Семинарский лист 7

Задание 1. Найдите ошибки и ограничения в применении методов и интерпретации в следующих примерах:

1. Исследователь оценил следующую регрессионную модель:

$$\widehat{promoted}_i = \hat{\beta}_1 gender_i + \hat{\beta}_2 age_i^2 + \hat{\beta}_4 performance_i$$

Зависимая переменная является бинарной: принимает значение 1, если сотрудник получил повышение в последние 2 года, 0 – в противном случае.
Переменная $gender$ – дамми-переменная, принимающая значение 1 для сотрудников-мужчин, 0 – соответственно, для сотрудниц-женщин;
Переменная age – возраст (в годах);
Переменная $performance_i$ – показатель производительности сотрудника, измеренный в интервальной шкале от 0 до 100

2. В 2023 г. регион А внедрил программу цифровизации, предоставив всем старшеклассникам планшеты с интерактивными учебниками. Соседний регион В подобной программы не внедрял. Были собраны панельные данные о средних оценках по математике для случайной выборки учащихся 10-х классов из обоих регионов за два учебных года —- до внедрения программы цифровизации (то есть, за 2022 г.) и после (то есть, за 2024 г.).

   Для оценки эффекта программы была предложена следующая модель линейной регрессии с использованием объединённых данных за два года (2022 и 2024 гг.):

$$\widehat{Math\_Score}_i = \hat{\beta}_0 + \hat{\beta}_1 Digital\_Program_i$$

   ,

   где:

   $Math\_Score_i$ — средняя оценка по математике i-го ученика;

   $Digital\_Program_i$ – дамми-переменная, принимающая значение 1, если ученик i является участником программы (то есть учится в регионе А в 2024 году), и 0 в противном случае (все остальные случаи: регион А в 2022, регион В в 2022 и 2024)

   Оценка методом наименьших квадратов дала результат: $\hat{\beta}_1 = 4.73$ (p-value = 0.02). Положительный и статистически значимый коэффициент при $Digital\_Program$ свидетельствует о том, что программа цифровизации повысила успеваемость по математике в среднем на 4.7 балла. Данная оценка продемонстрировала устойчивость к добавлению в модель контрольной переменной на пол ученика

3. Несколько предикторов – контрольных переменных – оказались в оцененной регрессионной модели незначимыми, поэтому для более экономной спецификации модели исследователи исключили указанные объясняющие переменные из уравнения регрессии.

   Для диагностики линейной регрессионной модели на мультиколлинеарность исследователи построили корреляционную матрицу для используемых в модели предикторов. Ни один из парных коэффициентов корреляции не превышал 0.9, исходя из чего исследователи сделали вывод, что проблемы сильной мультиколлинеарности не наблюдается, а значит, и оценки коэффициентов останутся несмещенными.

   Кроме этого, был проведен тест Голдфелда-Квандта для каждой переменной по отдельности. Так как p-value в каждом из данных тестов превысил 0.05, исследователи сделали вывод о том, что гетероскедастичности в данном случае не наблюдается.

Задание 2. Порассуждайте, с какими источниками эндогенности могут столкнуться исследователи (в одном примере может быть сразу несколько таких источников) в следующих случаях:

История 1. Было замечено, что люди, не снимающие обувь перед сном, испытывают утром головную боль. На основе этого делается вывод о том, что засыпать, не снимая обуви, вредно, так как приводит к головной боли.

История 2. Изучается взаимосвязь заработной платы, с одной стороны, и образования (количества лет образования), вида занятости, с другой стороны. Заработная плата выступает зависимой переменной в регрессионной модели.

История 3. Есть наблюдение, что в некотором городе А полиция чаще задерживает для обыска пешеходов афро- и латиноамериканцев по сравнению с белыми. При этом с точки зрения применения полицейскими силы по отношению к задержанным значимых различий не удалось выявить. На основе этого делается вывод, что дискриминации в применении силы полицией по расовому признаку в городе А не существует.

История 4. Изучается влияние приема лекарства на состояние здоровья индивида. В регрессионной модели контролируются пол, возраст индивида и побочные эффекты от лекарства.

История 5. Исследователь изучает, как экономический кризис повлиял на финансовые показатели производительности компаний. Для этого он сравнивает до- и пост-кризисные значения показателей компаний. Для работы с пропущенными значениями используется процедура listwise deletion *(то есть, удаляется вся строка наблюдения, если в нём отсутствует хотя бы одно значение для любой переменной, участвующей в анализе)*