Универсиада по эконометрике

9 апреля 2022 г.

Задача 1. Дедлайн (25%)

Консультант Антон оценил регрессию вида

$$y_i = x_i'\beta + \varepsilon_i,$$

где $x_i \in \mathbb{R}^5$, и все стандартные предпосылки теоремы Гаусса-Маркова выполнены, и по заданию от менеджера протестировал гипотезы

$$H_0 : \ \beta_1 + \beta_2 = 1$$

$$H_1 : \ \beta_1 + \beta_2 > 1$$

Получившееся p-value оказалось равно 0.52.

За 5 минут до дедлайна Антон заметил, что, сохраняя данные в Excel, он случайно скопировал их дважды. То есть вместо 124 наблюдений он использовал 248, так что каждое наблюдение повторяется ровно два раза.

1. Как это повлияло на его оценки $\hat{\beta}_1, \hat{\beta}_2$, их дисперсии и значимость?
2. Предположим, что менеджера волнует только ответ на вопрос, какую из гипотез следует принять. Стоит ли консультанту Антону исправлть свои выводы? Если ваш ответ да, объясните, как именно исправить, если нет, объясните почему.

Задача 2. Регрессия по двум точкам (25%)

Замечение: если не справляетесь с каким-то пунктом, следует принять его результат как данный и двигаться дальше.

Рассмотрим стандартную модель линейной регрессии

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

где x_i – детерминированные и ошибки $\varepsilon_i \sim$ i.i.d. (независимые одинаково распределённые) и удовлетворяют $\mathbb{E}(\varepsilon_i) = 0$ и $\mathbb{E}(\varepsilon_i^2) = \sigma^2$.

1. Является ли оценка $\hat{\beta}_{OLS}$ несмещенной? эффективной (в классе линейных по несмещённых оценок)? состоятельной? Обоснуйте Ваши ответы.

2. Рассмотрим альтернативную процедуру оценки: из выборки размера n берутся два наблюдения наугад и строится прямая, соединяющая взятые точки. Она и будет "линией регрессии". Выпишите уравнение для полученной таким образом оценки $\tilde{\beta}$, считая что выбраны наблюдения (x, y) и (x', y'). Является ли она линейной? несмещенной? эффективной (в классе линейных по несмещённых оценок)? состоятельной?

3. Пусть теперь $\tilde{\beta}$ строится следующим образом. Выберем пару наблюдений наугад, например (x_m, y_m) и (x'_m, y'_m) и применим к ним процедуру из предыдущего пункта (2). Назовем получившуюся оценку $\tilde{\beta}_m$. Повторим эту процедуру M раз (брать те же пары точек повторно разрешается, каждый раз выбираем независимо от предыдущих) и усредним
$$\tilde{\beta} = \frac{1}{M} \sum_{m=1}^{M} \tilde{\beta}_m$$
Является ли эта оценка линейной по y_i? несмещенной? Найдите предел по вероятности от $\tilde{\beta}$ при $M \to \infty$ и фиксированном n (Подсказка: выборка зафиксирована!)

4. Какая оценка «лучше», $\tilde{\beta}$ или $\hat{\beta}_{OLS}$ и по какому критерию? (Подсказка: есть ли какая-нибудь теорема, которая позволит их сравнить?)

Замечание: в ходе решения некоторых пунктов можно (но не обязательно) использовать без доказательства неравенство: для любой случайной величины X с конечным вторым моментом
$$P(|X| > a) = P\left(X^2 > a^2\right) \leqslant \frac{\mathbb{E}\left(X^2\right)}{a^2}$$

Задача 3. Комбинация прогнозов (25%)

Инвестор прогнозирует премию за риск инвестиций в акции и оценивает 2 регрессии:

$$y_t = \alpha_1 + \beta_1 x_{t-1} + \varepsilon 1_t$$

$$y_t = \alpha_2 + \beta_2 z_{t-1} + \varepsilon 2_t$$

где y_t – совокупная доходность индекса S&P 500; x_t – логарифм отношения дивидендов к цене; z_t – логарифм дивидендной доходности (разница между логарифмом дивидендов и логарифмом лагированных цен на акции). Инвестор определяет вектор ошибок прогноза как:

$$e_1 = y - f_1$$

$$e_2 = y - f_2$$

где f_1 , f_2 - 2 вектора расчётных значений из первой и второй регрессий; y – вектор наблюдаемых значений зависимой переменной.

Инвестор знает, что сложно найти единственный наилучший прогноз и что найдётся несколько моделей с сопоставимой точностью прогноза. Поэтому он строит комбинацию прогнозов:

$$f_c = \omega * f_1 + (1 - \omega) * f_2$$

$$e(\omega) = y - f_c$$

где $e(\omega)$ – ошибка прогноза, ω и $(1-\omega)$ – веса прогнозов. Среднеквадратическая ошибка прогноза (MSE) определяется как $E[e(\omega)^2]$

1. Предположим, что оба прогноза являются несмещёнными, т. е. $E[e_1] = E[e_2] = 0$. Пусть теоретические дисперсии и ковариация ошибок прогноза: $Var(e_1) = \sigma_1^2$, $Var(e_2) = \sigma_2^2$, $Cov(e_1, e_2) = \sigma_{12}$. Найдите оптимальные веса, которые минимизируют MSE. Выразите их через σ_1^2, σ_2^2, σ_{i2}.

2. Инвестор решил использовать для комбинации прогноза 2 типа весов:

   - Равные веса (по 0.5)
   - Метод Бейтса и Грейнджера: используя доступные данные, нашёл ковариационную матрицу ошибок: $\begin{pmatrix} 6662 & 6618 \\ 6618 & 6597 \end{pmatrix}$ и посчитал веса по формуле из

пункта (1).

Найдите MSE комбинированного прогноза, полученного этими двумя способами. *(Используйте данную ковариационную матрицу для обоих типов весов).*

3. Сравните эффективность двух подходов, использованных в предыдущем вопросе: Является ли комбинация прогнозов более предпочтительной, чем отдельный прогноз? Какой тип весов обеспечивает лучшую комбинацию прогнозов? Всегда ли один тип весов работает лучше, чем другой?

Задача 4. Переобучение безработных (25%)

Виктор интересуется, как программы переобучения безработных в стране A влияют на их зароботок после нового трудоустройства. В базе есть следующие показатели: Y − заработная плата (в условых единицах) после того как человек нашёл новую работу; D − бинарная переменная, равная 1, если безработный фактически прошёл программу переобучения, и 0, если нет; X_1 − возраст (лет), X_2 − бинарная переменная пола: 1 для мужчин и 0 для женщин; Z − бинарная переменная, равная 1, если в службе занятости страны A безработному предложили пройти программу переобучения, и 0, если не предложили (далее он мог согласиться или не согласиться или же без предложения самостоятельно попросить записать его на переобучение); n − число людей в однородных группах (с одинаковыми характеристиками).

Виктор считает, что пол и возраст влияют на то, прошёл ли безработный фактически программу переобучения. Он преложил оценить влияние переобучения безработных на заработную плату после трудоустройства следующим способом: Сначала с помощью модели бинарного выбора оценивается вероятность безработного фактически пройти программу переобучения в зависимости от пола и возраста. Затем берутся пары безработных (1 к 1) с одинаковой расчётной вероятностью такие, что один человек из пары фактически прошёл переобуение ($D = 1$), а другой - нет ($D = 0$). И среди получившихся пар надо посчитать оценку для выражения:

$$\alpha_1 = E[Y|D = 1] - E[Y|D = 0] = \frac{E[DY]}{P(D = 1)} - \frac{E[(1 - D)Y]}{P(D = 0)}$$

(Можно использовать выражение без доказательства.)

Он оценил пробит-модель бинарного выбора с зависимой переменной D и показал следующие частичные результаты в таблицах 1 и 2:

ТАБЛИЦА 1. Частичные результаты для тех, кто фактически прошёл переобучение ($D = 1$)

№ (название группы)	Y	$\hat{P}(D = 1\|X)$	X_1	X_2	n
G1	10	0.50	25	0	20
G2	9	0.84	21	1	30

и коэффициенты для пробит-регрессии $\hat{\beta}_0 = -1.25$, $\hat{\beta}_1 = 0.05$, $\hat{\beta}_2 = 1.2$

ТАБЛИЦА 2. Частичные результаты для тех, кто не прошёл переобучение ($D = 0$)

№ (название группы)	Y	$\hat{\beta}_0 + \hat{\beta}_1 * X_1 + \hat{\beta}_2 * X_2$	n
A1	7	-1.000	200
A2	8	0.000	30
A3	2	1.000	40
A4	8	1.668	30
A5	4	1.815	30

1. Найдите оценку влияния переобучения безработных на заработную плату после трудоустройства способом, который предлагает Виктор

2. Можно ли сказать, что выводы из оценки Виктора можно применить ко всем безработным в стране А? Почему?

3. Можно ли вместо способа Виктора посчитать оценку как разницу средних заработных плат между теми, кто прошёл и не прошёл переобучение? Почему?

4. Коллега Виктора сомневается в результатах его расчётов и считает, что между фактическим переобучением и будущей заработной платой есть эндогенность: существуют такие ненаблюдаемые характеристики человека, как мотивация и способности, а они могли повлиять и на решение переобучаться, и на будущую зарплату. Предложите способ оценки, который решал бы эту проблему. Какие предпосылки об имеющихся данных нужно сделать, чтобы воспользоваться этим методом?