

МФТИ, 2026/2027
Прикладная статистика
Домашнее задание №2
Дедлайн: 14 октября, 10:00

Задача 1

Предположим, что модель генеральной совокупности, определяющая Y , имеет вид

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + u_i,$$

и эта модель удовлетворяет предположениям Гаусса–Маркова. Однако мы оцениваем модель, в которой переменная X_3 пропущена. Пусть $\tilde{\beta}_0$, $\tilde{\beta}_1$ и $\tilde{\beta}_2$ — оценки МНК из регрессии Y на X_1 и X_2 .

- а) Покажите, что условное математическое ожидание $\tilde{\beta}_1$ (при заданных значениях регрессоров в выборке) равно

$$\mathbb{E}(\tilde{\beta}_1 \mid X_1, X_2, X_3) = \beta_1 + \beta_3 \frac{\sum_{i=1}^n \hat{r}_{i1} X_{i3}}{\sum_{i=1}^n \hat{r}_{i1}^2},$$

где $\hat{r}_{i1}$ — остатки МНК-регрессии X_1 на X_2 и константу.

- б) Выразите условную дисперсию этой оценки.
с) Предполагая конечность 2-х моментов у всех регрессоров и ошибок, найдите

$$\text{plim } \tilde{\beta}_1$$

и укажите условия, при которых $\tilde{\beta}_1$ состоятельна для β_1 , несмотря на пропуск X_3 .

Задача 2

Предположим, что вас интересует оценка взаимосвязи между Y и X_1 при прочих равных условиях. Для этой цели вы можете собрать данные по двум контрольным переменным: X_2 и X_3 (для наглядности можно считать, что Y — это балл за итоговый экзамен, X_1 — посещаемость занятий, X_2 — средний балл по итогам предыдущего семестра, а X_3 — средний балл ЕГЭ.)

Пусть $\tilde{\beta}_1$ — оценка коэффициента в простой регрессии Y на X_1 , а $\hat{\beta}_1$ — оценка коэффициента в множественной регрессии Y на X_1, X_2, X_3 .

- а) Если X_1 сильно коррелирована с X_2 и X_3 в выборке, а X_2 и X_3 имеют высокую корреляцию с Y , будете ли вы ожидать, что $\tilde{\beta}_1$ и $\hat{\beta}_1$ будут похожи или сильно различаться? Объясните.

- b) Если X_1 почти не коррелирована с X_2 и X_3 , но X_2 и X_3 сильно коррелированы между собой, будут ли $\tilde{\beta}_1$ и $\hat{\beta}_1$, скорее всего, похожи или сильно различаться? Объясните.
- c) Если X_1 сильно коррелирована с X_2 и X_3 , а X_2 и X_3 слабо связаны (скоррелированы) с Y , какую из стандартных ошибок вы ожидаете меньшей:

$$\text{s. e.}(\tilde{\beta}_1) \quad \text{или} \quad \text{s. e.}(\hat{\beta}_1)?$$

Объясните.

- d) Если X_1 почти не коррелирована с X_2 и X_3 , кроме того X_2 и X_3 сильно коррелируют с Y , и X_2 и X_3 сильно коррелированы между собой, какую из стандартных ошибок вы ожидаете меньшей:

$$\text{s. e.}(\tilde{\beta}_1) \quad \text{или} \quad \text{s. e.}(\hat{\beta}_1)?$$

Объясните.

Задача 3

Используя 526 наблюдений в датасете `WAGE1` (из библиотеки `Wooldridge` в R; также приложен отдельным файлом), рассмотрим линейную модель для $\log(wage)$ с объясняющими переменными `educ` (длительность учебы), `exper` (длительность пребывания на рынке труда) и `tenure` (стаж на текущей работе).

$$\widehat{\log(wage)} = 0.284 + 0.092 \text{educ} + 0.0041 \text{exper} + 0.022 \text{tenure}, \quad n = 526. \quad (2)$$

Под длинной моделью далее имеется в виду линейная модель (2). Короткая модель — регрессия $\log(wage)$ на остатки `educ`, полученные после регрессии этой переменной на `exper`, `tenure` и константу.

- a) Проведите процедуру исключения переменных в соответствии с FWL. Для этого вначале оцените зависимость `educ` от `exper`, `tenure` (обозначим остатки в линейной модели за $\hat{r}_1$). Далее оцените зависимость $\log(wage)$ от $\hat{r}_1$. Сравните коэффициент при $\hat{r}_1$ с коэффициентом при `educ` в регрессии $\log(wage)$ на `educ`, `exper` и `tenure`.
- b) Вычислите вручную значение несмещённой оценки σ^2 по длинной модели в предположении гомоскедастичности.
- c) Вычислите вручную значение оценки σ^2 по короткой модели.
- d) Почему значения SSR в этих двух моделях равны? Выразите их с использованием проекторов.
- e) Как думаете, в чём причина того, что статистический пакет выдаёт разные значения стандартных ошибок у оценок коэффициентов при `educ` в этих двух моделях?

- f) Вычислите вручную значение оценки условной ковариационной матрицы оценок коэффициентов длинной модели

$$\widehat{\text{Var}}(\hat{\beta} | X)$$

в предположении гетероскедастичности ошибок. Используйте

$$\text{Var}(\hat{\beta} | X) = (X^\top X)^{-1} X^\top \mathbb{E}(\varepsilon \varepsilon^\top | X) X (X^\top X)^{-1}.$$

Сравните найденные стандартные отклонения $\hat{\beta}$ со значениями, получаемыми с использованием встроенных функций (в R это `vcovHC` с опцией `HC1` из библиотеки `sandwich`; в Python — например стандартная функция из `statsmodels` с опцией `HC0`. Учтите, что разные варианты опций дают поправку на малые выборки; попробуйте найти вариант, точно совпадающий с вашей оценкой).

Задача 4

Файл `galton_data.csv` содержит данные, собранные Ф. Гальтоном в 1885 году для исследования связи между ростом родителей и их детей. В рамках этой задачи мы исследуем некоторые аспекты связи между этими переменными.

- a) Гальтон в оригинальном исследовании использовал исправленное среднее значение роста родителей, в котором рост матери брался с коэффициентом 1.08 (отражающим соотношение среднего роста мужчин и женщин в популяции). Постройте переменную “средний рост родителей” с такой поправкой (назовём её X), и сделайте таблицу описательных статистик для этой переменной и переменной роста детей (переменная Y).

- b) Стандартизируйте переменные X и Y (стандартизованные переменные назовём $\tilde{X}$ и $\tilde{Y}$ соответственно). Вычислите доверительный интервал для параметра

$$\rho := \text{Corr}(\tilde{X}, \tilde{Y}),$$

используя бутстрап. Сформулируйте гипотезу о некоррелированности X и Y и проверьте её с помощью построенного доверительного интервала.

- c) Оцените регрессионную модель

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

вначале для случая, когда X_i — средний рост родителей, а Y_i — рост детей, а потом для случая, когда Y и X поменяли местами (далее обозначаются как “прямая” и “обратная” модель).

Проинтерпретируйте полученные значения $\hat{\beta}_0$ и $\hat{\beta}_1$ в каждом из двух случаев. Можно ли коэффициенты β_1 интерпретировать как среднее изменение Y при изменении X на 1?

- d) Верно ли, что произведение полученных оценок β_1 в “прямой” и “обратной” модели имеет абсолютное значение меньше единицы? Докажите, что такое свойство у оценок в простой регрессии (с константой) должно выполняться всегда: для этого сравните произведение оценок коэффициентов наклона в прямой и обратной модели с выборочным коэффициентом корреляции.

Задача 5

Рассмотрим линейную модель

$$y = X\beta + \varepsilon, \quad \mathbb{E}[\varepsilon \mid X] = 0, \quad \mathbb{E}[\varepsilon\varepsilon^\top \mid X] = \text{diag}(\sigma_1^2, \dots, \sigma_n^2) =: \Omega,$$

где X — матрица размера $n \times k$ ранга k , $\hat{\beta}$ — МНК-оценка,

$$\hat{\varepsilon} = Y - X\hat{\beta}$$

— вектор остатков.

Предполагаются условия регулярности:

$$\frac{X^\top X}{n} \xrightarrow{p} Q > 0$$

и

$$\frac{1}{n} \sum_{i=1}^n X_i X_i^\top \sigma_i^2 \xrightarrow{p} \Omega_0,$$

так что (сэндвич Хубера–Уайта)

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, Q^{-1}\Omega_0 Q^{-1}).$$

Рассмотрим две бутстрап-схемы:

а) **Бутстрап на остатках:**

$$Y_i^* = X_i^\top \hat{\beta} + \varepsilon_i^*,$$

где $\varepsilon_1^*, \dots, \varepsilon_n^*$ н.о.р. из эмпирического распределения центрированных остатков

$$\tilde{\varepsilon}_i = \varepsilon_i - \bar{\varepsilon}.$$

б) **Парный бутстрап:**

строки

$$(Y_i^*, X_i^*)$$

— н.о.р. из эмпирического распределения

$$\{(Y_i, X_i)\}_{i=1}^n.$$

Для каждой схемы вычислите условную ковариационную матрицу

$$V_n^* = \text{Var}^*\left(\sqrt{n}(\hat{\beta}^* - \hat{\beta})\right),$$

найдите её предел по вероятности и сравните с сэндвичем

$$Q^{-1}\Omega_0 Q^{-1}.$$

Какая схема даёт состоятельную аппроксимацию асимптотического закона

$$\sqrt{n}(\hat{\beta} - \beta)$$

при гетероскедастичности? При каком условии схемы эквивалентны?

Сделайте Монте-Карло симуляцию для этих двух схем: используйте многомерное нормальное распределение для X и ε , причём

$$\sigma_i(X_i) = |X_i|^\top \gamma$$

для некоторого фиксированного вектора (положительных) параметров.

Вычислите бутстрапные доверительные интервалы. В каком из случаев фактический уровень покрытия ближе к расчётному?