Универсиада «Ломоносов» по эконометрике 2026

Отборочный этап, сумма 30 баллов

Решение и критерии

Задание

Заполните ПРОПУСКИ и напишите обоснование для каждого их них. На месте пропуска может стоять одно или несколько слов или фраза.

Каждый пропуск по 1 баллу, обоснование по 2 балла.

Сергей Т. – начинающий эконометрист и опытный HR. Сергей оценивает методом наименьших квадратов (МНК) сперва парную регрессию с константой $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$, где y_i – заработная плата i-го сотрудника компании, а x_i – опыт работы i-го сотрудника компании.
Затем он оценивает парную регрессию с константой $x_i = \alpha_1 + \alpha_2 y_i + u_i$, где ε_i и u_i – случайные шоки.
Между МНК-оценками, полученными Сергеем, будет выполняться соотношение $\hat{\alpha}_2 = \frac{1}{\hat{\beta}_2}$, если ПРОПУСК 1.

РЕШЕНИЕ
Способ от составителя:
Можно вспомнить свойство, что эр-квадрат в парной регрессии равен квадрату коэффициента корреляции, и он одинаков в двух регрессиях: y от x и x от y. Отсюда можно получить связь между МНК-оценками коэффициентов β_2 и α_2.
$\hat{\alpha}_2 = \frac{\widehat{\mathrm{cov}}(x,y)}{\widehat{\mathrm{var}}(y)}$, $\hat{\beta}_2 = \frac{\widehat{\mathrm{cov}}(x,y)}{\widehat{\mathrm{var}}(x)}$

$$R^2 = (\widehat{\mathrm{corr}}(x,y))^2 = \left(\frac{\widehat{\mathrm{cov}}(x,y)}{\sqrt[2]{\widehat{\mathrm{var}}(x) * \widehat{\mathrm{var}}(y)}}\right)^2 = \frac{\widehat{\mathrm{cov}}(x,y)}{\widehat{\mathrm{var}}(x)} * \frac{\widehat{\mathrm{cov}}(x,y)}{\widehat{\mathrm{var}}(y)} = \hat{\beta}_2 * \hat{\alpha}_2$$

Откуда: $\hat{\alpha}_2 = \frac{R^2}{\hat{\beta}_2}$.
Ответ: … если $R^2 = 1$ (или корреляция равно +1 или -1)
Зачёт по смыслу: если все точки с координатами (x,y) лежат на одной прямой линии и т.д.
Альтернативно:
$\hat{\alpha}_2 = \frac{1}{\hat{\beta}_2}$ выполняется, если линия регрессии проходит через ноль ($\beta_1 = 0$ и $\alpha_1 = 0$) или что $(\sum_{i=1}^n x_i y_i)^2 = \sum_{i=1}^n y_i^2 * \sum_{i=1}^n x_i^2$

С подробностями:
$y_i = \beta_2 x_i + \varepsilon_i$, откуда $\hat{\beta}_2 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$.
$x_i = \alpha_2 y_i + u_i$, откуда $\hat{\alpha}_2 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^2}$.

$$\hat{\alpha}_2 = \frac{1}{\hat{\beta}_2}$$

$$\frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n y_i^2} = \frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n x_i y_i}$$

$$(\sum_{i=1}^n x_i y_i)^2 = \sum_{i=1}^n y_i^2 * \sum_{i=1}^n x_i^2$$

Затем Сергей решает вместо уравнения $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ оценить регрессию по центрированным данным, т. е. в отклонениях заработной платы от средней по компании и отклонениях опыта работы от среднего по компании.

В результате МНК-оценка коэффициента β_1 ПРОПУСК 2, МНК-оценка коэффициента β_2 ПРОПУСК 3, коэффициент детерминации ПРОПУСК 4, стандартная ошибка МНК-оценки коэффициента β_2 ПРОПУСК 5.

РЕШЕНИЕ

Это задние по смыслу, как изменятся оценки коэффициентов и прочие величины, если из игрека вычесть константу и из икса вычесть константу.

До центрирования:

$$\hat{\beta}_2 = \frac{\widehat{\mathrm{cov}}(x,y)}{\widehat{\mathrm{var}}(x)},\ \hat{\beta}_1 = \bar{y} - \hat{\beta}_2 * \bar{x},\ R^2,\ s.e.(\hat{\beta}_2) = \sqrt{\frac{ESS/(n-2)}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

После центрирования:

$\hat{\beta}_{2центр} = \frac{\widehat{\mathrm{cov}}(x-\bar{x},y-\bar{y})}{\widehat{\mathrm{var}}(x-\bar{x})} = \frac{\widehat{\mathrm{cov}}(x,y)}{\widehat{\mathrm{var}}(x)} = \hat{\beta}_2$. От вычитания константы ни дисперсия, ни ковариация не изменяются. То есть МНК-оценка коэффициента наклона не изменяется.

После центрирования средние становятся нулями. Поэтому $\hat{\beta}_{1центр} = 0$.

От вычитания константы ни дисперсия, ни ковариация не изменяются, поэтому коэффициент детерминации не изменяется:

$$R^{2центр} = (\widehat{\mathrm{corr}}(x-\bar{x}, y-\bar{y}))^2 = (\widehat{\mathrm{corr}}(x,y))^2 = R^2$$

Сумма квадратов остатков (ESS) по той же причине не изменяется. Квадраты отклонения иксов от среднего значения тоже не изменяются, поэтому $s.e.(\hat{\beta}_{2центр})$ без изменений.

Ксения ищет работу аналитиком и спрашивает у Сергея, как обойти ИИ-алгоритм, который используют HR-ры, чтобы её резюме прошло отбор на живое собеседование. Ксения имела опыт работы в Жёлтом банке и умеет строить модели машинного обучения. Сергей планирует исследовать, как вероятность приглашения на живое собеседование зависит от указания в резюме опыта работы с моделями машинного обучения. Для этого он оценивает логит-модель:

$$P(z_i=1) = \Lambda\ (\alpha + \beta w_i + \varepsilon_i)$$

где:

z_i принимает только значения 1, если кандидата i пригласили, и 0, если нет; w_i принимает только значения 1, если у кандидата i есть опыт работы с моделями машинного обучения, и 0, если нет;

Λ (…) – логистическая функция; α и β – коэффициенты модели; ε_i – случайный шок.

Сергей собрал выборку. Сведения о количестве кандидатов с соответствующим сочетанием характеристик представлены в таблице:

	$z_i = 0$	$z_i = 1$
$w_i = 1$	m_1	m_2
$w_i = 0$	m_3	m_4

Сергей сможет получить оценки логит-модели бинарного выбора по этим данным, если ПРОПУСК 6.

РЕШЕНИЕ:

… ЕСЛИ все 4 числа m_1, m_2, m_3, m_4 не нулевые.

Пояснение:

Если формально обосновано отсутствие экстремума (ссылкой на монотонность или неразрешимость FOC) — 2 балла за объяснение

Если объяснение не полностью и не строгое, то 1 балл.

	$z_i = 0$	$z_i = 1$
$w_i = 1$	m_1	m_2
$w_i = 0$	m_3	m_4

Ищем оценки методом максимального правдоподобия:

$$L = \prod_{y=1}(\Lambda(\alpha + \beta w_i) * \prod_{y=0}(1 - \Lambda(\alpha + \beta w_i)) =$$

$$(\Lambda(\alpha))^{m_2} * \left(\Lambda(\alpha + \beta)\right)^{m_4} * (1 - \Lambda(\alpha))^{m_1} * (1 - \Lambda(\alpha + \beta))^{m_3} \underset{\alpha,\beta}{\rightarrow} \max$$

Логарифмируем, чтобы перейти к суммам вместо произведений:

$$\ln L = m_2 \ln \Lambda(\alpha) + m_4 \ln \Lambda(\alpha + \beta) + m_1 (1 - \ln \Lambda(\alpha)) + m_3 \ln(1 - \Lambda(\alpha + \beta)) \underset{\alpha,\beta}{\rightarrow} \max$$

обозначим $\gamma = \alpha + \beta$. Тогда:

$$\ln L = m_2 \ln \Lambda(\alpha) + m_4 \ln \Lambda(\gamma) + m_1 (1 - \ln \Lambda(\alpha)) + m_3 \ln(1 - \Lambda(\gamma)) \underset{\alpha,\gamma}{\rightarrow} \max$$

Обозначим f – производную от Λ.

Тогда условия первого порядка (FOC):

$$\frac{\partial \ln L}{\partial \alpha} = m_2 \frac{f(\alpha)}{\Lambda(\alpha)} - m_1 \frac{f(\alpha)}{1 - \Lambda(\alpha)} = 0$$

$$\frac{\partial \ln L}{\partial \gamma} = m_4 \frac{f(\gamma)}{\Lambda(\gamma)} - m_3 \frac{f(\gamma)}{1 - \Lambda(\gamma)} = 0$$

Приведём к общему знаменателю, сократим первое на $f(\alpha)$, второе на $f(\gamma)$. Имеем:

$m_2(1 - \Lambda(\alpha)) - m_1\Lambda(\alpha) = 0$, откуда $\Lambda(\alpha) = \frac{m_2}{m_2 + m_1}$.

Тогда если $m_2 = 0$, то $\Lambda(\alpha) = \frac{1}{1+e^{\alpha}} = 0$, что не может выполняться.

Если $m_1 = 0$, то $\Lambda(\alpha) = \frac{1}{1+e^{\alpha}} = 1$, то есть $e^{\alpha}=0$, что не может выполняться.

Аналогично со вторым выражением из FOC.

Выкладки могут отличаться, но смысл такой же.

Приятель Ксении по имени Никита обеспокоен растущими ценами. Он оценил регрессию:

$$\hat{Q}_i = 5000 + 0.08Y_i - 150PF_i + 90PNF_i$$

где Q_i – затраты на продукты питания в регионе i, Y_i – располагаемый доход в регионе i, PF_i – индекс цен на продукты питания в регионе i, PNF_i – индекс цен на непродовольственные товары в регионе i.

Расчётные t-статистики для оценок коэффициентов равны соответственно: 4.17; 1.6; -1.25 и 0.82.

Коэффициент детерминации равен 0.88, число регионов в выборке Никиты равно 30.

Подобные результаты возможны из-за ПРОПУСК 7. Преодолеть эту проблему ПРОПУСК 8.

Исходя из изначально полученных результатов, 95%-ный доверительный интервал для коэффициента при располагаемом доходе ПРОПУСК 9.

РЕШЕНИЕ

Подобные результаты возможны из-за МУЛИТИКОЛЛИНЕАРНОСТИ. Пояснение: признаками мультиколлинеарности являются незначимые коэффициенты и одновременно высокий эр-квадрат.

Преодолеть эту проблему МОЖНО, НАПРИМЕР, УДАЛИВ ОДИН ИЗ ИНДЕКСОВ ЦЕН ИЛИ ПЕРЕЙДЯ К РЕАЛЬНЫМ ПОКАЗАТЕЛЯМ ВМЕСТО НОМИНАЛЬНЫХ. Ещё одна проблема – это оценки по малому числу точек (30).

Исходя из изначально полученных результатов, 95%-ный доверительный интервал для коэффициента при располагаемом доходе СОСТАВЛЯЕТ (-0.0228; 0.1828). Расчёт: стандартная ошибка для оценки коэффициента при располагаемом доходе = 0.08/1.6 = 0.05. Значение из таблицы Стьюдента для 26 степеней свободы 2.056. Тогда доверительный интервал: 0.08 – 2.056 * 0.05; 0.08 + 2.06*0.05.

Никита переходит к изучению потребления на макроуровне. Он столкнулся с системой уравнений, которые описывают закрытую экономику:

$$Y_i = \beta_1 + \beta_2 C_i + \beta_3 I_i + \beta_4 G_i + \varepsilon_i$$

$$C_i = \alpha_1 + \alpha_2 Y_i + u_i,$$

где Y_i – ВВП, C_i – потребление, I_i – инвестиции, G_i – госрасходы. Случайные шоки ε_i и u_i, некоррелированные. Опытная макроэкономистка Лиза говорит Никите, что при оценке второго уравнения по МНК оценка коэффициента α_2 будет смещена в сторону ПРОПУСК 10.

Стандартная ситуация эндогенности:

$$Y_i = \beta_1 + \beta_2 C_i + \beta_3 I_i + \beta_4 G_i + \varepsilon_i = \beta_1 + \beta_2(\alpha_1 + \alpha_2 Y_i + u_i) + \beta_3 I_i + \beta_4 G_i + \varepsilon_i$$

$$Y_i = \frac{(\beta_1 + \beta_2\alpha_1) + \beta_3 I_i + \beta_4 G_i + \beta_2 u_i + \varepsilon_i}{1 - \beta_2\alpha_2}$$

$$cov(Y_i; u_i) = \frac{\beta_2 var(u_i)}{1 - \beta_2\alpha_2}$$

$$E(\widehat{\alpha}_2) = \mathrm{E}\frac{\widehat{\mathrm{cov}}(\mathrm{C,Y})}{\widehat{\mathrm{var}}(\mathrm{Y})} = \frac{cov(\alpha_1 + \alpha_2 Y_i + u_i, \mathrm{Y})}{var(\mathrm{Y})} = \alpha_2 + \frac{cov(u_i, \mathrm{Y})}{var(\mathrm{Y})} = \alpha_2 + \frac{\beta_2 var(u_i)}{(1-\beta_2\alpha_2)var(\mathrm{Y})}.$$

Дисперсии >0, $\beta_2 > 0$, и 0<$\alpha_2 < 1$ (это предельная склонность к потреблению). Соответственно, если $\beta_2 * \alpha_2 < 1$, то смещение вверх (завышение).
Штраф: если путаница с крышками - минус 0,5
Вероятностный предел + 1 балл
Правильное указание всех знаков + 0,5 баллов
Правильное утверждение про положительную корреляцию Y и u (можно словами, можно с помощью формулы) + 0,5 баллов
Правильный ответ сам по себе +1 балл (здесь мы засчитываем "зависит от величины коэффициентов" и "положительное смещение/завышена").