Лабораторна робота №4
з комп'ютерної статистики

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

15 листопада 2021 р.

1 Вступ.

Ця робота є заключною в серії робіт, що присвячені побудові лінійної регресійної моделі залежності цін закриття компанії Colgate-Palmolive від цін закриття інших компаній. У звіті наводяться результати з використання процедури оптимального відбору множини регресорів за допомогою критерію C_p Меллоуза. У кінці наведена стисла порівняльна характеристика нових моделей з найкращими з попередніх робіт.

2 Хід роботи.

2.1 Коротко про критерій C_p Меллоуза.

Нехай $\mathcal{X}^*$ - справжній набір регресорів у регресійній формулі

$$Y_j = \beta_0^* + \sum_{X^i \in \mathcal{X}^*} \beta_i^* X_j^i + \varepsilon_j$$

з математичним сподівання відгука рівним $Y_{j0} := \mathbb{E}[Y_j] = \beta_0^* + \sum_{X^i \in \mathcal{X}^*} \beta_i^* X_j^i$. Для деякого набору регресорів $\mathcal{X}$ запишемо прогноз за оцінками класичного МНК:

$$\hat{Y}_j(\mathcal{X}) = \hat{\beta}_0 + \sum_{X^i \in \mathcal{X}} \hat{\beta}_i X_j^i$$

Вводиться теоретичний функціонал якості обраного набору:

$$\Delta_p(\mathcal{X}) = \frac{1}{\sigma^2}\mathbb{E}\left[\sum_{j=1}^{N}\left(\hat{Y}_j(\mathcal{X}) - Y_j\right)^2\right] = (\#\mathcal{X} + 1) + \frac{SSB(\mathcal{X})}{\sigma^2},\ SSB(\mathcal{X}) = Y_0^T(I_N - P_{\text{л.о.}\{\mathcal{X}\}})Y_0$$

де $\#\mathcal{X}$ – кількість регресорів, що були взяті для прогнозування, $P_{\text{л.о.}\{\mathcal{X}\}}$ – оператор ортогонального проектування на лінійну оболонку, натягнуту на вектори з $\mathcal{X}$. Ми будемо використовувати оцінку для $\Delta_p(\mathcal{X})$:

$$C_p(\mathcal{X}) = \frac{1}{\hat{\sigma}^2}RSS(\mathcal{X}) + 2\#\mathcal{X} - N + 2$$

Хорошими вважатимемо ті набори, при яких $C_p \simeq \#\mathcal{X} + 1 =: p(\mathcal{X})$. Це ми будемо відслідковувати на діаграмі розсіювання $(p(\mathcal{X}), C_p(\mathcal{X}))$ та визначати де є передпідгонка чи недопідгонка, а де є оптимальні набори, які були б варті уваги.

2.2 Відбір регресорів.

2.2.1 Повні дані.

Покажемо $(p(\mathcal{X}), C_p(\mathcal{X}))$-діаграму.

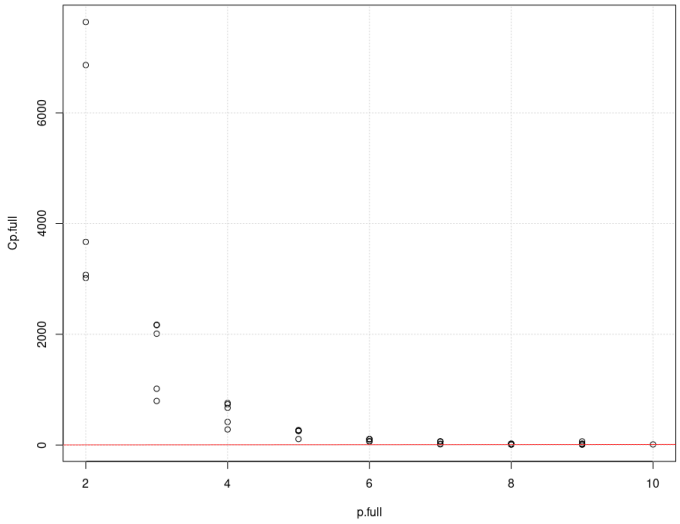

Рис. 1: $(p(\mathcal{X}), C_p(\mathcal{X}))$-діаграма для повних даних.

Як видно з діаграми, моделі, у яких кількість включених регерсорів становить менше п'яти, є недопідігнаними (зміщення, як видно, велике). Тим не менш, це нам заважає досліджувати більш оптимальні комбінації, якщо такі, взагалі кажучи, є. Звузимо огляд діаграми до таких C_p значень, що $0 \leq C_p \leq 30$.

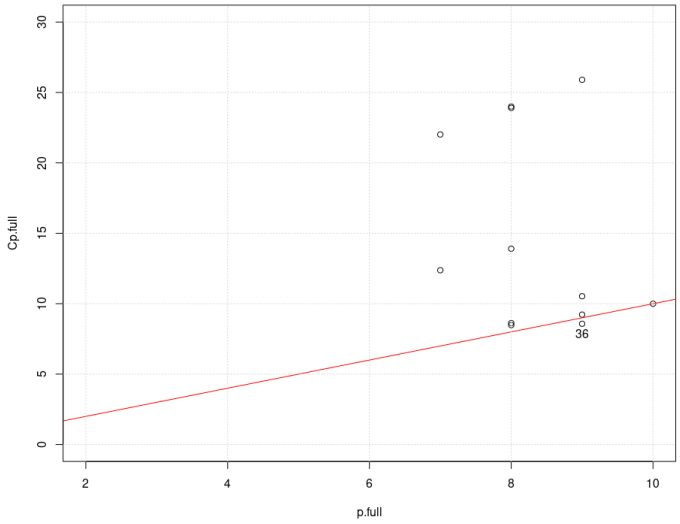

Рис. 2: $(p(\mathcal{X}), C_p(\mathcal{X}))$-діаграма для повних даних, масштабована.

Є одна "хороша" точка, якій відповідає 36-та комбінація регресорів: у ній наявні всі змінні, окрім cmcsa. У порівнянні з іншими точками вона є найкращою лише в сенсі найменшого значення C_p, а так то природньо вона схожа на комбінації, де наявний ефект перепідгонки.

2.2.2 Свіжі дані.

Покажемо $(p(\mathcal{X}), C_p(\mathcal{X}))$-діаграму.

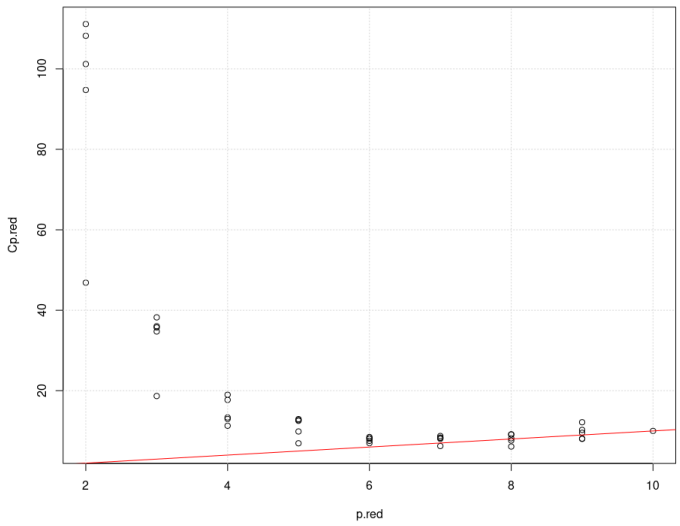

Рис. 3: $(p(\mathcal{X}), C_p(\mathcal{X}))$-діаграма для повних даних.

Як і раніше, нам заважає купа комбінацій, за якими модель буде недопідігнаною. Обмежимося лише тими комбінаціями, для яких $4 \leq C_p \leq 15$. Тоді картина стає більш виразною і можна побачити декілька непоганих кандидатів:

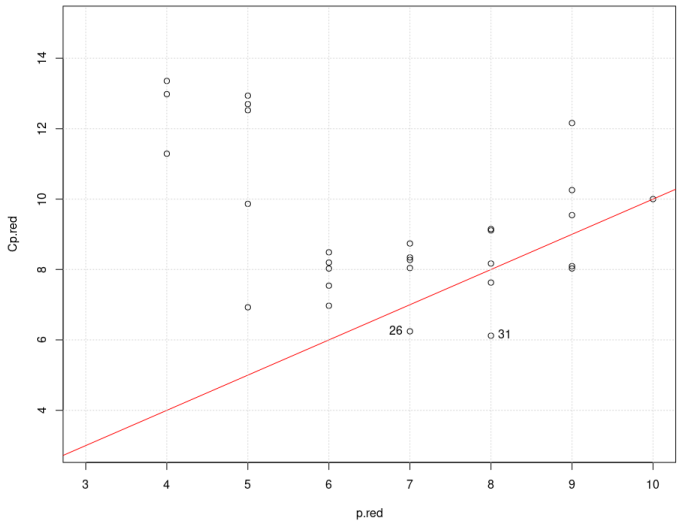

Рис. 4: $(p(\mathcal{X}), C_p(\mathcal{X}))$-діаграма для повних даних, масштабована.

Бачимо хороші точки, для яких $C_p \leq p$. Ми не будемо досліджувати моделі, де кількість використаних регресорів є досить великою в силу того, що спостережень не так багато. Ми розглянемо дві комбінації, для яких C_p є найменшим, причому для однієї з комбінацій не зовсім сильно відрізняється від p. А саме, дослідимо 26-ту та 31-шу комбінації: для них C_p дорівнює 6.244168 та 6.119751 відповідно.

2.3 Підгонка регресійної моделі.

2.3.1 Підгонка за всіма сесіями.

```
Call:
lm(formula = cl ~ . - cmcsa, data = x.dat)

Residuals:
    Min      1Q  Median      3Q     Max
-4.8072 -1.4758 -0.2341  1.3276  5.8942

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.591523   0.727757   4.935 8.72e-07 ***
clf          0.006653   0.004818   1.381    0.167
clx          0.458556   0.020249  22.646  < 2e-16 ***
cma         -0.330526   0.008846 -37.362  < 2e-16 ***
cme          0.055611   0.004481  12.411  < 2e-16 ***
cmg         -0.010693   0.001395  -7.665 2.87e-14 ***
cmi          0.031231   0.007480   4.175 3.12e-05 ***
cms          0.841184   0.068164  12.341  < 2e-16 ***
cnp          0.090693   0.063340   1.432    0.152
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.113 on 1868 degrees of freedom
Multiple R-squared:  0.9492,        Adjusted R-squared:  0.949
F-statistic:  4363 on 8 and 1868 DF,  p-value: < 2.2e-16
```

Оцінені значення коефіцієнтів у певному наближенні схожі до тих, що мали для моделі з усіма регресорами на повних даних. З першого погляду, результати хороші, але якщо спробувати заглибитися більше в аналіз якості прогнозу та залишків, то ми бачимо наслідки перепідгонки.

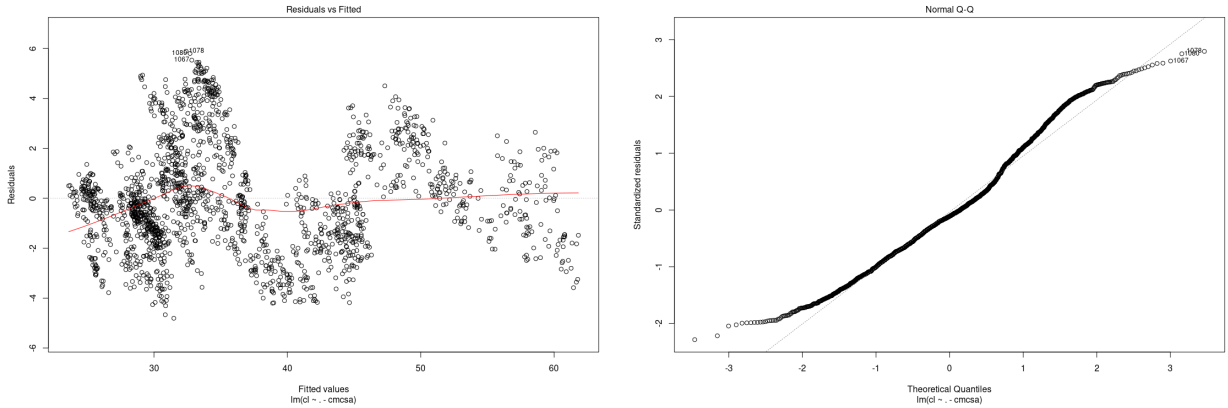

Рис. 5: Зліва-направо: Діаграма залишків та квантильна діаграма для залишків у моделі за повними даними.

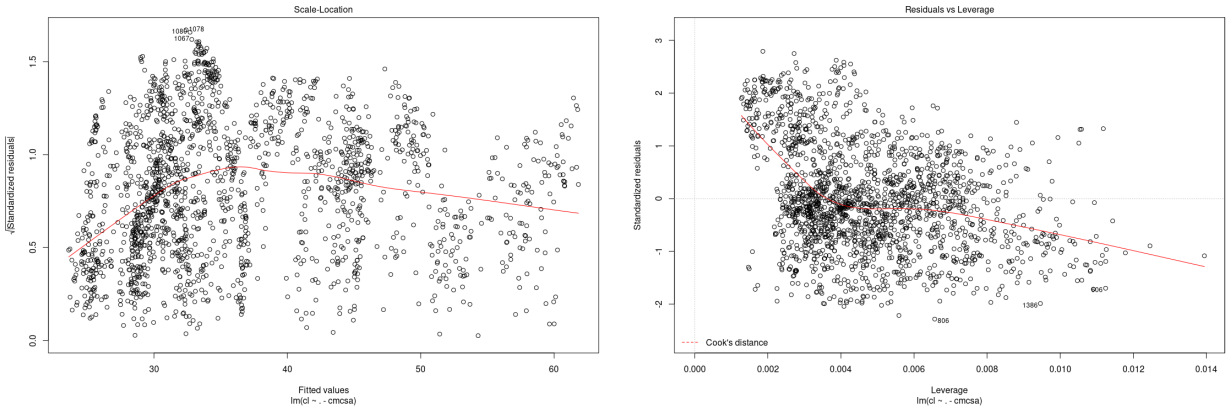

Рис. 6: Зліва-направо: Діаграма кореня стьюдентизованих залишків та діаграма впливу у моделі за повними даними.

2.3.2 Підгонка за "свіжими" даними.

Спочатку наводимо результати підгонки, де вилучається змінні clf, cmg та cmi.

```
Call:
lm(formula = cl ~ . - clf - cmg - cmi, data = x.red)

Residuals:
     Min       1Q   Median       3Q      Max
-1.47132 -0.32572  0.00807  0.40094  1.04133

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.99897    7.06869   0.990  0.32765
clx          0.44810    0.10117   4.429  6.4e-05 ***
cma          0.17613    0.08975   1.962  0.05620 .
cmcsa       -0.24954    0.13623  -1.832  0.07391 .
cme         -0.17330    0.02905  -5.966  4.1e-07 ***
cms         -0.69743    0.31872  -2.188  0.03413 *
cnp          2.06671    0.62998   3.281  0.00206 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5872 on 43 degrees of freedom
Multiple R-squared:  0.8887,        Adjusted R-squared:  0.8731
F-statistic:  57.2 on 6 and 43 DF,  p-value: < 2.2e-16
```

Проблеми зі значущою відмінністю коефіцієнтів від нуля наявні, хоча розкид оцінок коефіцієнтів невисокий (поки не подивитися на оцінку розкиду значень зсуву) та частка поясненої дисперсії прогнозом є достатньо хорошою. Якщо провести діагностику залишків, то результати непогані. Можна було б вилучити впливові спостереження (номери яких можна побачити далі), але цього робити не будемо.

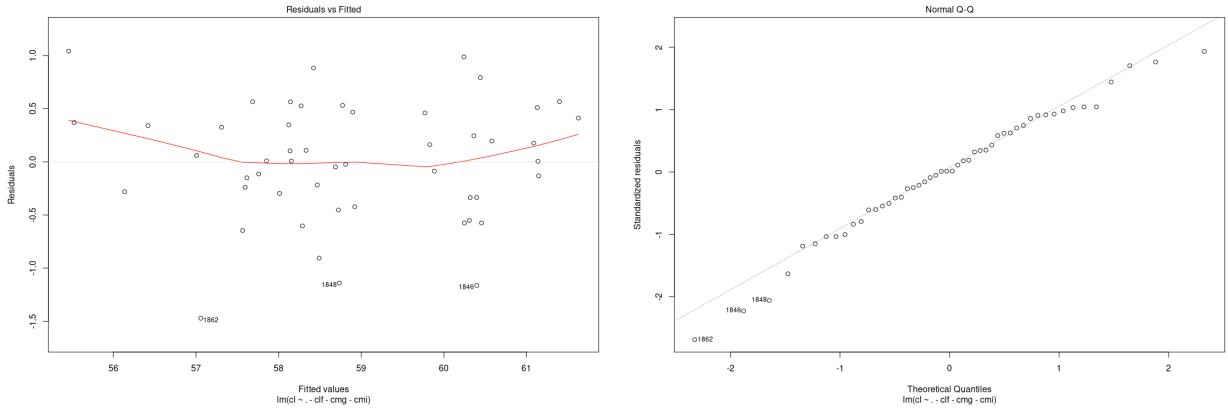

Рис. 7: Зліва-направо: Діаграма залишків та квантильна діаграма для залишків у моделі за свіжими даними. Перша модель.

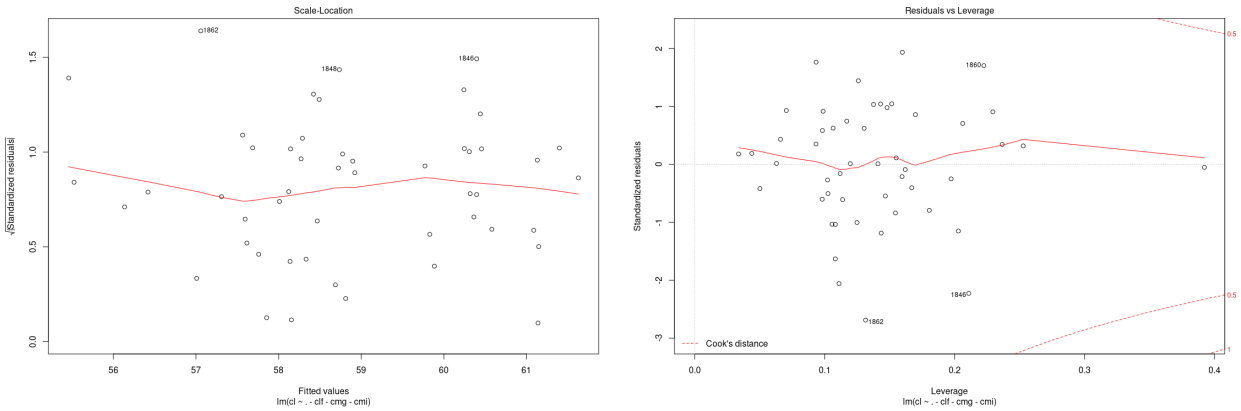

Рис. 8: Зліва-направо: Діаграма кореня стьюдентизованих залишків та діаграма впливу у моделі за свіжими даними. Перша модель.

Далі наведені результати підгонки, де вилучається змінні clf та cmg.

```
Call:
lm(formula = cl ~ . - clf - cmg, data = x.red)

Residuals:
     Min       1Q   Median       3Q      Max
-1.33234 -0.39164  0.05931  0.33190  1.07392

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.33023    7.31938   1.411 0.165506
clx          0.53934    0.11702   4.609 3.75e-05 ***
cma          0.18492    0.08869   2.085 0.043196 *
cmcsa       -0.32869    0.14443  -2.276 0.028024 *
cme         -0.19787    0.03304  -5.988 4.13e-07 ***
cmi         -0.04777    0.03203  -1.491 0.143356
cms         -1.05326    0.39459  -2.669 0.010761 *
cnp          2.43987    0.66970   3.643 0.000734 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5791 on 42 degrees of freedom
Multiple R-squared:  0.8943,        Adjusted R-squared:  0.8766
F-statistic: 50.74 on 7 and 42 DF,  p-value: < 2.2e-16
```

Проблеми із значущістю коефіцієнтів аналогічні до тих, що були у попередній моделі. Видно, що ця модель більше пояснює розкид відгука. Поведінка залишків хороша.

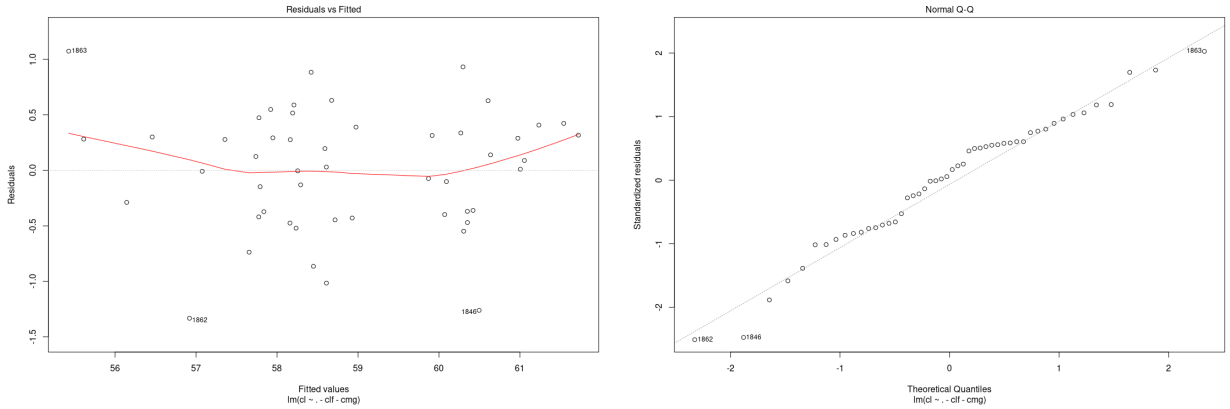

Рис. 9: Зліва-направо: Діаграма залишків та квантильна діаграма для залишків у моделі за свіжими даними. Друга модель.

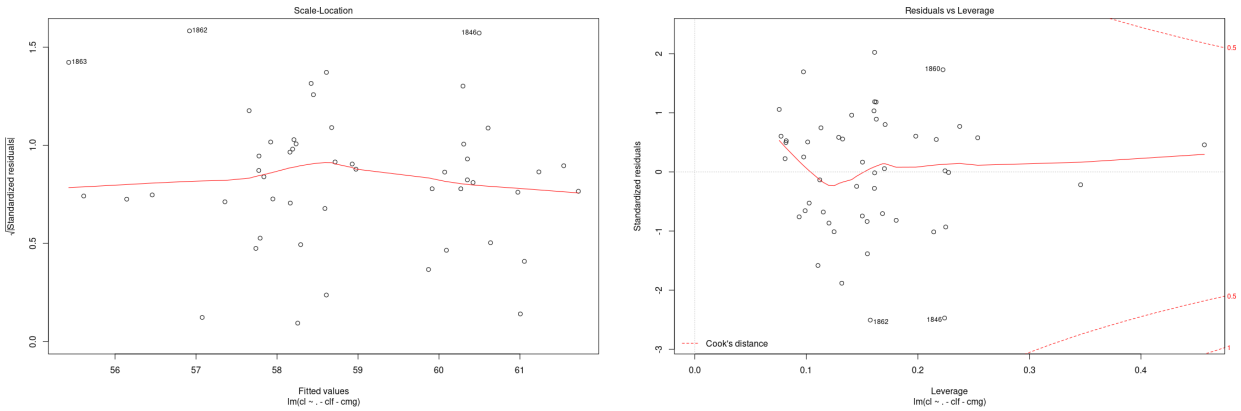

Рис. 10: Зліва-направо: Діаграма кореня стьюдентизованих залишків та діаграма впливу у моделі за свіжими даними. Друга модель.

Залишається зробити прогноз відгука на нових даних та порівняти моделі.

2.4 Прогнозування.

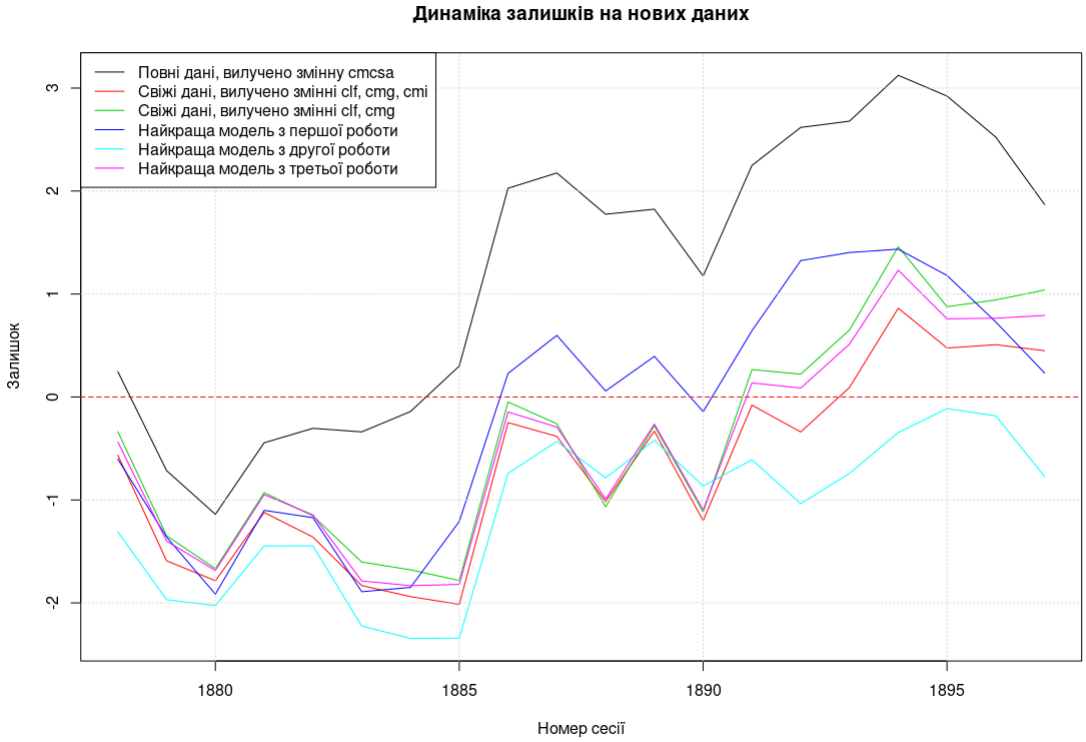

Рис. 11: Порівняння залишків прогнозу на нових даних для різних моделей.

Ефект перепідгонки для першої моделі дає про себе знати на нових даних: залишки зовсім великі у порівнянні з залишками інших моделей. Можна побачити, що моделі на свіжих даних з вилученням потрібних змінних, прогнозують приблизно настільки ж добре, як і найкраща рідж-модель (на свіжих даних, без штрафування зсуву). Отже серед усіх моделей, які наведені у цій роботі, найкращими (в термінах найменших залишків прогнозу) виявилися такі:

- Найкраща модель з першої роботи: свіжі дані, нульовий зсув та залежність від clx та cme; вилучені впливові спостереження;
- Найкраща модель з третьої роботи: свіжі дані, зсув не штрафується;
- Моделі з оптимальними в сенсі Меллоуза наборами регресорів: свіжі дані, вилучені clf, cmg (та cmi).

Питання полягає у тому, яку модель варто було б використовувати. Отримана рідж-модель непогана, однак оцінки для неї є зсунутими, тому невідомо як прогноз може поводитися на інших даних. Модель з першої роботи має зіпсований коефіцієнт детермінації завдяки тому, що зсув не враховувався. А тому точну оцінку того наскільки пояснений розкид відгука за цією моделлю не можна дати. Останні моделі можна було б брати до уваги, оскільки отримані показники для них є більш-менш адекватними. Тим не менш, у цьому незсунутість оцінок переважає над зсунутістю. Розкид, як ми побачили, для цих моделей невеликий (принаймні спостерігаємо на короткий термін часу).

3 Висновки.

Висновки щодо моделей, підходів оцінювання, результатів з попередніх робіт наведені у відповідних звітах. Щодо останньої роботи, за критерієм C_p Меллоуза вдалося побудувати, можливо, не найкращі моделі з точки зору знаучості значень отриманих оцінок, але з точки зору власне якості та адекватності прогнозування є добрими.