Лабораторна робота №2
з комп'ютерної статистики

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

15 листопада 2021 р.

1 Вступ.

Дана робота присвячена побудові регресійної моделі для прогнозування цін закриття на фондовій біржі за допомогою методу головних компонент. Дані, на основі яких робиться підгонка та прогноз, ті самі, як і в першій лабораторній роботі. Запропонована техніка використана для двох випадків: на повних даних та за останніми п'ятдесятьма сесіями. Діагностика моделей, оцінка якості прогнозу на нових спостереженнях, навдяться в кінці роботи.

2 Хід роботи.

2.1 Аналіз головних компонент.

Розподіл деяких змінних відрізнюяться один від одного як за формою, так і за значеннями. Доречно робити аналіз головних компонент з використанням кореляційної матриці, що врахує масштабування та інші лінійні перетворення.

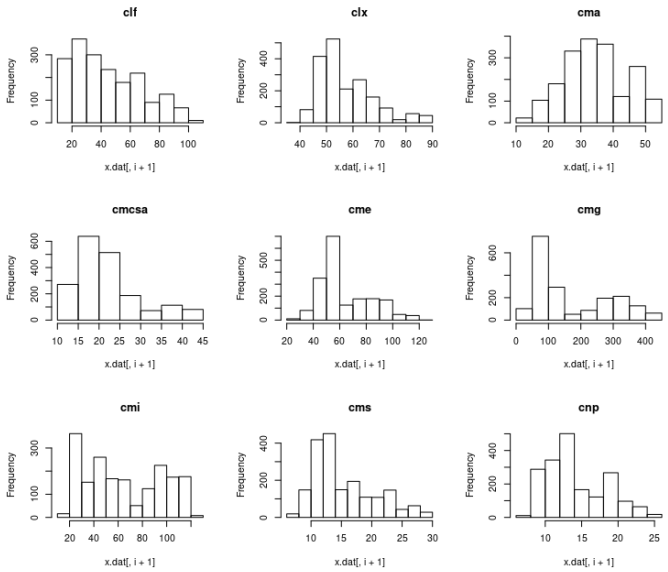

Рис. 1: Гістограми кожної змінної за повними даними (окрім останніх двадцяти сесій).

2.1.1 Повні дані.

Спочатку ми проаналізуємо головні компоненти власні числа та власні вектори за кореляційною матрицю (за Пірсоном в подальшому) за всіма даними окрім останніх двадцяти сесій.

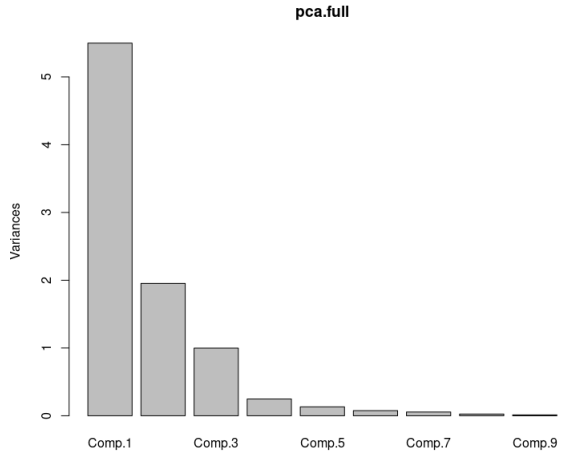

Рис. 2: Діаграма власних чисел за повними даними.

Найбільша частка збереженої дисперсії припадає на першу компоненту (приблизно 61% від загальної дисперсії). Злам відбувається після першої компоненти, хоча не менш інформативними є наступні чотири компоненти, на які припадає близько 33% збереженої дисперсії даних. Наступні компоненти зберегли досить мало інформації, що їх розгляд мало чого дасть. Тому при побудові регресійної моделі на головні компоненти у випадку повних даних, ми зорієнтуємося на перші чотири компоненти за спаданням.

	Comp.1	Comp.2	Comp.3	Comp.4	...
Standard deviation	2.3450767	1.3977712	0.9992432	0.49765095	...
Proportion of Variance	0.6110428	0.2170849	0.1109430	0.02751739	...
Cumulative Proportion	0.6110428	0.8281277	0.9390707	0.96658808	...

Табл. 1: Сингулярні числа (стандартні відхилення) та частки збережених дисперсій за повними даними.

	Comp.1	Comp.2	Comp.3	Comp.4
clf	0.13249271	0.31516210	0.83351196	0.02591704
clx	0.40549756	-0.09498865	-0.16764043	0.15814584
cma	-0.05058491	-0.63451076	0.28860019	0.66786810
cmcsa	0.35690945	-0.31663590	-0.17088810	-0.03748105
cme	-0.12870718	-0.59420053	0.31439950	-0.68034713
cmg	0.40448246	0.09663178	0.04615821	-0.11574630
cmi	0.40223034	0.09451492	0.23695597	0.08688631
cms	0.41660641	-0.11128538	-0.08613154	-0.02534032
cnp	0.41539436	-0.07112192	-0.01083000	-0.20592422

Табл. 2: Навантаження на початкові змінні за повними даними.

2.1.2 ”Свіжі” дані.

Виконаємо аналогічну процедуру аналізу головних компонент для 50 сесій, що передують 20-ти останнім.

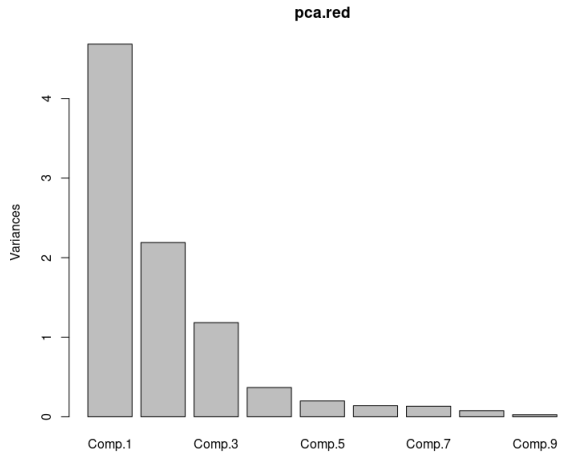

Рис. 3: Діаграма власних чисел за свіжими даними.

Картина майже не змінилася, хоча можна побачити, що трохи більше інформації про розкид розподілося по першим п’яти компонентам (загалом частка збереженої дисперсії за ними становить майже 96%). Для наступних двох компонент частка збереженої дисперсії збільшилася несуттєво. Оскільки спостережень для підгонки небагато, то для адекватності ми будемо робити підгонку за першими чотирьма компонентами.

	Comp.1	Comp.2	Comp.3	Comp.4	...
Standard deviation	2.1644389	1.4800242	1.0880246	0.6064966	...
Proportion of Variance	0.5205329	0.2433857	0.1315330	0.0408709	...
Cumulative Proportion	0.5205329	0.7639186	0.8954517	0.9363226	...

Табл. 3: Сингулярні числа (стандартні відхилення) та частки збережених дисперсій за свіжими даними.

	Comp.1	Comp.2	Comp.3	Comp.4
clf	0.39527793	0.25485258	0.15806822	0.27752621
clx	0.41172184	-0.04221731	0.23806385	-0.37551754
cma	-0.10836636	-0.60196593	0.13033868	0.45022141
cmcsa	0.37099745	-0.29410699	-0.12950435	0.45959962
cme	-0.38241014	-0.29165226	0.05055172	-0.42946339
cmg	0.14795575	-0.58653272	0.12749244	-0.23266708
cmi	0.05520124	0.10193705	0.88980499	0.02322559
cms	0.42793473	0.03801855	-0.25682621	-0.19464817
cnp	0.41298230	-0.20853708	-0.08976678	-0.30171932

Табл. 4: Навантаження на початкові змінні за свіжими даними.

2.2 Підгонка регресійної моделі.

2.2.1 Підгонка за всіма сесіями.

Як говорить назва цього підрозділу, робимо підгонку за повними даними, спочатку перейшовши до нового базису та узявши потрібні змінні. Підгонку робиться за всіма компонентами, поки що. Маємо такий результат:

```
Call:
lm(formula = x.dat$cl ~ ., data = x.dat.pc)

Residuals:
    Min      1Q  Median      3Q     Max
-4.5343 -1.8317 -0.7454  1.5098  7.1864

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.66482    0.05906  620.76   <2e-16 ***
Comp.1       3.75475    0.02519  149.08   <2e-16 ***
Comp.2       0.42738    0.04226   10.11   <2e-16 ***
Comp.3      -1.41567    0.05911  -23.95   <2e-16 ***
Comp.4      -2.02797    0.11869  -17.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.559 on 1872 degrees of freedom
Multiple R-squared:  0.9253,        Adjusted R-squared:  0.9252
F-statistic:  5798 on 4 and 1872 DF,  p-value: < 2.2e-16
```

Досить знайома ситуація, як і першій роботі для моделі за повними даними. Хороший коефіцієнт детермінації, залежність має місце, усі коефіцієнти при змінних значущо відрізняються від нуля...

Далі покажемо результати графічної діагностики моделі.

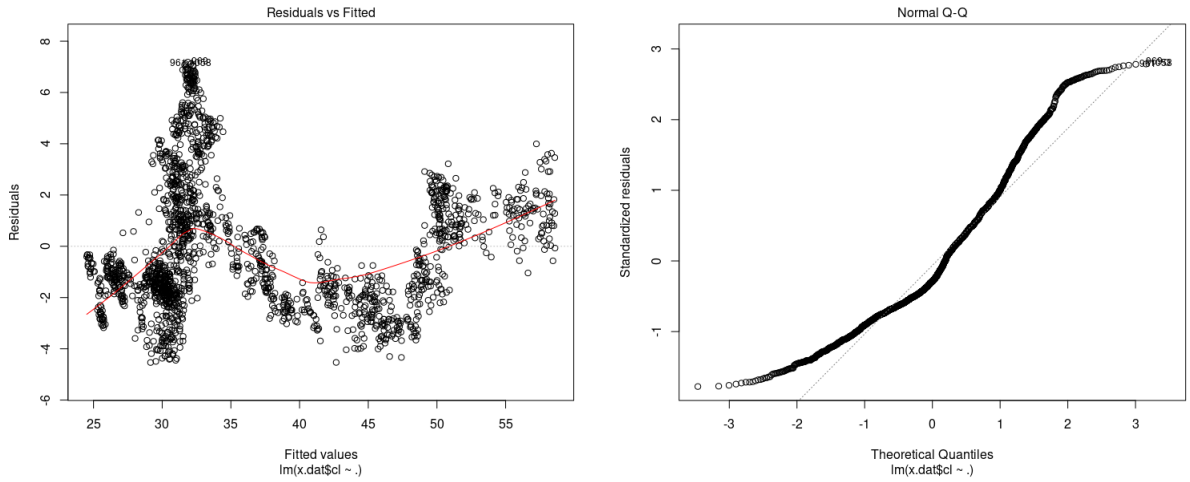

Рис. 4: Діаграма залишків та квантильна діаграма для залишків у повній моделі.

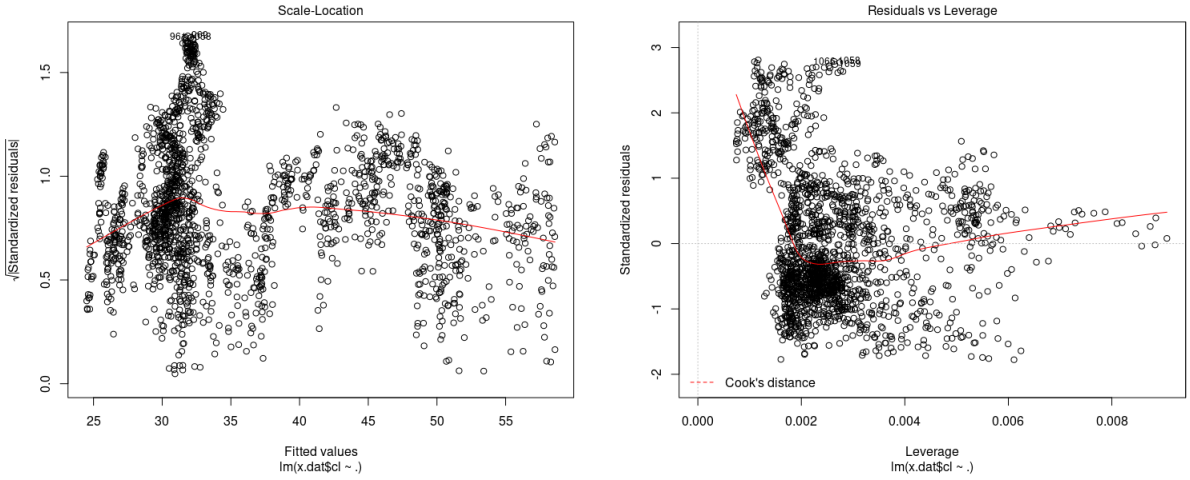

Рис. 5: Діаграма кореня стьюдентизованих залишків та діаграма впливу у повній моделі.

Справді, ситуація майже така сама, як і в попередній роботі: залишки не узгоджуються з нормальним розподілом, характер розкиду залишків є неоднорідним, залежність вимальовується на діаграмі "прогноз-залишки" досить добре. Для прогнозування навіть на короткостроковий час цю модель не варто використовувати.

Все що хотіли та змогли для повної моделі, то зробили, переходимо до підгонки за свіжими даними.

2.2.2 Підгонка за "свіжими" даними.

Техніка така сама, починаємо досліджувати модель залежності від усіх компонент (нагадаємо, що вибрали перші чотири). Ситуація така:

```
Call:
lm(formula = x.red$cl ~ ., data = x.red.pc)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3140 -0.5164  0.1447  0.5083  1.4314

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 58.88294    0.10297 571.867   <2e-16 ***
Comp.1       0.67564    0.04757  14.203   <2e-16 ***
Comp.2      -0.01160    0.06957  -0.167   0.8684
Comp.3       0.20066    0.09464   2.120   0.0395 *
Comp.4      -0.01437    0.16977  -0.085   0.9329
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7281 on 45 degrees of freedom
Multiple R-squared:  0.8209,        Adjusted R-squared:  0.805
F-statistic: 51.56 on 4 and 45 DF,  p-value: 3.051e-16
```

Залежність справді виявлена. Але тест Стьюдента (для кожного з коефіцієнтів) виявляє, що коефіцієнти при другій та четвертій компонентах значущо рівні нулю. Для третьої можна прийняти відмінність від нуля для рівня значущості $\alpha = 0,05$. Хорошою думкою є побудова моделей для різної кількості компонент. Наприклад, ми розглянемо модель, що підігнана на основі першої та третьої головних компонент, та лише за першою компонентою.

Лише з першою та третьою головними компонентами:

```
Call:
lm(formula = x.red$cl ~ Comp.1 + Comp.3, data = x.red.pc)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3148 -0.5128  0.1516  0.4985  1.4205

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 58.88294    0.10079 584.210   <2e-16 ***
Comp.1       0.67564    0.04657  14.509   <2e-16 ***
Comp.3       0.20066    0.09264   2.166   0.0354 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7127 on 47 degrees of freedom
Multiple R-squared:  0.8208,        Adjusted R-squared:  0.8131
F-statistic: 107.6 on 2 and 47 DF,  p-value: < 2.2e-16
```

Лише з першою головною компонентою:

```
Call:
lm(formula = x.red$cl ~ Comp.1, data = x.red.pc)

Residuals:
     Min       1Q   Median       3Q      Max
-2.19402 -0.50441  0.05208  0.62685  1.58838

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 58.88294    0.10460  562.96   <2e-16 ***
Comp.1       0.67564    0.04832   13.98   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7396 on 48 degrees of freedom
Multiple R-squared:  0.8029,        Adjusted R-squared:  0.7988
F-statistic: 195.5 on 1 and 48 DF,  p-value: < 2.2e-16
```

Очевидно, що із вилученням компонент зменшується коефіцієнт детермінації (хоча вражає, що коефіцієнт детермінації в моделі лише за однією компонентою, на яку припадає приблизно 52% інформації початкових даних, досить високий). З іншого боку, вилучення невпливових змінних може сприяти покращенню якості прогнозування, зменшуючи збурення значень коефіцієнтів моделі. Нижче наводимо графічну діагностику лише для моделі з першої головної компоненти.

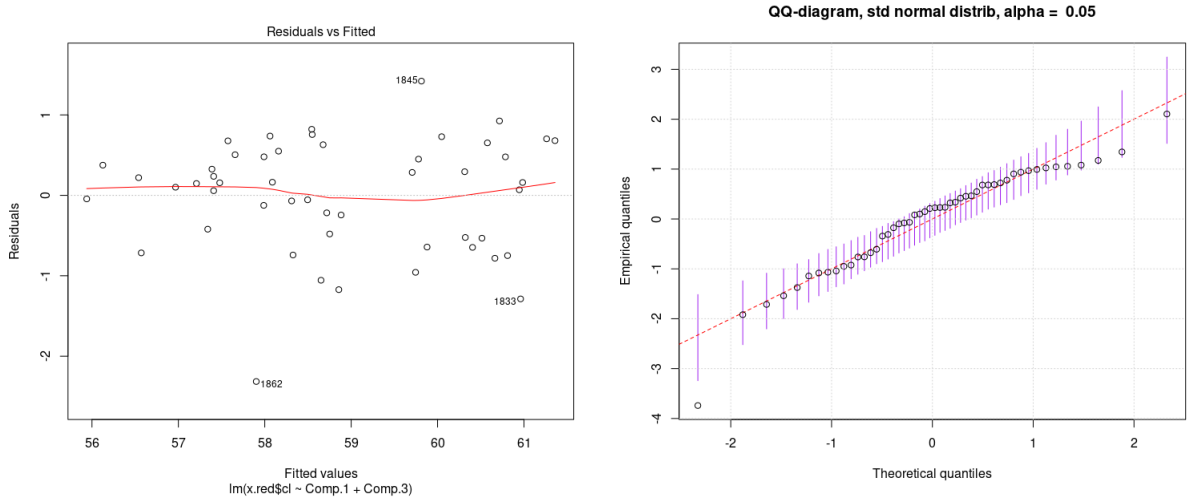

Рис. 6: Діаграма залишків та квантильна діаграма для залишків у "свіжій" моделі.

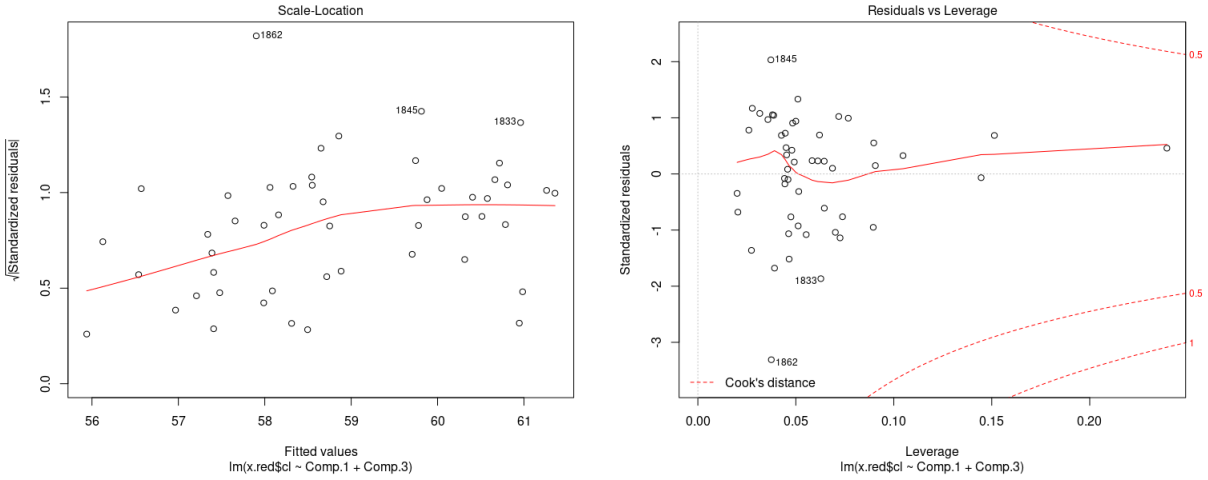

Рис. 7: Діаграма кореня стьюдентизованих залишків та діаграма впливу у "свіжій" моделі.

Результати більш приємні, ніж у попередньому випадку. Можливо, у тій моделі справді була перепідгонка? Але повертаючись до поточної, ми бачимо що залишки більш-менш узгоджуються з нашими пропущеннями, які притаманні гауссовій регресійній моделі. Ми бачимо, що є впливові спостереження (наприклад, під номером 1862 з діаграми стьюдентизованих залишків). Ми їх, згодом, вилучили, але на результат прогнозування на нових даних це не сильно вплинуло. Переходимо, власне, до останнього.

2.3 Прогнозування.

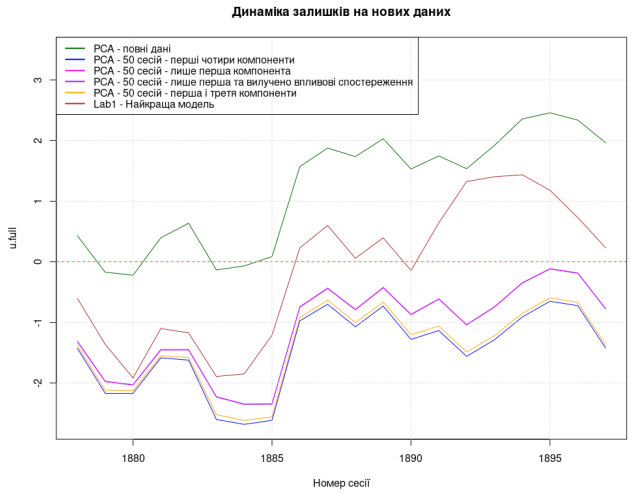

Рис. 8: Порівняння залишків прогнозу на нових даних для різних моделей.

С рисунку можна побачити, що найкраще себе показала так звана сумнівна остання регресійна модель з нульовим зсувом та за 50 сесіями з першої лабораторної роботи: коливання більш-менш відбуваються навколо нуля, відхилення не виходять за великі числа (наприклад 3 з емпіричної точки зору). Розкид помітно менший, ніж у моделях регресії на головні компоненти.

Декілька коментарів щодо якості прогнозу нових моделей. Для моделі регресії за повними даними, порівнюючи з першою моделлю з першої роботи, маємо досить схожу ситуацію: залишки сильно відхиляються від нуля з плином часу. Невдалість прогнозу можна пояснити тим, що залишки у цій моделі не узгоджуються з нормальним розподілом. Далі, серед трьох моделей регресії за 50-ма сесіями добре показала та, яка зроблена на першій головній компоненті. Інші моделі показали себе погано внаслідок включення незначущих компонент у формулу (але це вже виявили після підгонки) та, в загальному, втрати інформації про дані при переході від старих змінних (ціни компаній) до нових (в термінах головних компонент).

Отже серед усіх моделей регресії на головні компоненти більше всього вразила остання як найкраща серед найгірших. Перезапишемо її в термінах початкових змінних. Спочатку виведемо загальну форму:

$$Y_j \approx \beta^0 + \sum_{p=1}^{P} \beta^p U_j^p = \beta^0 + \sum_{p=1}^{P} \beta^p \sum_{l=1}^{d} c_p^l \left(\frac{X_j^l - \overline{X^l}}{S(X^l)} \right) = \beta^0 - \sum_{p=1}^{P} \beta^p \sum_{l=1}^{d} c_p^l \frac{\overline{X^l}}{S(X^l)} + \sum_{p=1}^{P} \beta^p \sum_{l=1}^{d} c_p^l \frac{X_j^l}{S(X^l)} =$$

$$= \beta^0 - \sum_{p=1}^{P} \sum_{l=1}^{d} \beta^p c_p^l \frac{\overline{X^l}}{S(X^l)} + \sum_{l=1}^{d} \sum_{p=1}^{P} \beta^p c_p^l \frac{X_j^l}{S(X^l)} = \alpha^0 + \sum_{l=1}^{d} \alpha^l X_j^l, \text{ де}$$

$$\alpha^0 = \beta^0 - \sum_{p=1}^{P} \sum_{l=1}^{d} \beta^p c_p^l \frac{\overline{X^l}}{S(X^l)}, \; \alpha^l = \sum_{p=1}^{P} \frac{\beta^p c_p^l}{S(X^l)}, \; l = \overline{1, d}$$

У цих позначеннях: $\beta = (\beta^0, \beta^1, \ldots, \beta^d)^\top$ – деякі оцінки коефіцієнтів регресії, $\forall p = \overline{1, P}$: вектор $c_p = (c_p^1, \ldots, c_p^d)^\top$ є власним вектором кореляційної матриці $\hat{\text{cor}}(X) = \{\hat{\text{cor}}(X^i, X^j)\}_{i,j=1}^d$, оцінки $\overline{X^l}$ та $S(X^l)$ є вибірковим середнім та середньоквадратичним відхиленням відповідно характеристик l-ої змінної за тими даними, які були використані для підгонки параметрів моделі (власне коефіцієнтів).

	means	sdevs
clf	18.79199	2.1567315
clx	84.23177	1.5313823
cma	38.90121	1.5289057
cmcsa	41.26472	1.3220544
cme	69.59979	6.4087799
cmg	368.89840	8.0824434
cmi	113.84772	4.0297661
cms	27.38658	0.9385598
cnp	23.58673	0.4798909

Табл. 5: Вибіркові середні та сердньоквадратичні відхилення характеристик l-ої змінної за тими даними, які були використані для підгонки моделі регресії (п'ятдесят сесій перед останніми двадцятьма).

Підставивши усі необхідні значення, виписуємо рівняння регресії через початкові змінні:

$$(\text{cl})_j \approx 11.38658 + 0.12382958 \cdot (\text{clf})_j + 0.18165117 \cdot (\text{clx})_j - 0.04788855 \cdot (\text{cma})_j + 0.18960051 \cdot (\text{cmcsa})_j - \\ - 0.04031549 \cdot (\text{cme})_j + 0.01236822 \cdot (\text{cmg})_j + 0.30805870 \cdot (\text{cms})_j + 0.58144255 \cdot (\text{cnp})_j$$

3 Висновки.

Прогнозування за моделями регресії на головні компоненти виявилося не зовсім вдалим рішенням внаслідок втрати інформації при переході до нових змінних. Позбутися проблеми з неадекватністю розподілу залишків прогнозу у першій моделі на головні компоненти, як у першій моделі за повними даними у попередній лабораторній роботі, не вдалося (це можна було б очікувати, оскільки, грубо кажучи, уся інформація була зміщена в іншу вимірність).