Универсиада по эконометрике 2026
Решения задач второго тура

Задача 1. (25 баллов)

(1)

Тестирование значимости влияния финансовой поддержки в рамках предлагаемой модели эквивалентно тестированию гипотезы об обнулении коэффициента регрессии при переменной «получение финансовой поддержки», при этом из текста задания

1. Для малых предприятий $\hat{\beta}_{\text{мал.}} = 1.65$ и $\widehat{\text{var}}\,\hat{\beta}_{\text{мал.}} = 0.09$, поэтому можно сделать стандартный t-тест:

$$t^{\text{мал.}}_{\text{расч.}} = \frac{\hat{\beta}_{\text{мал.}}}{\widehat{\text{se}}\,\hat{\beta}_{\text{мал.}}} = \frac{1.65}{\sqrt{0.09}} \approx 5.5 \qquad \text{(1 балл)}$$

   Поскольку выборка достаточно велика, 1%-ое критическое значение приблизительно равно 2.58, и $t^{\text{мал.}}_{\text{расч.}} > 2.58$. Иными словами, эффект получения финансовой поддержки для малых предприятий значим на уровне 1% (1 балл)

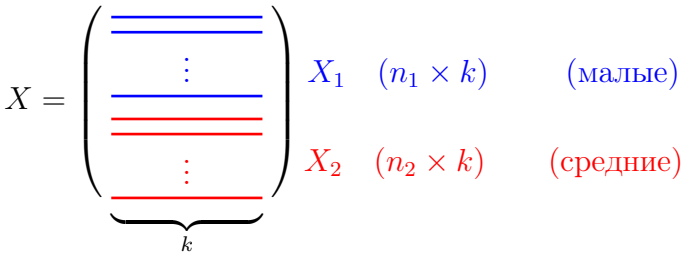
2. Для средних предприятий $\hat{\beta}_{\text{сред.}} = 0.55$ и $\widehat{\text{var}}\,\hat{\beta}_{\text{сред.}} = 0.063$, поэтому вновь можно сделать t-тест:

$$t^{\text{сред.}}_{\text{расч.}} = \frac{\hat{\beta}_{\text{сред.}}}{\widehat{\text{se}}\,\hat{\beta}_{\text{сред.}}} = \frac{0.55}{\sqrt{0.063}} \approx 2.19 \qquad \text{(1 балл)}$$

   Поскольку выборка достаточно велика, 5%-ое критическое значение приблизительно равно 1.96, и $1.96 < t^{\text{сред.}}_{\text{расч.}} < 2.58$. Иными словами, эффект получения финансовой поддержки для малых предприятий значим на уровне 5% (1 балл)

②

Восстановить оценки по полной выборке оказывается возможно, и помогает в этом блочная структура выборки. По условию Иван Сергеевич не располагает полными данными и вынужден работать исключительно с результатами оценивания регрессий на подвыборках. Рассмотрим сначала структуру оценки по всей выборке:



Из соображений блочной структуры выборки тогда имеем

$$X^\top X = X_1^\top X_1 + X_2^\top X_2 \qquad \text{и} \qquad X^\top y = X_1^\top y_1 + X_2^\top y_2$$

Получается, $\hat\beta = (X^\top X)^{-1} X^\top y = \left(X_1^\top X_1 + X_2^\top X_2\right)^{-1}\left(X_1^\top y_1 + X_2^\top y_2\right)$ (2 балла). Иными словами, если Иван Сергеевич сможем выразить $X_1^\top X_1$, $X_2^\top X_2$, $X_1^\top y_1$ и $X_2^\top y_2$ через известные ему данные, то он найдет и $\hat\beta$ – оценку по полной выборке.

Можно заметить, что[1]

1. Поскольку $\hat{\mathbb{V}}\hat\beta_1 = \frac{\text{ESS}_1}{n_1-k}(X_1^\top X_1)^{-1}$, то $X_1^\top X_1 = \left(\frac{n_1-k}{\text{ESS}_1}\hat{\mathbb{V}}\hat\beta_1\right)^{-1}$ – эту величину можно посчитать по тому, что предоставили стажеры (2 балла)
2. Поскольку $\hat\beta_1 = (X_1^\top X_1)^{-1} X_1^\top y_1$, то $X_1^\top y_1 = X_1^\top X_1\,\hat\beta_1$, но $X_1^\top X_1$ уже выражено выше, поэтому $X_1^\top y_1 = \left(\frac{n_1-k}{\text{ESS}_1}\hat{\mathbb{V}}\hat\beta_1\right)^{-1}\hat\beta_1$ (1 балл)
3. Аналогично $X_2^\top X_2 = \left(\frac{n_2-k}{\text{ESS}_2}\hat{\mathbb{V}}\hat\beta_2\right)^{-1}$ и $X_2^\top y_2 = \left(\frac{n_2-k}{\text{ESS}_2}\hat{\mathbb{V}}\hat\beta_2\right)^{-1}\hat\beta_2$

Получается, что для оценки $\hat\beta$ по полной выборке данных о регрессиях стажеров вполне хватает и результат выражается через них в виде

$$\hat\beta = \left[\left(\frac{n_1-k}{\text{ESS}_1}\hat{\mathbb{V}}\hat\beta_1\right)^{-1} + \left(\frac{n_2-k}{\text{ESS}_2}\hat{\mathbb{V}}\hat\beta_2\right)^{-1}\right]^{-1}\left[\left(\frac{n_1-k}{\text{ESS}_1}\hat{\mathbb{V}}\hat\beta_1\right)^{-1}\hat\beta_1 + \left(\frac{n_2-k}{\text{ESS}_2}\hat{\mathbb{V}}\hat\beta_2\right)^{-1}\hat\beta_2\right]$$

(1 балл)

[1] Здесь и далее под ESS имеется в виду сумма квадратов остатков (*errors sum of squares*), то есть в терминологии задачи $\text{ESS} = \hat e^\top \hat e = (y-\hat y)^\top (y-\hat y)$.

③

Для того чтобы тестировать посчитанные в предыдущем пункте оценки на значимость t-тестом, требуется знание (оценок) дисперсий $\widehat{\mathrm{var}}\,\hat\beta_j$, а для F-тестов на линейное ограничение – потенциально вся ковариационная матрица $\hat{\mathbb{V}}\,\hat\beta$ из модели, которая оценивалась на всех данных (1 балл). Покажем, что и ее тоже можно восстановить. Искомая матрица имеет вид

$$\hat{\mathbb{V}}\hat\beta = \frac{\mathrm{ESS}}{n_1+n_2-k}\big(X^\top X\big)^{-1} = \frac{\mathrm{ESS}}{n_1+n_2-k}\left[\left(\frac{n_1-k}{\mathrm{ESS}_1}\,\hat{\mathbb{V}}\hat\beta_1\right)^{-1} + \left(\frac{n_2-k}{\mathrm{ESS}_2}\,\hat{\mathbb{V}}\hat\beta_2\right)^{-1}\right]^{-1} \quad \text{(2 балла)}$$

Получается, Ивану Сергеевичу требуется вычислить ESS для модели, оцененной на объединенных данных (при этом надо отметить, что $\mathrm{ESS} \neq \mathrm{ESS}_1 + \mathrm{ESS}_2$, поэтому это нетривиальная задача). Рассмотрим формулу ESS:

$$\begin{aligned}\mathrm{ESS} = \hat e^\top \hat e = (y-\hat y)^\top (y-\hat y) = \big(y - X\hat\beta\big)^\top \big(y - X\hat\beta\big) = \\ = y^\top y - 2\hat\beta^\top X^\top y + \hat\beta^\top X^\top X\hat\beta = y^\top y - (X^\top y)^\top (X^\top X)^{-1} X^\top y \qquad \text{(1 балл)}\end{aligned}$$

В предыдущем пункте уже было показано, что $\hat\beta$, $X^\top X$ и $X^\top y$, посчитанные по полной выборке, выражаются через данные, предоставленные стажерами. Остается выразить через них и $y^\top y$. Поскольку $y^\top y = y_1^\top y_1 + y_2^\top y_2$, можно воспользоваться аналогичными формулами для ESS_1 и ESS_2 и выразить их через известные данные:

$$y_{1,2}^\top y_{1,2} = \mathrm{ESS}_{1,2} + (X_{1,2}^\top y_{1,2})^\top (X_{1,2}^\top X_{1,2})^{-1} X_{1,2}^\top y_{1,2} = \mathrm{ESS}_{1,2} + \frac{\mathrm{ESS}_{1,2}}{n_{1,2}-k}\,\hat\beta_{1,2}^\top \big(\hat{\mathbb{V}}\hat\beta_{1,2}\big)^{-1}\hat\beta_{1,2}$$

$$\Longrightarrow\; y^\top y = \mathrm{ESS}_1 + \frac{\mathrm{ESS}_1}{n_1-k}\,\hat\beta_1^\top \big(\hat{\mathbb{V}}\hat\beta_1\big)^{-1}\hat\beta_1 + \mathrm{ESS}_2 + \frac{\mathrm{ESS}_2}{n_2-k}\,\hat\beta_2^\top \big(\hat{\mathbb{V}}\hat\beta_2\big)^{-1}\hat\beta_2 \qquad \text{(2 балла)}$$

Подставляя это в формулу ESS выше, а результат – в формулу для $\hat{\mathbb{V}}\hat\beta$, находим ковариационную матрицу оценок коэффициентов, полученную по полным данным. Поскольку ее посчитать оказывается возможным, то и тесты на значимость и линейные ограничения тоже можно провести так, как будто модель оценивалась на полных данных (даже при их фактическом отсутствии).

④

В задаче подразумевается тест гипотезы $H_0 : \beta_{\text{мал.}} = \beta_{\text{сред.}}$ против составной альтернативы $H_1 : \beta_{\text{мал.}} \neq \beta_{\text{сред.}}$. Реализовать такой тест можно было по-разному.

Например, можно было рассмотреть k тестов типа Стьюдента. Поскольку выборки были независимыми, независимы и оценки $\hat{\beta}_{\text{мал.}}$ и $\hat{\beta}_{\text{сред.}}$, то есть в таком случае

$$\operatorname{var}\left(\hat{\beta}^{(j)}_{\text{мал.}} - \hat{\beta}^{(j)}_{\text{сред.}}\right) = \operatorname{var}\hat{\beta}^{(j)}_{\text{мал.}} + \operatorname{var}\hat{\beta}^{(j)}_{\text{сред.}} - 2\underbrace{\operatorname{cov}\left(\hat{\beta}^{(j)}_{\text{мал.}}, \hat{\beta}^{(j)}_{\text{сред.}}\right)}_{=0} \qquad \text{(1 балл)}$$

А тогда можно составить k отдельных статистик типа Стьюдента и сделать на их основе k отдельных теста по одному на каждую пару коэффициентов:

$$t^{\text{расч.}}_j = \frac{\hat{\beta}^{(j)}_{\text{мал.}} - \hat{\beta}^{(j)}_{\text{сред.}}}{\widehat{\operatorname{se}}\left(\hat{\beta}^{(j)}_{\text{мал.}} - \hat{\beta}^{(j)}_{\text{сред.}}\right)} = \frac{\hat{\beta}^{(j)}_{\text{мал.}} - \hat{\beta}^{(j)}_{\text{сред.}}}{\sqrt{\widehat{\operatorname{var}}\,\hat{\beta}^{(j)}_{\text{мал.}} + \widehat{\operatorname{var}}\,\hat{\beta}^{(j)}_{\text{сред.}}}} \overset{\text{As.}}{\sim} \mathcal{N}(0,1) \qquad \text{(2 балла)}$$

Один из существенных недочетов такого подхода состоит в том, что k тестов для каждого коэффициента по отдельности неэквивалентны тесту на отличие между векторами в целом из-за так называемой проблемы множественного тестирования. Чтобы устранить эту проблему, участники могли предложить использовать коррекцию Бонферрони, процедуру Хольма или какие-нибудь еще методы корректного множественного тестирования (2 балла).

Альтернативно можно было предположить единый векторный вариант этого теста. Ковариационная матрица разности оценок из-за независимости подвыборок имеет диагональный вид с суммами дисперсий на диагонали. А тогда можно сделать тест со статистикой типа Вальда:

$$F^{\text{расч.}} = \left(\hat{\beta}_{\text{мал.}} - \hat{\beta}_{\text{сред.}}\right)^{\top} \left(\mathbb{V}\left(\hat{\beta}_{\text{мал.}} - \hat{\beta}_{\text{сред.}}\right)\right)^{-1} \left(\hat{\beta}_{\text{мал.}} - \hat{\beta}_{\text{сред.}}\right) \overset{\text{As.}}{\sim} \chi^2_k \qquad \text{(5 баллов)}$$

Этот путь лишен недостатка множественного тестирования, причем можно отметить, что его тестовая F-статистика совпадает с суммой квадратов t-статистик из предыдущего рассуждения.

Отдельно надо упомянуть, что оба этих подхода доступны Ивану Сергеевичу вне зависимости от того, доступны ему оценки ковариационных матриц оценок коэффициентов моделей или же, как предложено в этом пункте, только t-статистики (поскольку стандартные ошибки можно выразить через коэффициенты и их t-статистики).

Очень похожий результат можно было бы получить с помощью теста Чоу, который проверяет в точности требуемую гипотезу. Однако в данном конкретном случае этот тест неприменим, поскольку в новых условиях у Ивана Сергеевича больше нет полных ковариационных матриц оценок коэффициентов для двух моделей. Соответственно, он не может восстановить ESS объединенной модели, который требуется для теста Чоу. Несмотря на это, участникам, предложившим этот тест, все равно выставлялся (неполный) балл за этот пункт, поскольку тест в самом деле имеет непосредственное отношение к вопросу задачи.

(5)

Пусть Иван Сергеевич в новых условиях отсутствия ковариационных матриц собирается искать взвешенную «интегральную» в виде $\alpha_1\hat\beta_1 + \alpha_2\hat\beta_2$. В качестве ответа участников принимались любые разумные рассуждения, обосновывающие выбор каких-нибудь конкретных весов (2 балла) с дальнейшим исследованием статистических свойств предлагаемого оценщика (2 балла). Далее приводится авторский пример такого рассуждения.

Можно исследовать взвешенную оценку на несмещенность в общем виде[2]:

$$\mathbb{E}\left(\alpha_1\hat\beta_1 + \alpha_2\hat\beta_2\right) = \alpha_1\beta + \alpha_2\beta = (\alpha_1+\alpha_2)\beta$$

Стало быть, оценка будет несмещенной, *если и только если* $\alpha_1 + \alpha_2 = 1$. Тогда можно определить взвешенный оценщик как

$$\hat\beta_\alpha = \alpha\hat\beta_1 + (1-\alpha)\hat\beta_2$$

Получается, при любом фиксированном α такая оценка является несмещенной. Теперь Иван Сергеевич может подумать о том, какое именно α лучше всего выбрать. Например, он может ориентироваться на среднеквадратичный риск – MSE (*mean squared error*). Теоретический MSE этой оценки равен сумме дисперсий оценок коэффициентов (в общей ситуации для векторной оценки MSE равен скалярному квадрату вектора смещений плюс след ковариационной матрицы):

$$\begin{aligned}\mathrm{MSE}\big(\hat\beta_\alpha\big) = \operatorname{tr}\mathbb{V}\hat\beta_\alpha = \operatorname{tr}\Big(\alpha^2\,\mathbb{V}\hat\beta_1 + 2\alpha(1-\alpha)\overbrace{\operatorname{cov}\big(\hat\beta_1,\hat\beta_2\big)}^{=0} + (1-\alpha)^2\,\mathbb{V}\hat\beta_2\Big) = \\ = \operatorname{tr}\left(\alpha^2\,\mathbb{V}\hat\beta_1 + (1-\alpha)^2\,\mathbb{V}\hat\beta_2\right) = \alpha^2\operatorname{tr}\mathbb{V}\hat\beta_1 + (1-\alpha)^2\operatorname{tr}\mathbb{V}\hat\beta_2\end{aligned}$$

Можно отыскать $\alpha^* = \underset{\alpha}{\operatorname{argmin}}\ \mathrm{MSE}\big(\hat\beta_\alpha\big)$ – оптимальное значение веса в $\hat\beta_\alpha$:

$$\frac{\partial}{\partial\alpha}\mathrm{MSE}\big(\hat\beta_\alpha\big) = 0 \iff \alpha\operatorname{tr}\mathbb{V}\hat\beta_1 - (1-\alpha)\operatorname{tr}\mathbb{V}\hat\beta_2 = 0 \iff \alpha^* = \frac{\operatorname{tr}\mathbb{V}\hat\beta_2}{\operatorname{tr}\mathbb{V}\hat\beta_1 + \operatorname{tr}\mathbb{V}\hat\beta_2}$$

Может показаться, что непосредственное вычисление этой α^* Ивану Сергеевичу недоступно, потому что для нахождения $\operatorname{tr}\mathbb{V}\hat\beta_{1,2} = \sigma^2\left(X_{1,2}^\top X_{1,2}\right)^{-1}$ нужно знать матрицы $X_{1,2}^\top X_{1,2}$ (которых в новых условиях у Ивана Сергеевича больше нет) и истинную σ^2 – дисперсию шума в модели.

[2] Здесь и далее предполагается, что две подвыборки однородны, то есть в самом деле являются выборками из одного и того же распределения, а тогда теорема Гаусса–Маркова гарантирует несмещенность обеих оценок.

Однако, поскольку σ^2 присутствует и в числителе, и в знаменателе, можно вычислить α^* так:

$$\alpha^* = \frac{\operatorname{tr}\left[\cancel{\sigma^2}(X_2^\top X_2)^{-1}\right]}{\operatorname{tr}\left[\cancel{\sigma^2}(X_1^\top X_1)^{-1}\right] + \operatorname{tr}\left[\cancel{\sigma^2}(X_2^\top X_2)^{-1}\right]} = \frac{\operatorname{tr}\left[\frac{n_2-k}{\mathrm{ESS}_2}\hat{\mathbb{V}}\hat{\beta}_2\right]}{\operatorname{tr}\left[\frac{n_1-k}{\mathrm{ESS}_1}\hat{\mathbb{V}}\hat{\beta}_1\right] + \operatorname{tr}\left[\frac{n_2-k}{\mathrm{ESS}_2}\hat{\mathbb{V}}\hat{\beta}_2\right]} =$$
$$= \frac{\frac{n_2-k}{\mathrm{ESS}_2}\operatorname{tr}\left[\hat{\mathbb{V}}\hat{\beta}_2\right]}{\frac{n_1-k}{\mathrm{ESS}_1}\operatorname{tr}\left[\hat{\mathbb{V}}\hat{\beta}_1\right] + \frac{n_2-k}{\mathrm{ESS}_2}\operatorname{tr}\left[\hat{\mathbb{V}}\hat{\beta}_2\right]}$$

Если Ивану Сергеевичу дана только регрессионная табличка с t-статистиками, то вновь, поскольку $t_{\text{расч.}} = \frac{\hat{\beta}}{\hat{\mathrm{se}}\,\hat{\beta}}$, имеем $\widehat{\mathrm{var}}\,\hat{\beta} = \frac{\hat{\beta}^2}{t^2_{\text{расч.}}}$. Иными словами, из матрицы $\hat{\mathbb{V}}\hat{\beta}$ Ивану Сергеевичу известны *только элементы на главной диагонали*, но этого достаточно для того, чтобы посчитать следы этих матриц. Это означает, что оптимальное значение α^* доступно для вычисления и может быть использовано в том виде, к которому преобразовано выше.

Задача 2. (25 баллов)

①

Рассмотрим совместное распределение (Y, D) и формально определим невязку $\varepsilon \stackrel{\text{def}}{=\!=} Y - \mathbb{E}[Y \mid D]$, тогда

$$Y = \mathbb{E}[Y \mid D] + \varepsilon$$

Поскольку $\mathbb{E}[Y \mid D]$ – функция от D, то $\mathbb{E}\big[\mathbb{E}[Y \mid D] \;\big|\; D\big] = \mathbb{E}[Y \mid D]$, откуда получаем известное свойство условного математического ожидания

$$\mathbb{E}[\varepsilon \mid D] = \mathbb{E}[Y - \mathbb{E}[Y \mid D] \mid D] = 0$$

Теперь обратим внимание на то, что $\mathbb{E}[Y \mid D]$ можно расписать по случаям $D \in \{0, 1\}$ следующим образом

$$\mathbb{E}[Y \mid D] = \begin{cases} \mathbb{E}[Y \mid D = 0], & D = 0 \\ \mathbb{E}[Y \mid D = 1], & D = 1 \end{cases} = \mathbb{E}[Y \mid D = 0](1 - D) + \mathbb{E}[Y \mid D = 1]D =$$

$$= \mathbb{E}[Y \mid D = 0] + (\mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0])D$$

В таком случае, получается, что Y представляется в виде

$$Y = \mathbb{E}[Y \mid D = 0] + \big(\mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0]\big)D + \varepsilon$$

Непосредственно сравнивая полученное представление с уравнением линейной регрессии $Y = \beta_0 + \beta_1 D + \varepsilon$ из условия, получаем, что

$$\beta_0 = \mathbb{E}[Y \mid D = 0] \qquad \text{(2 балла)}$$
$$\beta_1 = \mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0] \qquad \text{(2 балла)}$$

②

Для краткости обозначим $p = \mathbb{P}[D = 1]$. Теперь можно расписать τ по закону повторного математического ожидания и преобразовать результат как

$$\begin{aligned} \tau &= \mathbb{E}[Y(1) - Y(0)] = \\ &= p\,\mathbb{E}[Y(1) - Y(0) \mid D = 1] + (1 - p)\,\mathbb{E}[Y(1) - Y(0) \mid D = 0] = \\ &= p\,\mathbb{E}[Y(1) \mid D = 1] + (1 - p)\,\mathbb{E}[Y(1) \mid D = 0] - \\ &\quad - p\,\mathbb{E}[Y(0) \mid D = 1] - (1 - p)\,\mathbb{E}[Y(0) \mid D = 0] \end{aligned} \tag{1}$$

С другой стороны, так как $Y(1) = Y$ при $D = 1$, и $Y(0) = Y$ при $D = 0$, значит,

$$\begin{aligned}\beta_1 &= \mathbb{E}[Y \mid D = 1] - \mathbb{E}[Y \mid D = 0] = \\ &= \mathbb{E}[Y(1) \mid D = 1] - \mathbb{E}[Y(0) \mid D = 0]\end{aligned}$$

Из (1) и последнего равенства для β_1 следует, что

$$\begin{aligned}\beta_1 - \tau = (1-p)\big(\mathbb{E}[Y(1) \mid D = 1] - \mathbb{E}[Y(1) \mid D = 0]\big) + \\ + p\big(\mathbb{E}[Y(0) \mid D = 1] - \mathbb{E}[Y(0) \mid D = 0]\big)\end{aligned} \tag{2}$$

В общем случае правая часть (2) не равна нулю (3 балла), этот эффект иногда называется смещением из-за самоотбора (*selection bias*).

Что касается сравнения β_1 и τ, то эмпирически правдоподобно, что работники с более высокими способностями чаще получают высшее образование и что способности положительно влияют на потенциальные заработные платы как при $D = 1$, так и при $D = 0$. Тогда обычно ожидается

$$\mathbb{E}[Y(1) \mid D = 1] \geq \mathbb{E}[Y(1) \mid D = 0] \quad \text{и} \quad \mathbb{E}[Y(0) \mid D = 1] \geq \mathbb{E}[Y(0) \mid D = 0]$$

Поэтому в такой ситуации из (2) следует $\beta_1 \geq \tau$, то есть β_1 завышена (2 балла).

③

Предварительно удачно было бы заметить, что при фиксированном $p = \mathbb{P}[D = 1]$ имеют место равенства

$$E[D] = p \tag{3}$$

$$\begin{aligned}\mathbb{E}[DY] &= \mathbb{E}[Y \mid D = 1]\,\mathbb{P}[D = 1] + 0 \cdot \mathbb{P}[D = 0] = \\ &= \mathbb{E}[Y(1) \mid D = 1]\,\mathbb{P}[D = 1] = p\,\mathbb{E}[Y(1) \mid D = 1]\end{aligned} \tag{4}$$

$$\begin{aligned}\mathbb{E}[(1-D)Y] &= 0 \cdot \mathbb{P}[D = 1] + \mathbb{E}[Y \mid D = 0]\,\mathbb{P}[D = 0] = \\ &= \mathbb{E}[Y(0) \mid D = 0]\,\mathbb{P}[D = 0] = (1-p)\,\mathbb{E}[Y(0) \mid D = 0]\end{aligned} \tag{5}$$

Далее, каким бы ни было распределение (Y, D), в любом случае при $Y \in [0, 1]$ имеют место тривиальные неравенства

$$0 \leq \mathbb{E}[Y(0) \mid D = 1] \leq 1$$

$$0 \leq \mathbb{E}[Y(1) \mid D = 0] \leq 1$$

Используя красные неравенства, представление (1) и (3–5), получаем нижнюю границу для τ вида

$$\begin{aligned}\tau &= p\,\mathbb{E}[Y(1)\mid D=1] + (1-p)\,\underbrace{\mathbb{E}[Y(1)\mid D=0]}_{\geq 0} - \\ &- p\,\underbrace{\mathbb{E}[Y(0)\mid D=1]}_{\leq 1} - (1-p)\,\mathbb{E}[Y(0)\mid D=0] \geq \\ &\geq \underbrace{p\,\mathbb{E}[Y(1)\mid D=1]}_{\mathbb{E}[DY]} - \underbrace{(1-p)\,\mathbb{E}[Y(0)\mid D=0]}_{\mathbb{E}[(1-D)Y]} - \underbrace{p}_{\mathbb{E}[D]} = \\ &= \mathbb{E}[DY - (1-D)Y - D] = \mathbb{E}[(2D-1)Y - D] = \underline{\beta} \qquad \text{(2 балла)}\end{aligned}$$

Аналогично, используя синие неравенства, представление (1) и (3–5), получаем и верхнюю границу для τ как

$$\begin{aligned}\tau &= p\,\mathbb{E}[Y(1)\mid D=1] + (1-p)\,\underbrace{\mathbb{E}[Y(1)\mid D=0]}_{\leq 1} - \\ &- p\,\underbrace{\mathbb{E}[Y(0)\mid D=1]}_{\geq 0} - (1-p)\,\mathbb{E}[Y(0)\mid D=0] \leq \\ &\leq \underbrace{p\,\mathbb{E}[Y(1)\mid D=1]}_{\mathbb{E}[DY]} - \underbrace{(1-p)\,\mathbb{E}[Y(0)\mid D=0]}_{\mathbb{E}[(1-D)Y]} + \underbrace{(1-p)}_{\mathbb{E}[1-D]} = \\ &= \mathbb{E}[DY - (1-D)Y + 1 - D] = \mathbb{E}[(2D-1)Y + (1-D)] = \overline{\beta} \qquad \text{(2 балла)}\end{aligned}$$

Иными словами, действительно имеет место неравенство $\underline{\beta} \leq \tau \leq \overline{\beta}$, которое и требовалось получить (1 балл).

(4)

Поскольку $\underline{\beta} = \mathbb{E}[(2D-1)Y - D]$, то наиболее естественный способ построить для нее доверительный интервал – сконструировать какую-нибудь статистику, хорошо оценивающую это математическое ожидание. Тогда довольно естественно обозначить $W_i \overset{\text{def}}{=\!=} (2D_i - 1)Y_i - D_i$ для $i = 1, \ldots, n$ и ввести статистики

$$\overline{W}_n = \frac{1}{n}\sum_{i=1}^{n} W_i \quad \text{и} \quad s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\big(W_i - \overline{W}_n\big)^2$$

В таком случае видим, что $\mathbb{E}\big[\overline{W}_n\big] = \underline{\beta}$, то есть $\overline{W}_n$ несмещенно оценивает $\underline{\beta}$ и является средним одинаково распределенных случайных величин с конечной дисперсией. Поскольку из соображения закона больших чисел $s_n^2 \xrightarrow{p} \mathbb{V}[W_i]$, теорема

Слуцкого и центральная предельная теорема гарантируют, что

$$\frac{\sqrt{n}\left(\overline{W}_n - \underline{\beta}\right)}{s_n} \xrightarrow{d} \mathcal{N}(0,1)$$

На основе этой асимптотики можно построить требуемый 95%-й односторонний асимптотический доверительный интервал:

$$\mathbb{P}\left[\underline{\beta} \geq \overline{W}_n - \underbrace{1.645}_{z_{\text{крит.}}} \frac{s_n}{\sqrt{n}}\right] = \mathbb{P}\left[\frac{\sqrt{n}\left(\overline{W}_n - \underline{\beta}\right)}{s_n} \leq 1.645\right] \xrightarrow{n\to\infty} 0.95$$

Поэтому в качестве нижней границы можно взять $c_n = \overline{W}_n - 1.645\,\frac{s_n}{\sqrt{n}}$, и тогда $I_n = [c_n,\, +\infty)$ является искомым доверительным множеством (5 баллов).

⑤

Из дополнительного условия в этому пункту и того, что $Y \in [0,\,1]$, следуют неравенства

$$\mathbb{E}[Y(0) \mid D=0] \leq \mathbb{E}[Y(0) \mid D=1] \leq 1 \tag{6}$$

$$0 \leq \mathbb{E}[Y(1) \mid D=0] \leq \mathbb{E}[Y(1) \mid D=1] \tag{7}$$

Повторяя шаги из пункта ③ для синих неравенств, приведенных выше, получаем уточненную верхнюю границу для τ:

$$\begin{aligned}
\tau = {} & p\,\mathbb{E}[Y(1) \mid D=1] + (1-p)\,\underbrace{\mathbb{E}[Y(1) \mid D=0]}_{\leq \mathbb{E}[Y(1)\mid D=1]} - \\
& - p\,\underbrace{\mathbb{E}[Y(0) \mid D=1]}_{\geq \mathbb{E}[Y(0)\mid D=0]} - (1-p)\,\mathbb{E}[Y(0) \mid D=0] \leq \\
\leq {} & \underbrace{p\,\mathbb{E}[Y(1) \mid D=1] + (1-p)\,\mathbb{E}[Y(1) \mid D=1]}_{= \mathbb{E}[Y(1)\mid D=1] = \mathbb{E}[Y \mid D=1]} - \\
& - \underbrace{\left(p\,\mathbb{E}[Y(0) \mid D=0] + (1-p)\,\mathbb{E}[Y(0) \mid D=0]\right)}_{= \mathbb{E}[Y(0)\mid D=0] = \mathbb{E}[Y \mid D=0]} = \mathbb{E}[Y \mid D=1] - \mathbb{E}[Y \mid D=0] = \beta_1
\end{aligned}$$

Использование другой пары неравенств приводит к уже известному результату из пункта ③, поэтому в итоге имеем уточненное неравенство $\underline{\beta} \leq \tau \leq \beta_1$ (3 балла).

⑥

Если $\hat{\beta}_1$ – МНК-оценка коэффициента β_1 в модели $Y = \beta_0 + \beta_1 D + \varepsilon$, то из соображений, аналогичных шагам в пункте ④, имеем

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_n/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

Поэтому в качестве верхней односторонней доверительной границы можно взять $c_n^* = \hat{\beta}_1 + 1.645\,\frac{\hat{\sigma}_n}{\sqrt{n}}$, так что множество $I_n^* = (-\infty, c_n^*]$ удовлетворяет требованию на нужную (асимптотическую) доверительную (3 балла).

Задача 3. (25 баллов)

(1)

Имея стандартную ошибку и t-статистику для переменной EXP, можно легко восстановить сам коэффициент, то есть пропуск A:

$$t_{\text{расч.}} = \frac{\hat{\beta}_1}{\widehat{\text{se}}\,\hat{\beta}_1} \implies \hat{\beta}_1 = t_{\text{расч.}} \cdot \widehat{\text{se}}\,\hat{\beta}_1 = 3.637 \cdot 0.0081 = 0.0295$$

$$\implies \boxed{\text{A} = 0.0295} \qquad \text{(2 балла)}$$

Пропуск C можно заполнить, если заметить, что p-значение теста на значимость равно 0.01, то есть t-статистика *по модулю* совпадает с табличным значением 1%-го уровня. Более того, коэффициент регрессии имеет такой же знак, как t-статистика, и для переменной EXP*FEMALE он отрицателен. С учетом этого, а также того, что число наблюдений $n = 540$ достаточно велико, можно воспользоваться критическим значением нормального распределения:

$$t^{1\%}_{\text{крит.}} \approx z^{1\%}_{\text{крит.}} = -2.58 \implies \boxed{\text{C} = -2.58} \qquad \text{(2 балла)}$$

Имея C – t-статистику для теста на значимость коэффициента при переменной EXP*FEMALE, пропуск B восстанавливается тем же путем, что и A:

$$t_{\text{расч.}} = \frac{\hat{\beta}_4}{\widehat{\text{se}}\,\hat{\beta}_4} \implies \widehat{\text{se}}\,\hat{\beta}_4 = \frac{\hat{\beta}_4}{t_{\text{расч.}}} = \frac{-0.0258}{-2.58} = 0.01$$

$$\implies \boxed{\text{B} = 0.01} \qquad \text{(2 балла)}$$

Пропуск D – коэффициент R^2, который можно посчитать в терминах ESS (суммы квадратов остатков) и TSS (общей суммы квадратов):

$$\begin{aligned} \text{ESS} &= 126.24 \quad \text{(дано в таблице)} \\ \text{TSS} &= n\,\widehat{\text{var}}\,y = 540 \cdot 0.588^2 = 186.70 \end{aligned} \implies R^2 = 1 - \frac{126.24}{186.70} = 0.324$$

$$\implies \boxed{\text{D} = 0.324} \qquad \text{(2 балла)}$$

Наконец, E – исправленный R^2, который вычисляется, например, по формуле

$$R^2_{\text{adj}} = R^2 - \frac{k-1}{n-k}\left(1-R^2\right) = 0.324 - \frac{5-1}{540-5}\cdot(1-0.324) = 0.319$$

$$\Longrightarrow \boxed{\text{E} = 0.319} \qquad \text{(2 балла)}$$

(2)

Поскольку зависимая переменная в регрессии – логарифм почасовой заработной платы, коэффициент отражает процентное изменение заработной платы. Иными словами, если при прочих равных *у мужчины* опыт работы больше на 1 год, то его почасовая зарплата выше приблизительно на $0.0295 \cdot 100\% \approx 2.95\%$ (2 балла).

(3)

Как известно, нулевая гипотеза теста Уайта предполагает отсутствие гетероскедастичности; соответственно, тест уверенно показал *наличие* гетероскедастичности какого-то характера в модели. Тест Голдфелда–Квандта, в свою очередь, выявляет гетероскедастичность, монотонно ассоциированную с той переменной, по которой производилось упорядочивание, – то есть с годами обучения S.

Понятно, что выявленная гетероскедастичность может быть связана с S заметно немонотонным образом или же вообще может не быть связана с S (в модель включено множество переменных, кроме S). Это означает, что ни один из возможных выводов теста Голдфелда–Квандта не будет противоречить выводу теста Уйта. Соответственно, тестовая статистика может принимать любое положительное значение, то есть быть как меньше критического $F_{\text{крит.}}$, так и больше (4 балла).

В случае, если участники предлагали, что тестовая статистика теста Голдфелда–Квандта должна быть больше табличной, поскольку тест должен выявить гетероскедастичность и согласиться с тестом Уайта, выставлялся 1 балл.

(4)

В такой ситуации применяется F-тест на сравнение вложенных моделей («короткая» против «длинной», или «*restricted vs. unrestricted*»). Здесь «длинная» и «короткая» («*unrestricted*» и «*restricted*» соответственно) модели отличаются включением

переменных FEMALE и EXP*FEMALE, потому что половая дискриминация в точности означает, что либо пол имеет непосредственное влияние на доход, либо год опыта работы для мужчин и женщин ценится по-разному.

Модели, которые требуются для оценки, имеют вид:

$$\text{Модедь 1 (UR)}: \quad \ln \text{EARNINGS} = \beta_0 + \beta_1 \text{EXP} + \beta_2 \text{S} + \beta_3 \text{FEMALE} + \beta_4 \, \text{EXP*FEMALE} + \varepsilon$$

$$\text{Модедь 2 (R)}: \quad \ln \text{EARNINGS} = \alpha_0 + \alpha_1 \text{EXP} + \alpha_2 \text{S} + \varepsilon$$

Формально описанный тест в данном конкретном случае можно формализовать следующим образом:

$$H_0 \ : \ \beta_3 = \beta_4 = 0$$
$$H_1 \ : \ \beta_3 \neq 0 \ \text{ или } \ \beta_4 \neq 0 \qquad F_{\text{расч.}} = \frac{\left(R^2_{\text{UR}} - R^2_{\text{R}}\right)/q}{\left(1 - R^2_{\text{UR}}\right)/(n-k)} \sim F_{q,\, n-k}$$

Выше n и k – числа наблюдений и переменных (в «длинной» модели) соответственно, а q – число коэффициентов, обнуление которых тестируется. В данном конкретном примере тогда

$$F_{\text{расч.}} = \frac{(0.324 - 0.25)/2}{(1 - 0.324)/(540 - 5)} \approx 29.28 > 3.01$$

Получается, нулевая гипотеза отвергается в пользу альтернативной: модель 1 («длинная») заметно лучше описывает данные. Иными словами, исследование Сергея показывает, что дискриминация есть (5 баллов).

(5)

VIF-коэффициенты рассчитываются на основе R^2-коэффициентов вспомогательных регрессий одной объясняющей переменной на другие. Поскольку в модели 2 всего две объясняющие переменные — S и EXP, то для вычисления VIF-коэффициентов требуется оценка R^2 двух регрессий:

$$\text{EXP} = \gamma_0 + \gamma_1 \text{S} + \varepsilon \qquad (1)$$

$$\text{S} = \delta_0 + \delta_1 \text{EXP} + \varepsilon \qquad (2)$$

В данном случае VIF-коэффициенты должны совпадать, потому что они измеряют степень мультиколлинеарности между одной и той же парой переменных, поскольку обе регрессии – парные (2 балла).

Однако это можно показать и чисто формально. Как известно, для парной линейной регрессии коэффициент детерминации R^2 равен квадрату коэффициента корреляции между объясняющей и объясняемой переменными. Это означает, что в регрессиях (1)–(2) коэффициенты R^2 совпадают (так как оба равны $\widehat{\text{corr}}^2(\text{EXP}, \text{S})$). Но тогда и VIF-коэффициенты переменных должны совпадать, то есть

$$\text{VIF}_\text{S} = \text{VIF}_\text{EXP} = 1.050 \qquad \text{(2 балла)}$$

Содержательный вывод, который можно из этого сделать, таков: оба VIF-коэффициента существенно меньше 10, то есть в модели не наблюдается сколько-нибудь существенной мультиколлинеарности.

Задача 4. (25 баллов)

①

5 баллов за полное обоснование. Вообще говоря, предложенное в этом пункте утверждение – известный факт из теории вероятностей. Для его обоснования предположим дополнительно, что F_Y и F_X строго возрастают, и отыщем функцию распределения случайной величины $\eta = \frac{P_i}{100}$ (для $\frac{R_i}{100}$ рассуждение полностью аналогичное):

$$F_\eta(y) = \mathbb{P}\left(\frac{P_i}{100} \le y\right) = \mathbb{P}\big(F_Y(Y_i) \le y\big) = \begin{cases} 1, & y \ge 1 \\ \mathbb{P}\big(Y_i \le F_Y^{-1}(y)\big), & y \in (0,\,1) \\ 0, & y \le 0 \end{cases} =$$

$$= \begin{cases} 1, & y \ge 1 \\ F_Y\left(F_Y^{-1}(y)\right), & y \in (0,\,1) \\ 0, & y \le 0 \end{cases} = \begin{cases} 1, & y \ge 1 \\ y, & y \in (0,\,1) \\ 0, & y \le 0 \end{cases}$$

Это в точности функция распределения случайной величины, равномерно распределенной на $[0,\,1]$, иными словами, $\frac{P_i}{100} \sim \text{Uniform}[0,\,1]$ и абсолютно аналогично $\frac{R_i}{100} \sim \text{Uniform}[0,\,1]$.

②

5 баллов за полное обоснование, включая ссылку на равенство дисперсий. Рассмотрим теоретическую модель линейной регрессии $R = \alpha + \beta P + \varepsilon$ и отыщем коэффициент β:

$$\operatorname{cov}(P_i, R_i) = \operatorname{cov}(P_i, \alpha + \beta P_i + \varepsilon_i) = \beta \underbrace{\operatorname{cov}(P_i, P_i)}_{\mathbb{V}P_i} + \underbrace{\operatorname{cov}(P_i, \varepsilon_i)}_{0} = \beta\,\mathbb{V}P_i$$

$$\Longrightarrow\ \beta = \frac{\operatorname{cov}(P_i, R_i)}{\mathbb{V}P_i}$$

Осталось заметить, что, поскольку P_i и R_i одинаково распределены, их дисперсии совпадают и равны[3] $\mathbb{V}P_i = \mathbb{V}R_i = \frac{100^2}{12}$. Тогда можно преобразовать коэффициент корреляции и получить в точности требуемое:

$$\rho = \operatorname{corr}(P_i, R_i) = \frac{\operatorname{cov}(P_i, R_i)}{\sqrt{\mathbb{V}P_i\,\mathbb{V}R_i}} = \frac{\operatorname{cov}(P_i, R_i)}{100^2/12} = \frac{\operatorname{cov}(P_i, R_i)}{\mathbb{V}P_i} = \beta$$

[3] Поскольку в первом пункте показано, что $\eta = \frac{P_i}{100} \sim \text{Uniform}[0,\,1]$, то $\mathbb{V}P_i = \mathbb{V}(100\eta) = 100^2 \cdot \mathbb{V}\eta = \frac{100^2}{12}$, и для R_i верно то же самое.

(3)

5 баллов за полное обоснование, из них 3 за докзательство и 2 за ответ с обоснованием про ребёнка из семьи с медианным доходом. Если в этом пункте только верная формула для свободного члена (альфа), то 1 балл. Поскольку R_i и P_i распределены как $100 \cdot \text{Uniform}[0,\, 1]$, их математические ожидания существуют и равны $100 \cdot \frac{1}{2} = 50$.

Тогда для определения α вновь рассмотрим уравнение линейной регрессии $R_i = \alpha + \beta P_i + \varepsilon$ и вычислим математическое ожидание левой и правой части:

$$\underbrace{\mathbb{E}\, R_i}_{=50} = \mathbb{E}(\alpha + \beta P_i + \varepsilon_i) = \alpha + \beta \underbrace{\mathbb{E}\, P_i}_{=50} + \underbrace{\mathbb{E}\, \varepsilon_i}_{=0} = \alpha + 50\beta$$

$$50 = \alpha + 50\beta \implies \alpha = 50(1 - \beta)$$

Что касается ожидаемого ранга ребенка, родившегося в семье с медианным доходом, для него $P = 100 \cdot \frac{1}{2} = 50$ и $\mathbb{E}\,(\varepsilon \mid P = 50) = 0$, поэтому из равенства выше имеем

$$\mathbb{E}(R \mid P = 50) = \alpha + 50\beta = 50$$

Интересно, что этот результат не зависит от значений α и β: в среднем, у родителей, зарабатывающих медианный доход (доход ранга 50), рождаются дети, которые будут иметь такой же медианный ранг дохода в своей когорте.

(4)

Задание предполагает много разных способов ответа. 10 баллов распределяются следующим образом:

- 4 балла за выписанную модель, по которой можно проверить гипотезу из условия. В задании сказано о превышении ранга дохода ребёнка над рангом родителей, поэтому зависимой переменной может быть, например, разность между рангами или их отношение. Может быть модель бинарного выбора, где событие - это превышение ранга ребенка над рангом родителей. Просто показатель дохода или ранг дохода ребёнка в качестве зависимой переменной не подходит (-1 балл). В правой части уравнения должны присутствовать, как минимум, образование родителей и госрасходы (здесь можно было написать на вкус участника, например, получение гранта/субсидии), и их произведение. Чтобы первая часть гипотезы из условия выполнялась, необходимо чтобы коэффициент при произведении был значимым и отрицательным. Для проверки второй части гипотезы можно ввести, например (возможны и другие варианты) показатель доходов ниже медианы (или другого уровня) и ввести в модель произведение образования,

госрасходов и бинарной переменной низких доходов. Если по выписанной модели нельзя проверить гипотезу из условия, баллы снимаются.

- 2 балла за обсуждение проблемы эндогенности: она может возникать, например, из-за пропуска ненаблюдаемых существенных характеристик (талант, способности и т.д.), а также из-за ошибок измерения дохода.

- 2 балла за описание стратегии оценки: 1 балл за необходимые данные (это панель или пространственные данные, перечень статистических показателей) и 1 балл за метод оценки. Раз участник пишет об эндогенности, то необходимо написать, как попытаться решить эту проблему.

- 2 балла за уместные тесты и критерии качества модели с хотя бы минимальными объяснениями, почему именно ни нужны в данной ситуации.