Универсиада по эконометрике

экономического факультета МГУ имени М. В. Ломоносова

при поддержке Департамента экономической политики и развития города Москвы

13 апреля 2024 г.

100 баллов, каждая задача по 25 баллов

У Д А Ч И !

Задача 1. Размышления о p-значении

Аналитик из Департамента экономической политики и развития города Москвы изучает влияние открытий станций Большой Кольцевой Линии метрополитена (БКЛ) на динамику цен на жилую недвижимость в Москве в 2023 г. Он делает случайную выборку домов в Москве, собирает информацию о ценах на квадратный метр в квартирах, выставленных на продажу в каждом доме из выборки, а также считает расстояния от дома до ближайшей станции БКЛ. Используя эти данные, он оценивает следующую модель:

$$Y_i = \beta_1 + \beta_2 * W_i + \varepsilon_i$$

где Y_i — это изменение средней цены за квадратный метр жилой недвижимости в доме i, W_i — бинарная переменная, принимающая значение 1, если в 2023 г. открылась станция БКЛ в пределах 15 минут от дома i, ε_i — случайный шок. Дополнительно предполагаем, что случайный шок не коррелирует с W_i и распределён нормально.

1. Исследователь рассчитал МНК-оценки и провёл статистический тест Стьюдента на значимость коэффициента β_2, получив p-значение, равное 0,007. Дайте интерпретацию этому результату.
2. Объясните, что такое p-значение.
3. С какой вероятностью p значение
   - меньше 10 процентов для нормального ε_i ?
   - меньше 10 процентов доя любого ε_i ?
   - меньше, чем некоторый x?
4. Докажите, что при верной нулевой гипотезе в тесте Стьюдента на значимость коэффициента β_2 p-значение распределено равномерно на отрезке [0,1].
5. Изобразите графически плотность распределения p-значения при верной нулевой гипотезе и при верной альтернативной гипотезе ($\beta_2 \neq 0$). Дайте необходимые пояснения.
6. Изобразите на том же графике уровень значимости α (некоторый заданный).

Задача 2. Boosted ridge or Ridged boosting?

Начинающий эконометрист Андрей располагает данными X (матрица $n \times k$) и собирается предсказать по ним y (столбец $n \times 1$)[1]. После оценки линейной модели $y = X\beta + \varepsilon$ с помощью МНК он видит, что дисперсия его прогноза $\widehat{\mathrm{Var}}\,\hat{y}$ очень велика, и хочет как-то это исправить. Андрей ознакомился с учебником по многомерному статистическому анализу и узнал о существовании следующего метода улучшения оценок:

1. В качестве начального прогноза выбирается уже построенный Андреем $\hat{y}^{(0)} = \hat{y}(X)$
2. Затем вычисляется вектор его ошибок $e^{(0)} = y - \hat{y}^{(0)}$ и оценивается отдельная модель, предсказывающая e по X (обычно выбирается модель того же типа, что и для $\hat{y}(X)$)
3. Вычисляются прогнозы этой модели $\hat{e}^{(0)} = \hat{e}(X)$
4. Новый прогноз вычисляется как сумма старого прогноза и его предсказанных ошибок, то есть $\hat{y}^{(1)} = \hat{y}^{(0)} + \hat{e}^{(0)}$

Повторяя шаги 1—4 для прогноза $\hat{y}^{(1)}$, можно получить дважды улучшенный прогноз $\hat{y}^{(2)}$. Повторяя процедуру m раз, получается m раз улучшенный прогноз, для которого на каждом шаге $e^{(k)} = y - \hat{y}^{(k)}$ и $\hat{y}^{(k+1)} = \hat{y}^{(k)} + \hat{e}^{(k)}$.

a. Пусть в качестве алгоритма предсказания на всех этапах оценивания, включая промежуточные регрессии ошибок, Андрей использует линейную регрессию, оцененную с помощью МНК. Выразите его прогноз $\hat{y}^{(m)}$ через X, y и m. Был ли смысл в улучшении?

Пусть Андрей понял свою ошибку и решил использовать так называемую ridge-регуляризованную линейную регрессию[2] на каждом шаге (включая оценку $\hat{y}^{(0)}$) с некоторым $\lambda > 0$ (одним и тем же на всех этапах).

b. Выразите прогноз $\hat{y}^{(0)}$ через X, y и λ. Свой ответ обоснуйте.
c. Какой вид теперь имеет $\hat{y}^{(m)}$? Выразите его через X, y, m и λ.
d. Вычислите ковариационную матрицу $\hat{y}^{(m)}$.
e. Является ли прогноз $\hat{y}^{(m)}$ смещенным? Если да, то чему равно его смещение?
f. Покажите, что при $m \to \infty$ смещение прогноза стремится к нулю[3] при любом $\lambda > 0$.
g. Как вы считаете, что происходит с $\hat{y}^{(m)}$ при $m \to \infty$? Почему вы так считаете?

[1] Будем считать наблюдения (X_i, y_i) независимыми и одинаково распределенными
[2] Для оценивания ridge-регрессии решается задача $e^\top e + \lambda\beta^\top\beta = \sum(y_i - X_i\beta)^2 + \lambda\sum\beta_j^2 \to \min(\beta)$
[3] Возможно, здесь будет удобно использовать сингулярное разложение и вычислить смещение в терминах сингулярных чисел матрицы X.

Задача 3. Стать мракоборцем

Студенты факультета Гриффиндор продолжают выбирать карьерный путь. Гарри Поттер хочет стать мракоборцем. Он прогнозирует долю тех студентов Гриффиндора, которые решат продолжить обучение на курсах мракоборцев и проводит анализ выпусков факультета предыдущих восьми лет, в ходе которого регистрировались значения переменных: $x_i^{(1)}$ — среднедушевые доходы в семье i-го выпускника (тысяч галеонов в год), $x_i^{(2)}$ — средний балл в его дипломе выпускника, π — доля выпускников Гриффиндора, решивших продолжить обучение на курсах мракоборцев (от общего числа окончивших факультет). Были подсчитаны следующие оценки:

$$\bar{X}(1) = \begin{pmatrix} \bar{x}^{(1)}(1) \\ \bar{x}^{(2)}(1) \end{pmatrix} = \begin{pmatrix} 50 \\ 4 \end{pmatrix} \text{ и } \bar{X}(2) = \begin{pmatrix} \bar{x}^{(1)}(2) \\ \bar{x}^{(2)}(2) \end{pmatrix} = \begin{pmatrix} 40 \\ 3 \end{pmatrix}$$

средние значения переменных $X = \left(x_i^{(1)}, x_i^{(2)}\right)^T$ отдельно по выпускникам, решившим продолжить обучение на курсах мракоборцев (класс 1), и по остальным выпускникам (класс 2), а также — ковариационные матрицы

$$\sum(1) = \sum(2) = \begin{pmatrix} 100 & 3 \\ 3 & 1 \end{pmatrix}$$

и доля $\pi = 0{,}3$.

Гарри также проверил, что предположение о нормальности распределения двумерной случайной величины X внутри каждого из классов не противоречит имеющимся исходным данным. Помогите Гарри Поттеру:

1) Оценить долю тех выпускников Гриффиндора, имея среднедушевой доход в семье 60 тыс. галеонов. и средний балл 4,2, решат продолжить обучение на курсах мракоборцев

2) Какой из известных Вам моделей соответствует способ вычисления искомой доли, следующий из решения задачи?

Задача 4. О взвешивании

Рассмотрим регрессионную модель, в которой 2n наблюдений разбиты на 2 равные группы.

В матричном виде она записывается как:$y = X\beta + \varepsilon$, где $E(\varepsilon) = \bar{0}$, $Cov(\varepsilon_s, \varepsilon_t) = 0$,

$V(\varepsilon_i) = \sigma_1^2$ для i = 1, 2,…,n, $V(\varepsilon_i) = \sigma_2^2$ для i = n+1, n+2, …, 2n

Матрицы разбиты на блоки:

$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$, $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$, $\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \end{bmatrix}$

$y_1, y_2, \varepsilon_1, \varepsilon_2$ - векторы-столбцы n*1, X_1, X_2 - матрицы n*k

А) Докажите, что МНК-оценка по всем 2n наблюдениям – это «взвешенное среднее» МНК-оценок по двум группам:$\hat{\beta} = A_1\hat{\beta}_1 + A_2\hat{\beta}_2$, где $A_1 + A_2 = I_k$ (единичная матрица размерности k*k).

Б) Докажите, что GLS (ОМНК)-оценка имеет вид: $\hat{\beta}^{GLS} = \left(\frac{X_1^T X_1}{\sigma_1^2} + \frac{X_2^T X_2}{\sigma_2^2}\right)^{-1} \left(\frac{X_1^T y_1}{\sigma_1^2} + \frac{X_2^T y_2}{\sigma_2^2}\right)$.