Лабораторна робота №5
з комп'ютерної статистики
Варіант №4

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

4 грудня 2021 р.

1 Вступ.

У даній роботі розглядаються методи боротьби з гетероскедастичністю похибок у лінійній регресійній моделі вигляду

$$Y_j = a + bX_j + \varepsilon_j,\ j = \overline{1, N} \tag{1}$$

Де X_j – реалізації $X \sim N(m, \sigma^2)$, а $\varepsilon_j \sim N(0, g(X_j))$, де $g(t) = (t-1)^2/2$. У цьому варіанті розглядаються $a = -2$, $b = 1/2$, $m = -1$, $\sigma^2 = 2$. Використано адаптивний двокроковий метод найменших квадратів, у якому на роль пілотної оцінки береться оцінка методу найменших квадратів, а для підгонки залежності дисперсій похибок від регресора X використовується поліноміальна регресія другого порядку. Дослідження проводилося для різних обсягів вибірки: $N = 25, 50, 100, 500, 1000, 2000$. Для оцінення зміщення та дисперсії використано імітаційне моделювання.

2 Хід роботи.

2.1 Демонстрація адаптивного МНК.

Спочатку розглянемо на конкретному випадку як акуратно застосувати адаптивний МНК. Зафіксуємо значення зернини для генератора псевдовипадкових чисел рівним 1 та згенеруємо вибірку $X = (X_j)_{j=\overline{1,N}}$ з $N = 1000$ елементів.

```
# Фіксуємо зернину
set.seed(1)
# Функція "кореня з дисперсії"
g <- function(t)
{
  sqrt(0.5)*abs(t-1)
}
# Параметри зсуву і нахилу відповідно
a <- -2
b <- 0.5
```

```
# Мат. сподівання, дисперсія X ~ N(m, s)
m <- -1
s <- sqrt(2)
# Обсяг вибірки
N <- 1000
# Генеруємо дані
# Регресор - набір з реалізацій X
X <- data.matrix(rnorm(N, mean=m, sd=s))
# Похибки
E <- data.matrix(rnorm(N, mean=0, sd=g(X)))
# Моделюємо відгук
Y <- a + b*X + E
```

Цікаво подивитися на діаграму розсіювання відгука відносно регресора.

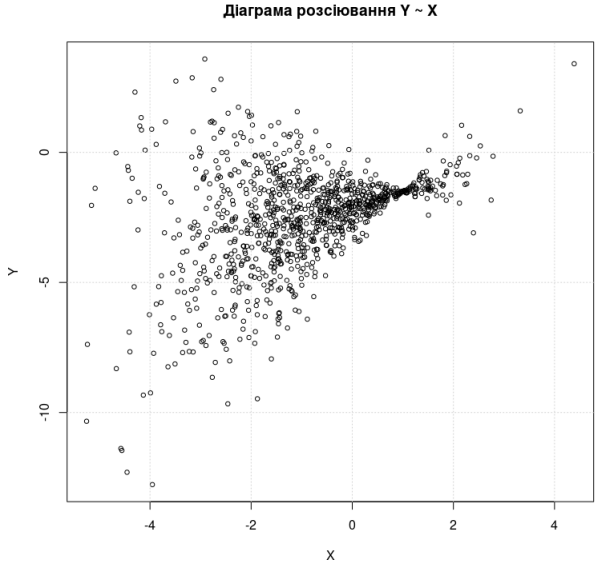

Рис. 1: Діаграма розсіювання відгука Y відносно регресора X.

З діаграми добре видно, що дисперсія похибок ε_j різко зменшується при значеннях X_j в околі одиниці (це зрозуміло з вибору функції, що задає дисперсію похибок). Лінійна тенденція, звісно, прослідковується, але для такого "метелика" класичний МНК не годиться, але для старту його потрібно застосувати.

Добре відомо, що оцінки класичного МНК для одновимірного регресора такі:

$$\hat{b}_{OLS} = \frac{\widehat{\mathrm{cov}}\,(X,Y)}{S^2(X)},\ \hat{a}_{OLS} = \overline{Y} - \hat{b}_{OLS} \cdot \overline{X}$$

Оцінимо коефіцієнти зсуву та нахилу в (1).

```
# Класичний МНК
mX = mean(X)
mY = mean(Y)

b_ols = cov(X, Y)/var(X)
a_ols = mY - b_ols * mX

abline(a_ols, b_ols, col="red")
print(c(a_ols, b_ols))
# -1.9839270  0.5261309

b <- data.matrix(c(a_ols, b_ols))
```

Значення оцінок за МНК такі: $\hat{b}_{OLS} = 0.5261309$, $\hat{a}_{OLS} = -1.9839270$. Значення оцінок досить близькі до справжніх.

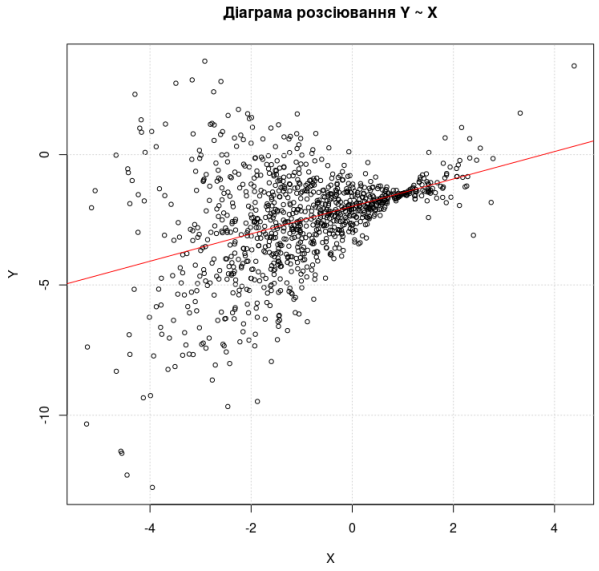

Рис. 2: Діаграма розсіювання відгука Y відносно регресора X. Червона лінія – пряма регресії.

На око можна сказати, що пряма за отриманими значеннями оцінок в певному наближенні ділить навпіл метелика. Розглянемо поведінку залишків прогнозу U_{OLS} на діаграмі "Прогноз-залишки".

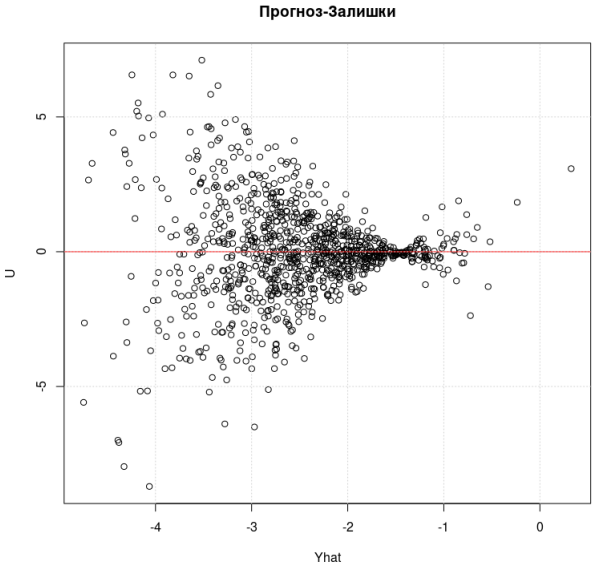

Рис. 3: Діаграма "Прогноз-залишки".

Чого, власне, можна було і очікувати: розкид залишків є неоднорінним. Потрібно враховувати неоднорідність – це зробимо за допомогою адаптивного двокрокового МНК. Дослідимо поведінку U^2_{OLS} від X:

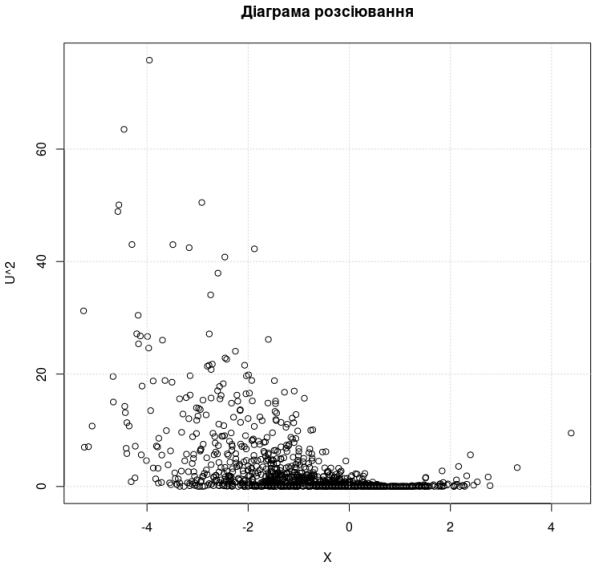

Рис. 4: Діаграма розсіювання U^2_{OLS} відносно X.

Зрозуміло, що відстежується поліноміальна залежність значень квадрату залишків від значень регресора. Припустимо, що залежність є поліноміальною другого порядку:

$$U^2_{OLS,j} = \alpha + \beta X_j + \gamma X_j^2 + \eta_j,\ j = \overline{1, N}$$

де η_j - випадкова похибка. Тоді можна спробувати застосувати МНК для оцінювання α, β і γ, що ми і зробимо.

```
# Підгонка параметрів у моделі U^2 = alpha + beta*X + gamma*X^2
X.poly <- cbind(1,X,X^2)
A.poly <- t(X.poly)%*%X.poly
greeks <- solve(A.poly)%*%t(X.poly)%*%(U_ols^2)
print(greeks)

f <- function(t) {c(1, t, t^2)%*%greeks}
fv <- function(v) {sapply(v, f)}
curve(fv(x), col="red", add=T)
```

Оцінки коефіцієнтів такі: $\hat{\alpha} = 0.4024425$, $\hat{\beta} = -0.9798103$, $\hat{\gamma} = 0.6247096$. Досить віддалено нагадують справжні значення, які використовувалися при моделюванні.

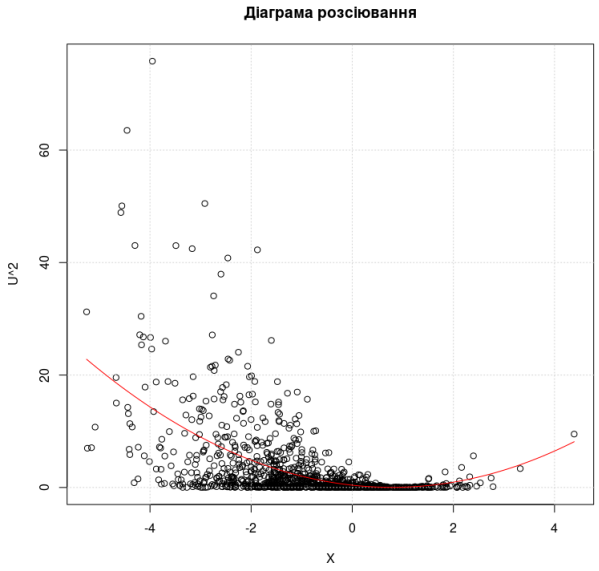

Рис. 5: Діаграма розсіювання U^2_{OLS} відносно X. Домальовано графік $P(t) = \hat{\alpha} + \hat{\beta}t + \hat{\gamma}t^2$.

Значюи прогноз для квадратів залишків, можемо побудувати вагову матрицю W та застосувати навантажений МНК для можливого покращення ситуації. Звернімо увагу на те, що знак компонент W відіграє важливу роль при мінімізації навантаженого МНК-функціонала та і аналізу моделі в цілому. Можливі випадки (залежить від початкових даних), коли $P(t)$ набуває від'ємних значень. У нашому випадку такого щастя немає, усі спрогнозовані значення є додатними:

```
> Usq_fitted <- X.poly%*%greeks
> print(sum(Usq_fitted < 0))
[1] 0
```

Але у випадку від'ємних значень прогнозу потрібно щось робити. Першим, що спадає на думку, це зануляти відповідні компоненти вагової матриці $W = \operatorname{diag}\left(\left(\hat{U}_j^2\right)^{-1},\ j = \overline{1,N}\right)$. Таким чином ми не тільки зможемо зберегти початкові властивості функціонала (наприклад опуклість), а й зберегти змістовність отриманих результатів. Чому це справді важливо, розглянемо на конкретному випадку.

Наприклад, візьмемо таку зернину, за якою можна було б отримати ті дані, підгонка за якими давала б від'ємні вагові коефіцієнти W. Таку ситуацію виявили для початкового значення рівному 10000. У такому разі матриця не є невід'ємно визначеною (хоча є симетричною), тому опуклість тут порушується. Тоді постає проблема, що єдина критична точка необов'язково є точкою мінімуму функціонала.

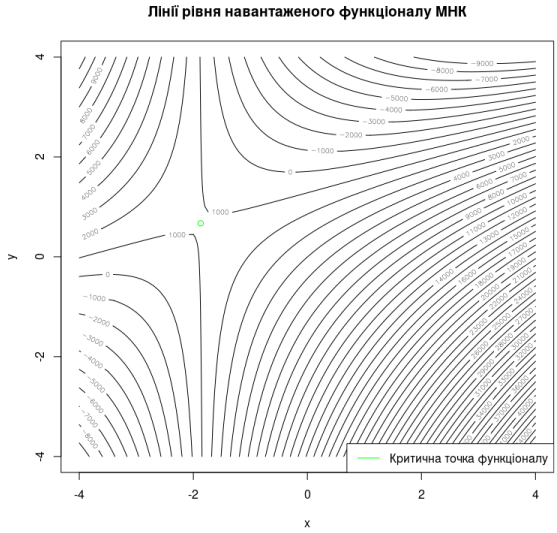

Рис. 6: Лінії рівня навантаженого функціоналу, зернина: 10000. Функціонал не є опуклим, критична точка не доставляє мінімум, а є сідловою.

Якщо зануляти від'ємні компоненти, то маємо палицю з двома кінцями. Зануляючи відповідні діагональні елементи, ми не враховуємо квадрати відхилень для тих спостережень, номер яких відповідав від'ємним значенням матриці W. Такий підхід є ризиковим, бо втрачається точність оцінювання, тому так варто робити при значних обсягах вибірки та з певним знанням про те, що ймовірність того, що прогноз значення U^2 набуває від'ємного значення є відносно малою.

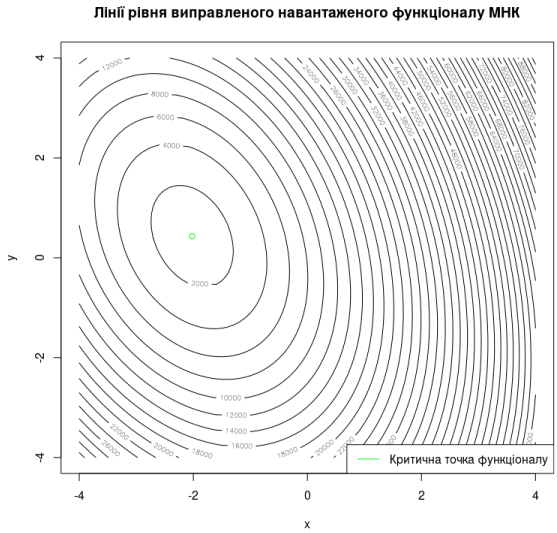

Рис. 7: Лінії рівня виправленого навантаженого функціоналу, зернина: 10000. Функціонал є опуклим, критична точка доставляє мінімум.

Якщо знову звернутися до випадку з зерниною рівною одиниці, то, як зазначили раніше, функціонал не псується, бо спрогнозовані квадрати залишків є додатними. Критична точка доставляє мінімум.

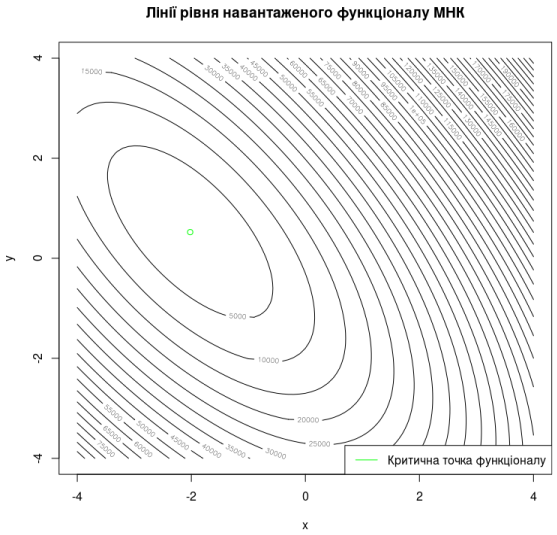

Рис. 8: Лінії рівня навантаженого функціоналу, зернина: 1. Функціонал є опуклим, критична точка доставляє глобальний мінімум.

```
Aw <- t(cbind(1,X))%*%W.res%*%cbind(1,X)
bw <- solve(Aw)%*%t(cbind(1,X))%*%W.res%*%Y
```

Зокрема значення оцінки навантаженого МНК дорівнює:

$$\hat{a}_{GLS} = -2.0202088,\ \hat{b}_{GLS} = 0.5229783$$

Дослідимо залишки прогнозу на трансформованих даних. Зробимо центрування та нормування значень відгука та регресора на корені діагональних елементів W, тобто:

$$\frac{(Y_j - \overline{Y})}{\sqrt{\hat{U}_j^2}} = b \cdot \frac{(X_j - \overline{X})}{\sqrt{\hat{U}_j^2}} + \tilde{\varepsilon}_j,\ j = \overline{1, N}$$

Тоді можна побачити, що ефект гетероскедастичності начебто затухає. Це видно як з діаграми розсіювання відгука відносно регресора, так і на діаграмі "Прогноз-залишки". Діаграми наведені далі.

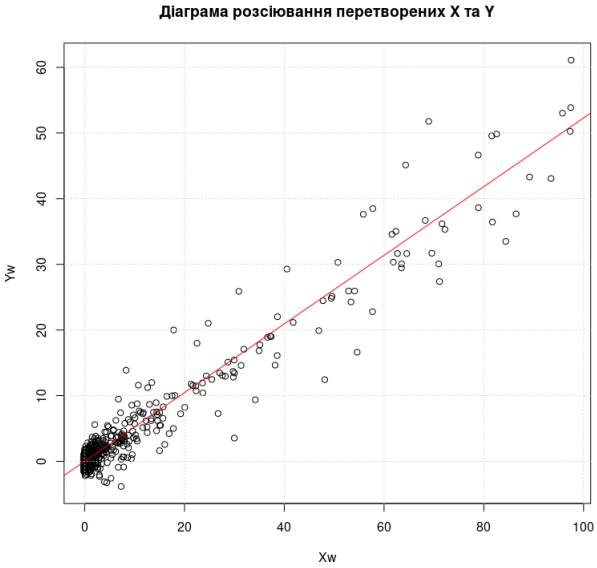

Рис. 9: Діаграма розсіювання перетворених даних. Червона лінія – пряма регресії за оцінками навантаженого МНК.

Метелика вже немає, його перетворили на комету. Звернімо увагу на те, що точки сильно розкидані – це може пояснюватися тим, що на стику метелика дисперсія близька до нуля, а тому відповідний ваговий коефіцієнт стає великим.

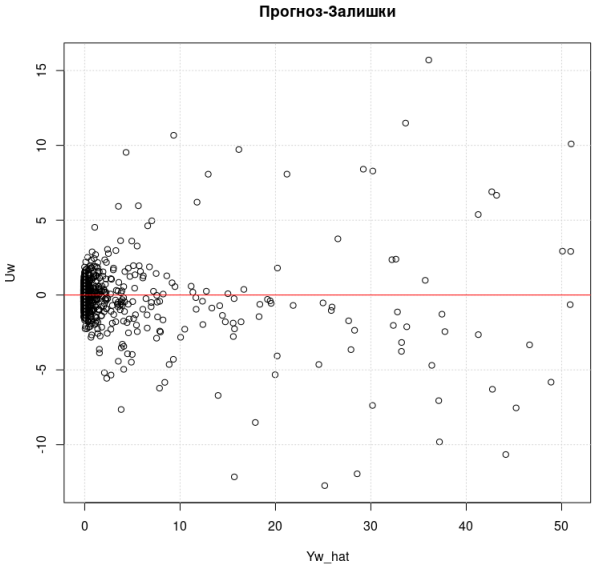

Рис. 10: Діаграма "Прогноз-залишки" на перетворених даних.

Видно, що ефект неоднорідності розкиду похибок послаблено. Однак бачимо, що перетворення є в певному сенсі грубим: можна звернути увагу на те, що в околі початку координат скупчується ціла хмара точок. Розглянемо такий варіант виправлених ваговх коефіцієнтів:

$$\forall j = \overline{1,N}: \; W^0_{jj} := \begin{cases} 0 & W_{jj} \leq 0, \\ W_{jj} & 0 < W_{jj} \leq 1, \\ 1 & W_{jj} > 1. \end{cases} \tag{2}$$

Який результат будемо з цього мати?

Зокрема значення оцінки навантаженого МНК з ваговою матрицею W^0 дорівнює:

$$\hat{a}^0_{GLS} = -2.0109172,\ \hat{b}^0_{GLS} = 0.5181089$$

На перетворених даних (тільки нормування робиться з використанням нових виправлених вагів) знову розглянемо залишки.

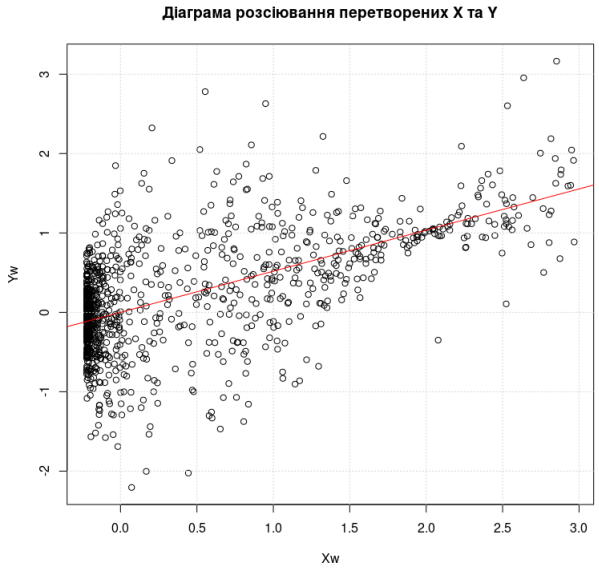

Рис. 11: Діаграма розсіювання перетворених даних. Червона лінія – пряма регресії за оцінками навантаженого МНК.

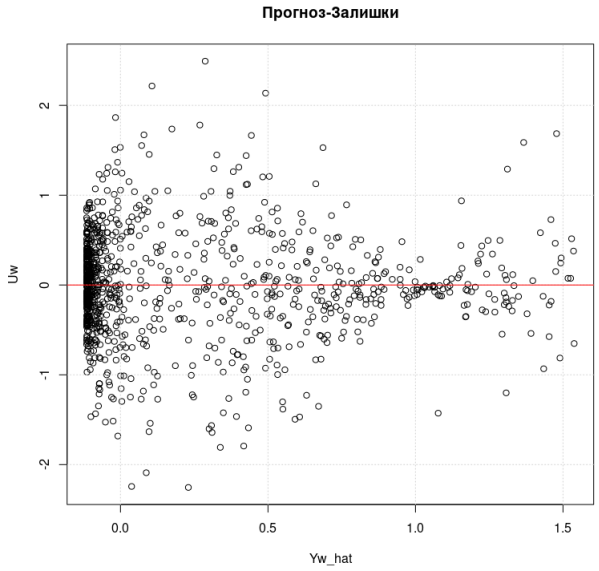

Рис. 12: Діаграма "Прогноз-залишки" на перетворених даних.

Розкид залишків набагато зменшився, неоднорідність знову трохи приглушена: проблемним місцем все ще залишається точка стику метелика. Отже для імітаційного моделювання ми будемо використовувати перетворення на вагові коефіцієнти, визначені за правилом (2).

2.2 Імітаційний експеримент.

Для кожного $N = 25, 50, 100, 500, 1000, 2000$ проводимо по $B = 1000$ разів раніше наведені операції, а саме моделювання даних, оцінюємо невідомі коефіцієнти за допомогою класичного МНК та навантаженого МНК, де вагова матриця є діагональною з компонентами вигляду (2). Зміщення оцінки від справжнього значення обчислюється за формулою:

$$\text{bias}(\hat{a}, \hat{b}) = \left(a - \overline{\hat{a}}, b - \overline{\hat{b}}\right),$$

де $\hat{a}, \hat{b}$ - вибірки зі значень оцінок коефіцієнтів зсуву і нахилу відповідно. Середньоквадратичні відхилення обчислюються через виправлену вибіркову дисперсію.

	bias, a		bias, b		sd, a		sd, b	
N	OLS	GLS	OLS	GLS	OLS	GLS	OLS	GLS
25	0.005470819	0.0002969087	-0.0003688455	-0.008537819	0.2617768	0.2482386	0.3260139	0.2595942
50	-0.005483008	-0.003730721	0.001901961	0.004038391	0.182512	0.1789935	0.2278854	0.1791523
100	0.00215513	0.0003223568	0.006089087	0.0036317966	0.126474	0.1247484	0.1606803	0.1212157
500	0.001678134	0.001733323	0.001358381	0.001893053	0.05339896	0.05173705	0.06889201	0.05104343
1000	0.001996208	0.0020768085	-0.001645795	-0.0008999662	0.03931147	0.03597701	0.05101843	0.03738925
2000	-0.0003739584	-0.0002864551	0.0006117176	0.0012356139	0.02729818	0.02484811	0.03402068	0.02383824

Табл. 1: Оцінки зміщення та середньоквадратичних відхилень оцінок коефіцієнтів за класичним та навантаженим МНК.

Видно, що теорія узгоджується з практикою: при збільшенні обсягів вибірки маємо кращі показники зсунутості та розкиду для оцінок за навантаженим МНК.

3 Висновки.

Адаптивний двокроковий МНК слід використовувати на вибірках більшого обсягу, оскільки для великих вибірок може бути менша дисперсія оцінок у порівнянні з оцінками класичного МНК. Навантажений МНК працює чудово лише тоді, коли вдається його правильно застосовувати. Тим не менш, основною проблемою є виявлення форми гетерскедастичності похибок, коректна побудова вагової матриці та обмеження на обсяги вибірки на практиці.