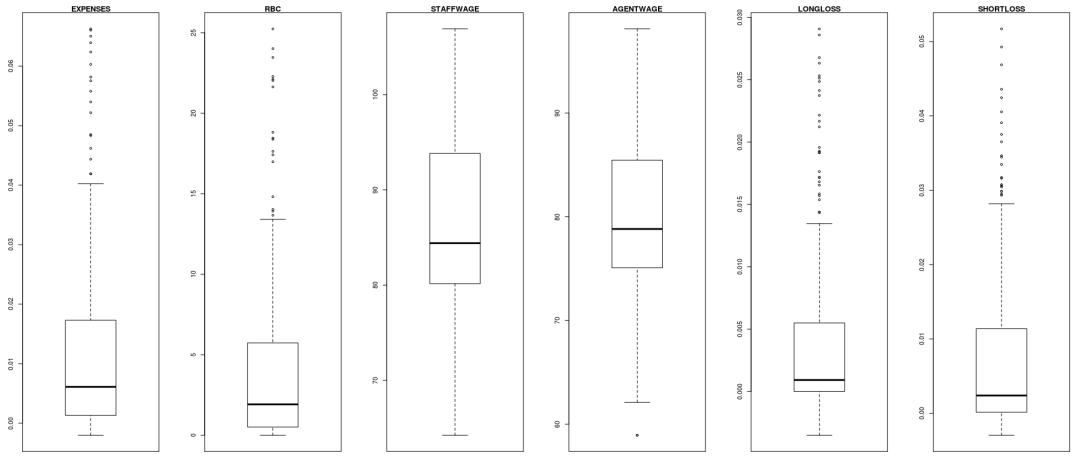

Рис. 2: Скриньки з вусами.

Дійсно, на гістограмах для 'EXPENSES', 'RBC', 'LONGLOSS' та 'SHORTLOSS' помітний тяжкий правий хвіст. Це зумовлено значною кількістю викидів у наборі даних. В цьому можна переконатись, якщо подивитися на графіки скриньок з вусами. Подібні розподіли маємо для групи 'EXPENSES', 'RBC', 'LONGLOSS', 'SHORTLOSS' та 'STAFFWAGE', 'AGENTWAGE'.

Корельованість регресорів. Проаналізуємо залежність між регресорами. На матриці розсіювання можна побачити, що існує лінійна залежність між даними колонок 'STAFFWAGE', 'AGENTWAGE' та 'LONGLOSS', 'SHORTLOSS'. Побудуємо матриці кореляцій за кореляційними функціями Пірсона та Спірмена.

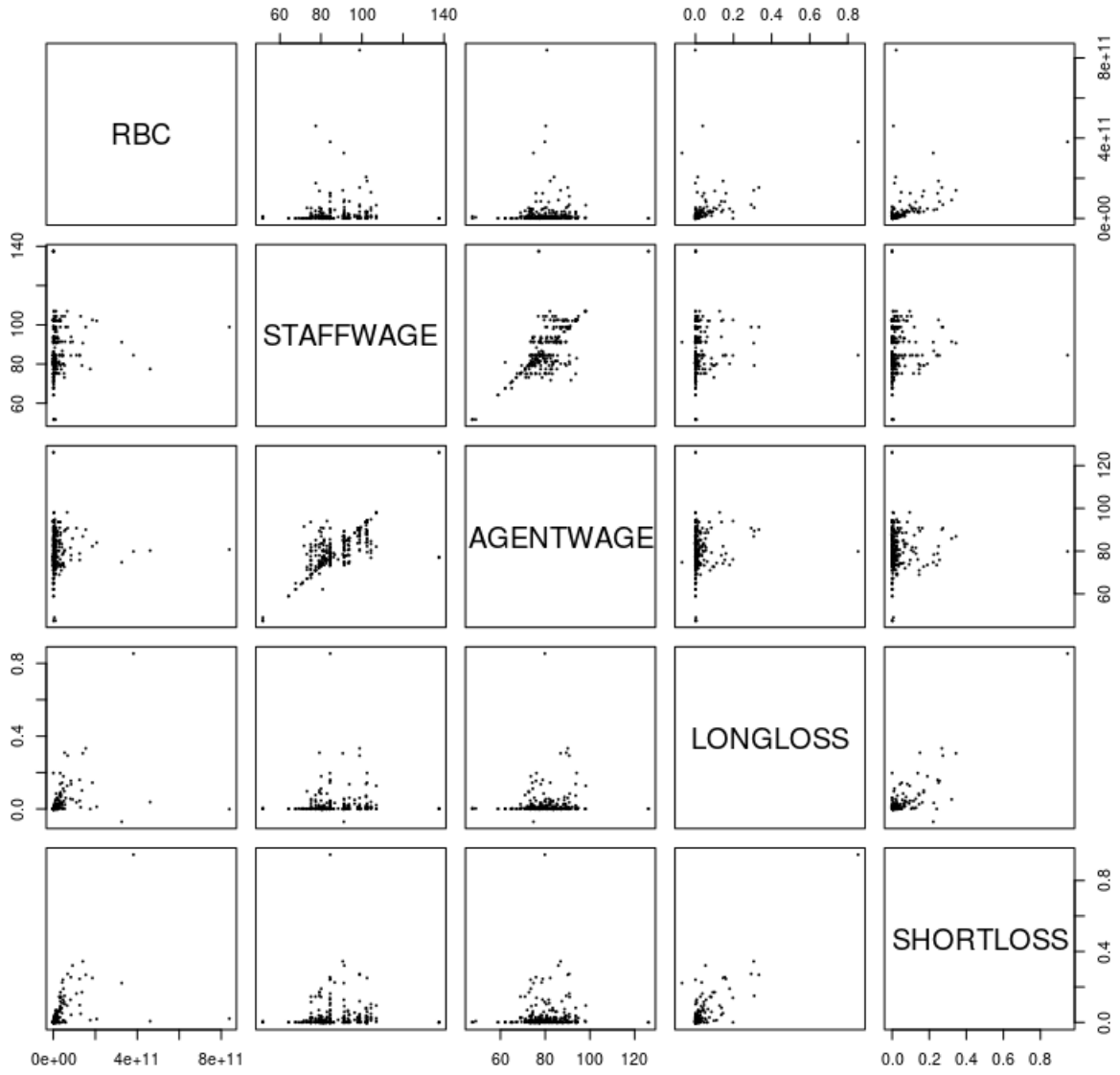

Рис. 3: Матриця розсіювання для заданих регресорів.

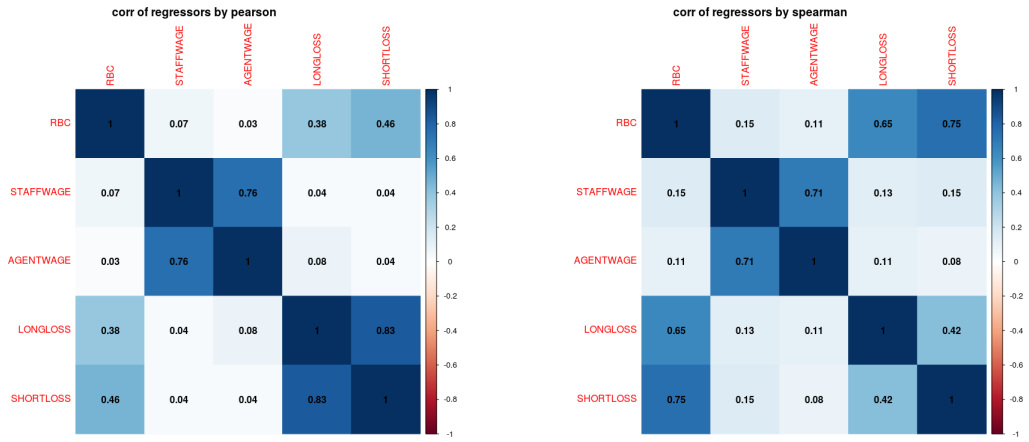

Рис. 4: Матриці кореляцій.

Нелінійна залежність між 'LONGLOSS' ('SHORTLOSS') та 'RBC' сильніша за лінійну. Для інших корельованих пар навпаки. Врахуємо ці моменти під час побудови регресійної моделі. В загальному, можна застосувати модель лінійної регресії для тих даних, що маємо.

Класична модель лінійної регресії. Рівняння моделі має наступний вигляд:

$$\text{EXPENSES} = \beta_0 + \beta_1 \times \text{RBC} + \beta_2 \times \text{STAFFWAGE} + \beta_3 \times \text{AGENTWAGE} + \\ + \beta_4 \times \text{LONGLOSS} + \beta_5 \times \text{SHORTLOSS}$$

Дані про залишки для моделі:

Min	1Q	Median	3Q	Max
-0.256886	-0.007686	-0.006125	-0.000020	0.195474

Отримані оцінки та інші показники (перша колонка - коефіцієнти моделі, друга - середньоквадратичне відхилення, останні дві колонки відносяться до результатів t-критерія Стьюдента):

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	5.177e-03	1.773e-02	0.292	0.771
RBC	1.602e-04	3.337e-05	4.802	2.53e-06
STAFFWAGE	8.662e-05	2.670e-04	0.324	0.746
AGENTWAGE	-6.015e-05	3.436e-04	-0.175	0.861
LONGLOSS	7.303e-01	4.657e-02	15.682	$<$ 2e-16
SHORTLOSS	3.715e-01	3.519e-02	10.557	$<$ 2e-16

Нижче отримані значення для R^2, R^2_{adj}, стандартного відхилення залишків. Також наведено результат тесту Фішера:

Residual standard error: 0.03491 on 286 degrees of freedom
Multiple R-squared: 0.9141, Adjusted R-squared: 0.9126
F-statistic: 608.8 on 5 and 286 DF, p-value: $<$ 2.2e-16

Для такої моделі отримали $MSE = 0.001193346$.

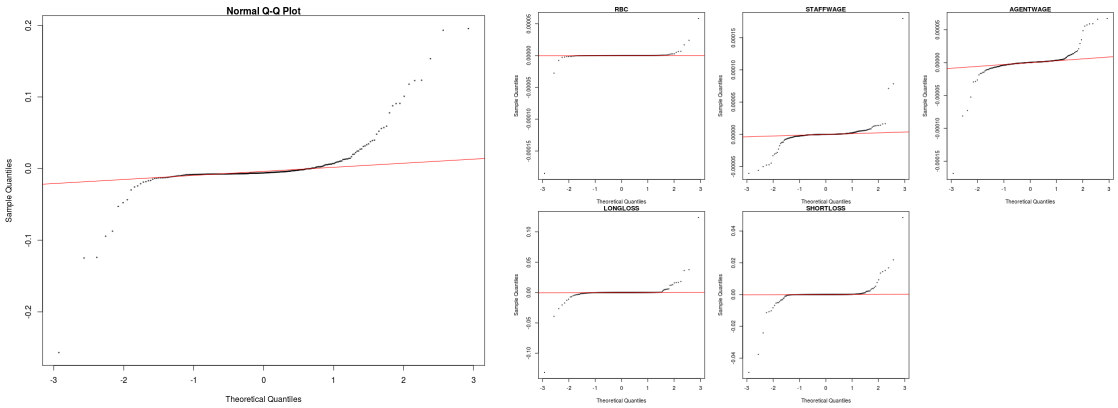

Рис. 5: 1: QQ-діаграма залишків моделі; 2: QQ-діаграми коефіцієнтів першої моделі.

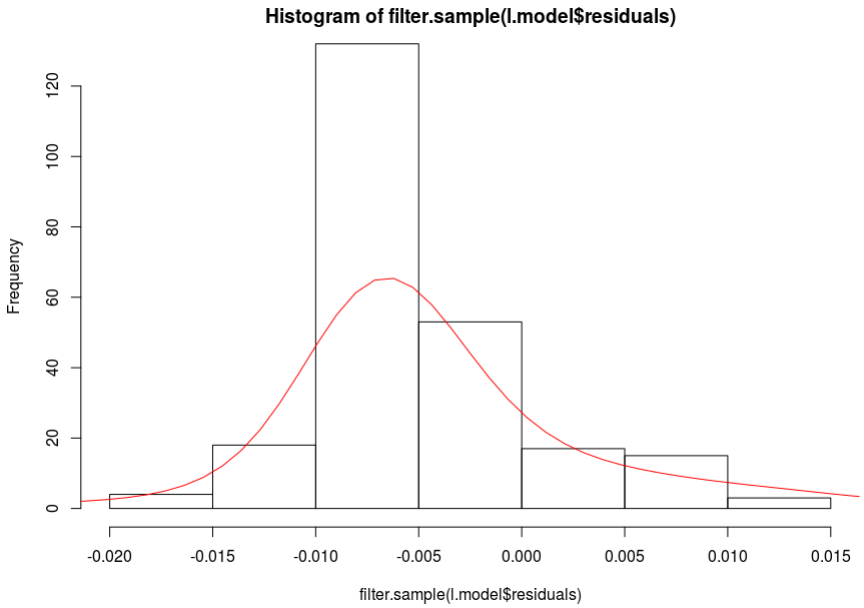

Рис. 6: Гістограма залишків моделі, щільність.

У цій моделі значення вільного члена та коефіцієнти при 'STAFFWAGE', 'AGENTWAGE' майже не відрізняються від 0. Отже в новій моделі використаємо лише 'RBC', 'LONGLOSS', 'SHORTLOSS'. З p-значення тесту Фішера можемо стверджувати, що кореляція між регресорами існує.

Розподіл залишків близький до нормального з незначним зсувом у ліву сторону.

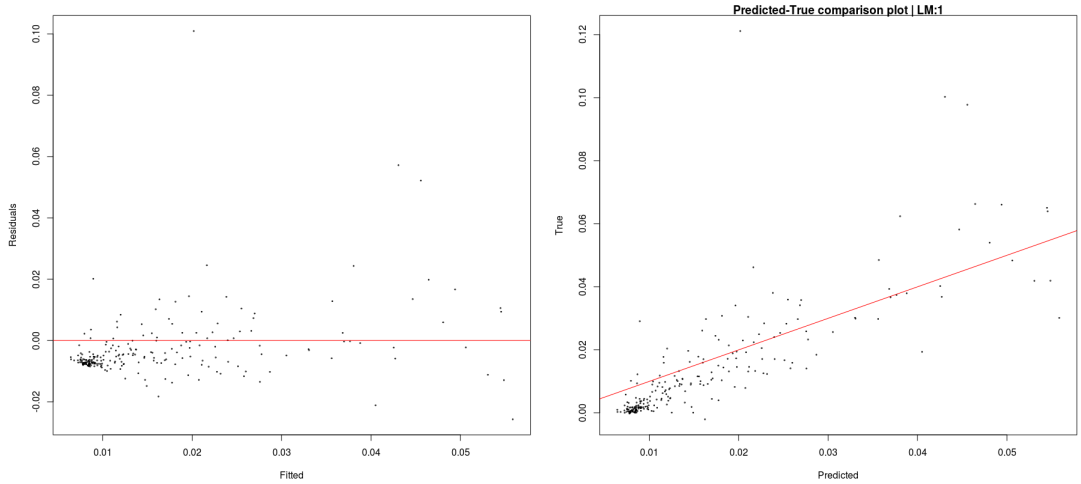

Рис. 7: 1: графік 'прогноз-залишок'; 2: графік 'прогноз-відгук'.

Дивлячись на графік "прогноз-відгук знову переконуємося в тому, що необхідно використати центровану модель регресії.

Перша регресійна модель показала себе непогано, але потрібно врахувати недоліки.

Модифікація моделі регресії. Усунення зайвих регресорів. Враховуючи результати першої моделі, спробуємо усунути коефіцієнт при 'STAFFWAGE', 'AGENTWAGE' та покладемо вільний член рівняння рівному нулеві. Маємо наступне рівняння моделі регресії:

$$\text{EXPENSES} = \beta_1 \times \text{RBC} + \beta_2 \times \text{LONGLOSS} + \beta_3 \times \text{SHORTLOSS}$$

Отримаємо такі результати:

Характеризація залишків:

Min	1Q	Median	3Q	Max
-0.264958	0.000071	0.001625	0.007542	0.202792

Оцінки нової моделі:

	Estimate	Std. Error	t value	Pr(>|t|)
RBC	1.802e-04	3.343e-05	5.391	1.46e-07
LONGLOSS	7.352e-01	4.694e-02	15.661	$<$ 2e-16
SHORTLOSS	3.809e-01	3.555e-02	10.715	$<$ 2e-16

Метрики, тест Фішера:

Residual standard error: 0.03552 on 289 degrees of freedom
Multiple R-squared: 0.9219, Adjusted R-squared: 0.921
F-statistic: 1136 on 3 and 289 DF, p-value: $<$ 2.2e-16

Тут $MSE = 0.001248781$.

Дійсно, результати стали краще. За вказаними раніше тестами, кореляція зберігається, коефіцієнти значущо відмінні від нуля. Збільшилась середньоквадратична похибка моделі.

Перевіримо розподіл залишків:

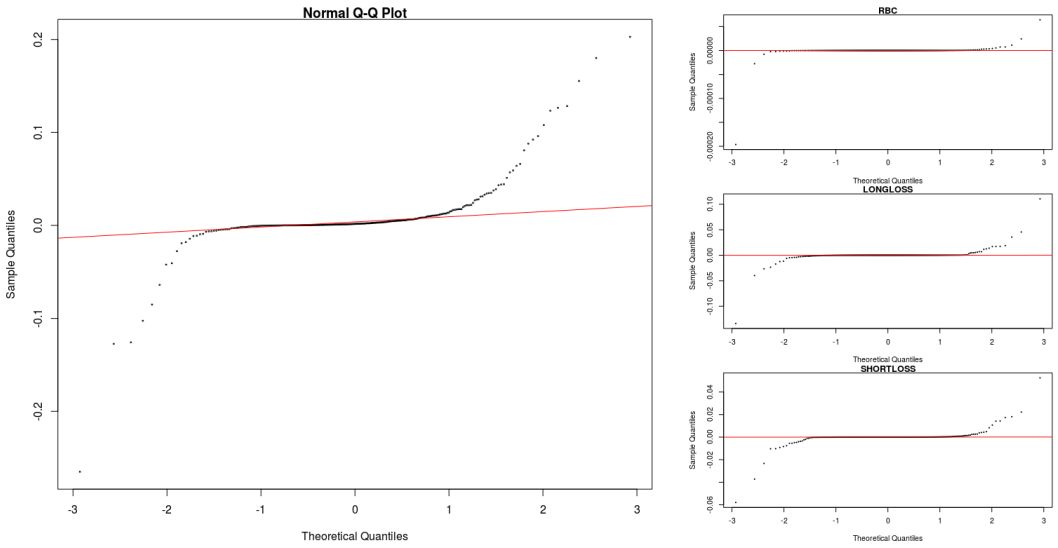

Рис. 8: 1: QQ-діаграма залишків моделі; 2: QQ-діаграми коефіцієнтів нової моделі.

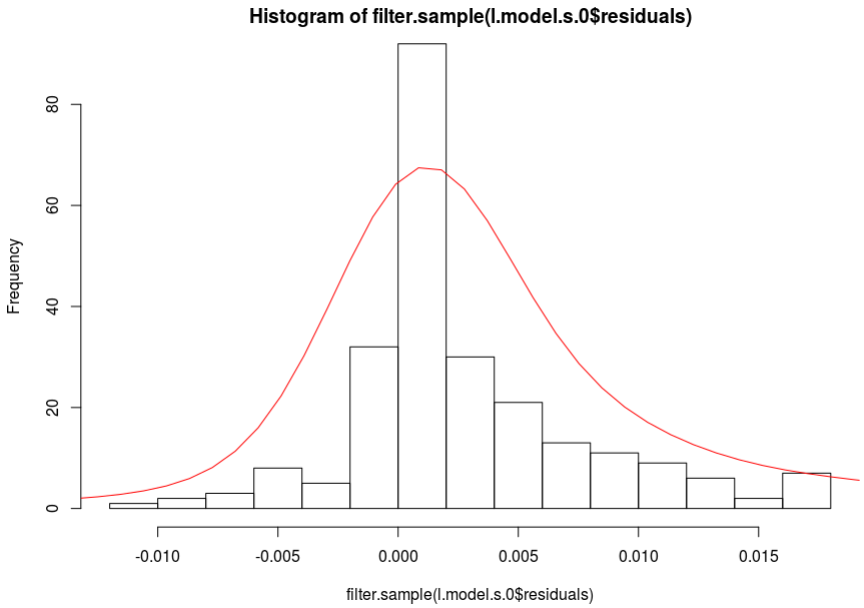

Рис. 9: Гістограма залишків моделі, щільність.

У новій моделі спостерігаємо аналогічну ситуацію: розподіл близький до нормального з середнім біля нуля.

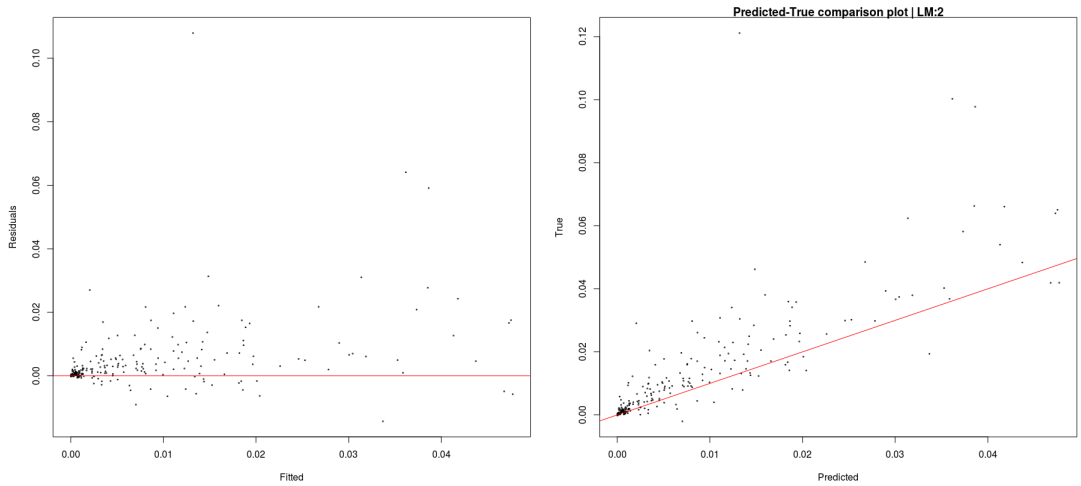

Рис. 10: 1: графік 'прогноз-залишок'; 2: графік 'прогноз-відгук'.

Модифікація моделі регресії. Припущення про нелінійну залежність даних. Спробуємо покращити модель іншим шляхом. Раніше було зазначено, що об'єкт 'RBC' має сильну нелінійну залежність від 'LONGLOSS' ('SHORTLOSS'). Розглянемо ближче діаграму розсіювання однієї з цих пар.

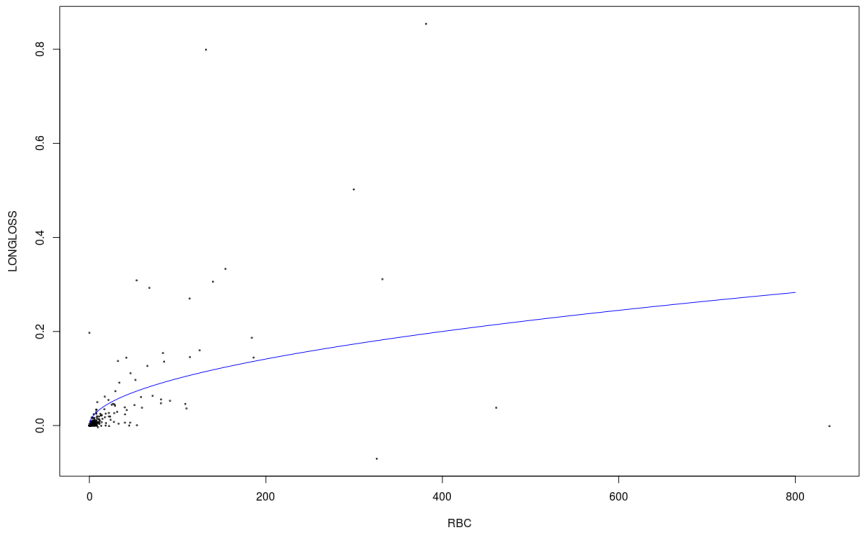

Рис. 11: Діаграма розсіювання даних колонок 'RBC' та 'LONGLOSS'. Також зображено графік функції $y(t) = \sqrt{0.0001 \times t}$, підібраний інтуїтивно.

Неважко припустити, що показану залежність можна описати степеневою функцією $y(t) = (t)^a$. У нашому випадку, беремо $a = \frac{1}{2}$. Враховуючи написане, вводимо нову модель, яка задана рівнянням:

$$\text{EXPENSES} = \beta_1 \times \text{LONGLOSS} + \beta_2 \times \text{SHORTLOSS} + \beta_3 \times \sqrt{|\ \text{RBC} + \text{SHORTLOSS}\ |}$$

Отримаємо такі результати:
Характеризація залишків:

Min	1Q	Median	3Q	Max
-0.235650	-0.005742	-0.002598	-0.000226	0.217082

Оцінки нової моделі:

	Estimate	Std. Error	t value	Pr(>|t|)
LONGLOSS	0.7154317	0.0439276	16.287	< 2e-16
SHORTLOSS	0.3324302	0.0339581	9.789	< 2e-16
$\sqrt{\mid \text{RBC} + \text{SHORTLOSS} \mid}$	0.0048290	0.0005543	8.712	2.35e-16

Метрики, тест Фішера:

Residual standard error: 0.03316 on 289 degrees of freedom
Multiple R-squared: 0.9319, Adjusted R-squared: 0.9312
F-statistic: 1318 on 3 and 289 DF, p-value: < 2.2e-16

Тут $MSE = 0.001088484$.
Похибка зменшилась, інші показники покращились. Кореляція існує, оцінки знайдені.

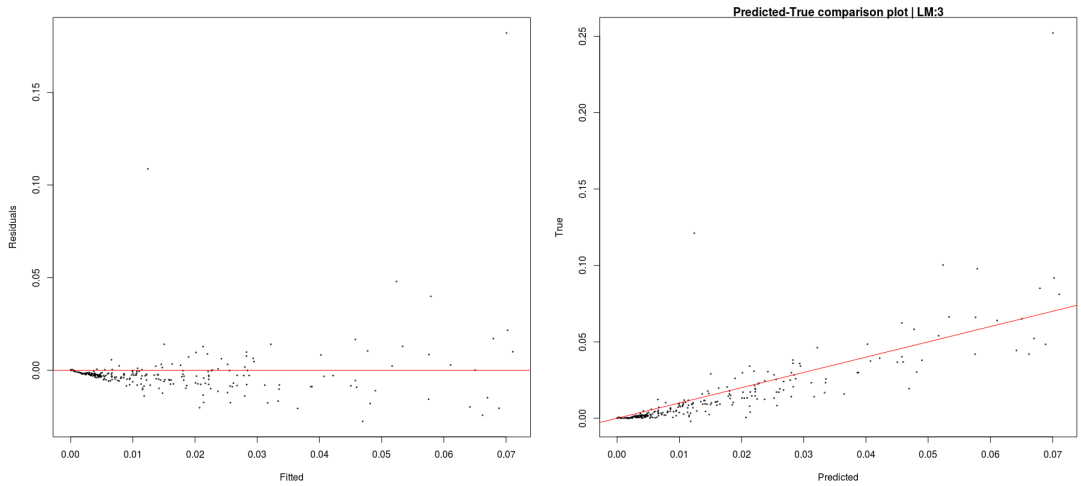

Рис. 12: 1: графік 'прогноз-залишок'; 2: графік 'прогноз-відгук'.

Модель непогана, однак з нелінійністю можна було б ще погратися.
Подивимося на розподіл залишків.

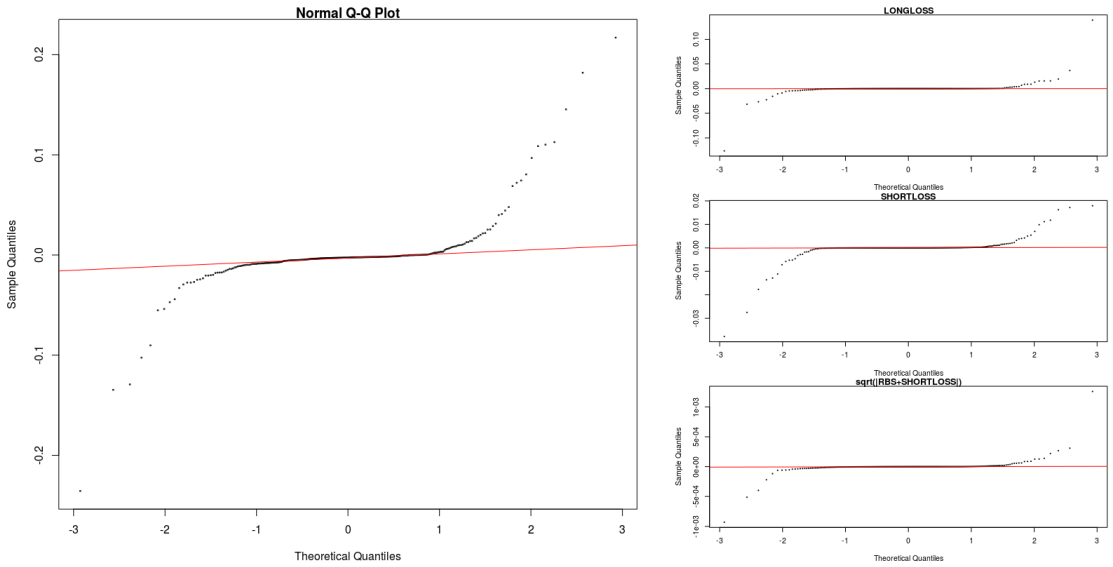

Рис. 13: 1: QQ-діаграма залишків моделі; 2: QQ-діаграми коефіцієнтів нової моделі.

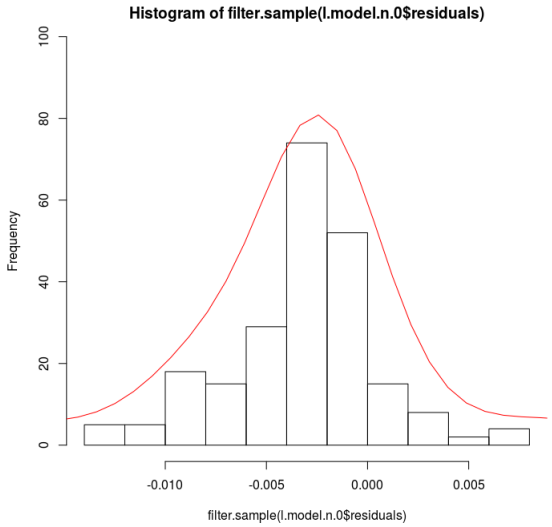

Рис. 14: Гістограма залишків моделі, щільність.

Бачимо незначний зсув від нуля, подібно першій моделі.

Гребенева регресія. *Для рідж-регресії була використана бібліотека 'glm'.*

Першою задачею було знаходження такого $\hat{\lambda}$, з використанням якого модель гребеневої регресії працювала би найкраще. В загальному, будемо порівнювати по результатам значення середньоквадратичної похибки. У даній роботі вектор параметрів λ був описаний наступним чином:

$$\vec{\lambda} = 10^{seq(-4,1,by=.1)} = \{10^{-4}, 10^{-3.9}, \ldots, 10\}$$

Тут розглядається модель гребеневої регресії тільки для загального випадку:

$$\begin{aligned}\text{EXPENSES} = \beta_0 + \beta_1 \times \text{RBC} + \beta_2 \times \text{STAFFWAGE} + \beta_3 \times \text{AGENTWAGE} + \\ + \beta_4 \times \text{LONGLOSS} + \beta_5 \times \text{SHORTLOSS}\end{aligned}$$

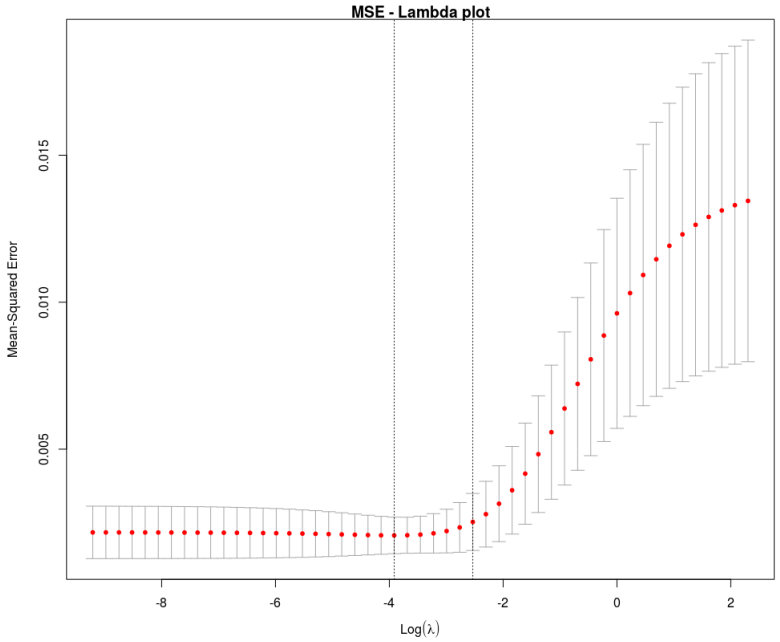

Рис. 15: Графік зміни середньоквадратичної похибки від значення λ. Тут $\ln \hat{\lambda} = -3.914395$.

Отримали оптимальне значення $\hat{\lambda} = 0.01995262$. Використмаємо його в новій моделі та розглянемо отримані оцінки та вимірювання:

Column	Estimate
RBC	1.947086e-04
STAFFWAGE	3.824234e-05
AGENTWAGE	6.241833e-05
LONGLOSS	6.111124e-01
SHORTLOSS	3.730934e-01

Тут $R^2 = 0.9197094$, $R^2_{adj} = 0.9183057$, $MSE = 0.001283118$. Порівняно з першою моделлю регресії, метрики стали кращими, однак збільшилась похибка. p-значення F-тесту (483.8151): $6.345933 \times 10^{-137} < \alpha$, для довільного адекватного рівня значущості α гіпотеза про залежність справджується.

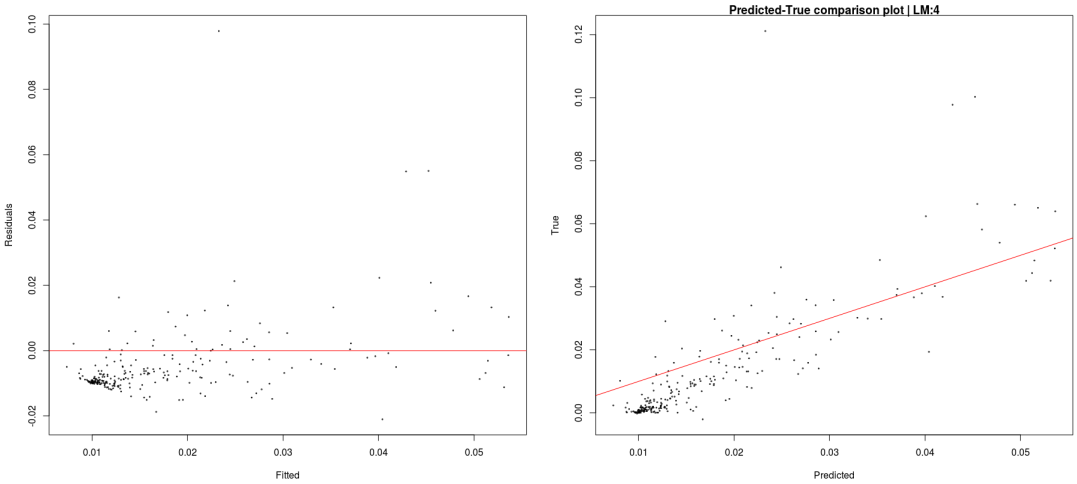

Рис. 16: 1: графік 'прогноз-залишок'; 2: графік 'прогноз-відгук'.

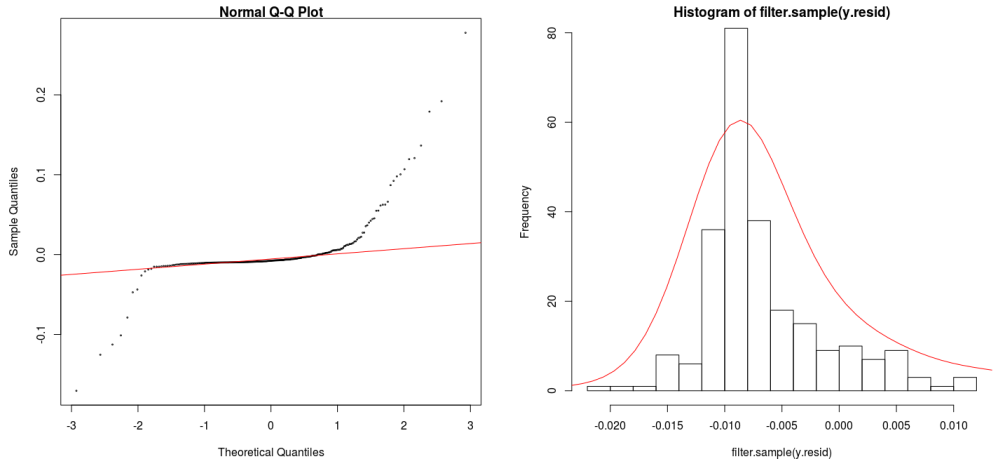

Рис. 17: Гістограма залишків моделі, щільність.

Залишки нормально розподілені - припущення має місце, враховуючи QQ-діаграму та гістограму цих залишків. Помітний зсув на графіках 'прогноз-залишок' та 'прогноз-відгук'.

Аналіз головних компонент, застосування. Використали метод головних компонент.

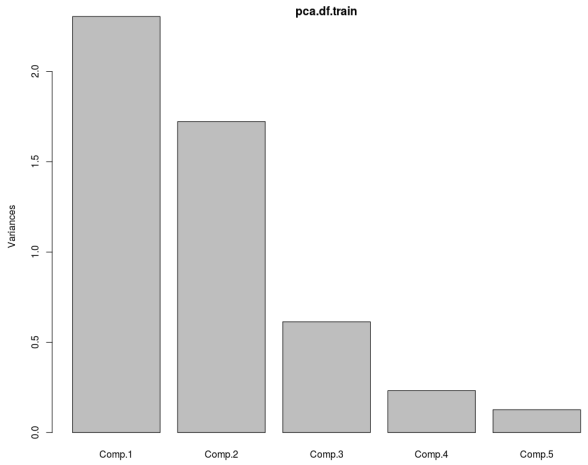

Рис. 18: Дисперсії головних компонент. Злам починається з третьої компоненти.

Тоді рівняння нової моделі матиме вигляд:

$$\text{EXPENSES} = \beta_0 + \beta_1 \times PC_1 + \beta_2 \times PC_2 + \beta_3 \times PC_3$$

Аналіз залишків:

Min	1Q	Median	3Q	Max
-0.248233	-0.007621	-0.005702	0.000431	0.209518

Оцінки грегресії:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.045678	0.002054	22.24	<2e-16
pca.scores[, 1:3]Comp.1	0.070467	0.001353	52.08	<2e-16
pca.scores[, 1:3]Comp.2	0.020689	0.001565	13.22	<2e-16
pca.scores[, 1:3]Comp.3	-0.028522	0.002622	-10.88	<2e-16

Метрики, F-тест:

Residual standard error: 0.0351 on 288 degrees of freedom Multiple R-squared: 0.9126, Adjusted R-squared: 0.9116 F-statistic: 1002 on 3 and 288 DF, p-value: $<$ 2.2e-16

Тут $MSE = 0.001215016$. Зміни хороші, дисперсія збільшена.

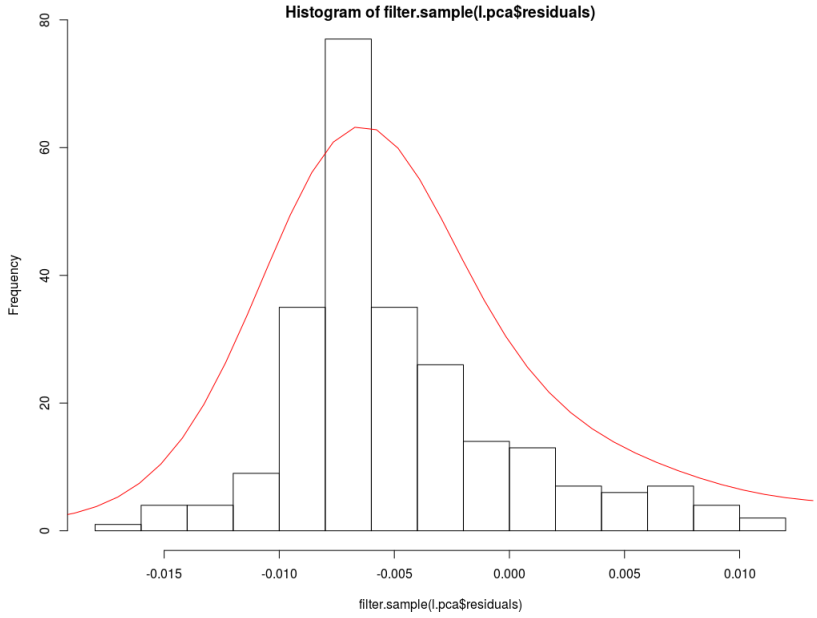

Рис. 19: Гістограма залишків моделі, щільність.

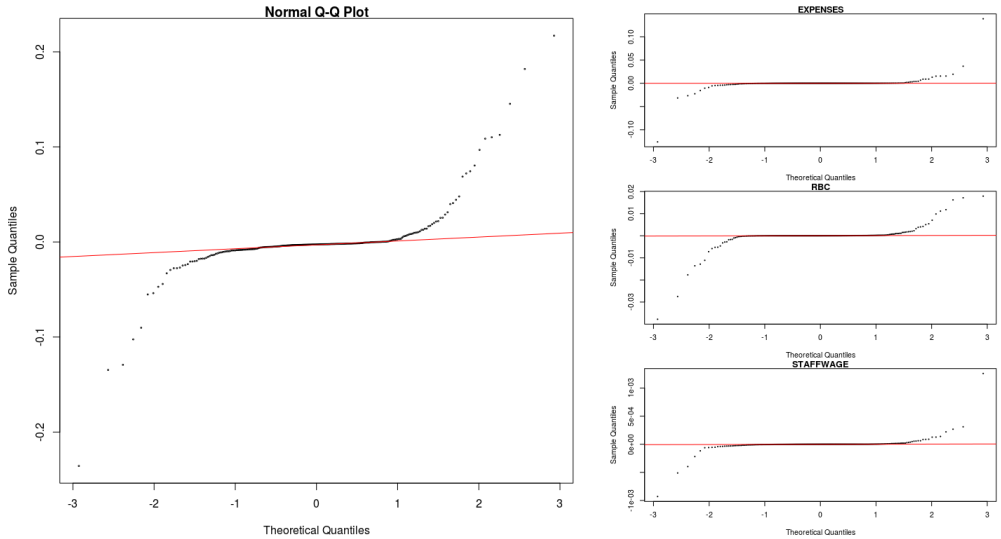

Рис. 20: 1: QQ-діаграма залишків моделі; 2: QQ-діаграми коефіцієнтів нової моделі.

Попередні припущення виконуються.

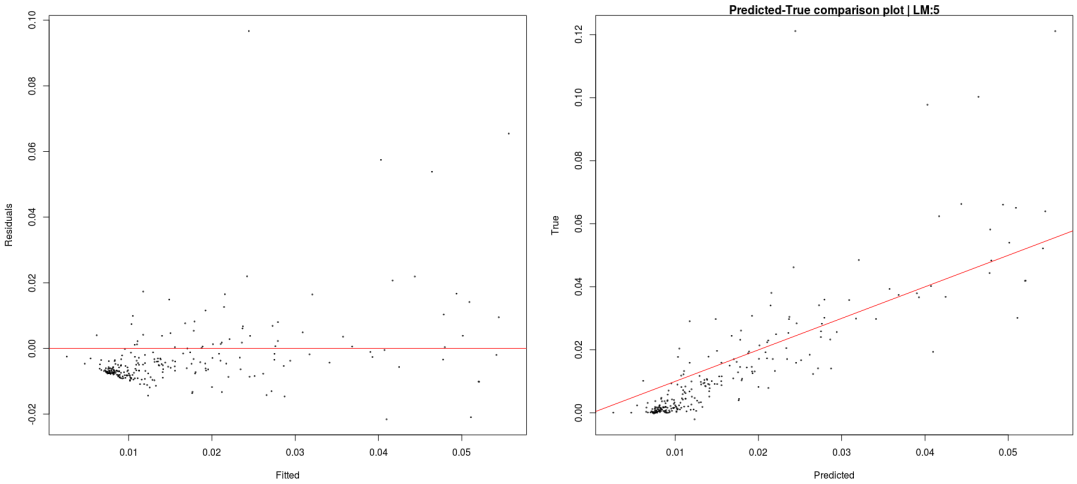

Рис. 21: 1: графік 'прогноз-залишок'; 2: графік 'прогноз-відгук'.

Висновки. У першій лабораторній роботи вдалося провести регресійний аналіз деяких даних. Найкращі результати були отримані з трьох моделей: L3, L4 та L5. Доречно буде використовувати останні дві, оскільки припущення про форму нелінійної залежності не було точним. Не зважаючи на велику кількість викидів та інших проблем, які виникали під час аналізу, були реалізовані досить хороші регресійні моделі. Слід зауважити, що в роботі не використовували алгоритм оптимального відбору регресорів: це пояснюється очевидною залежністю об'єктів та невеликим обсягом даних.