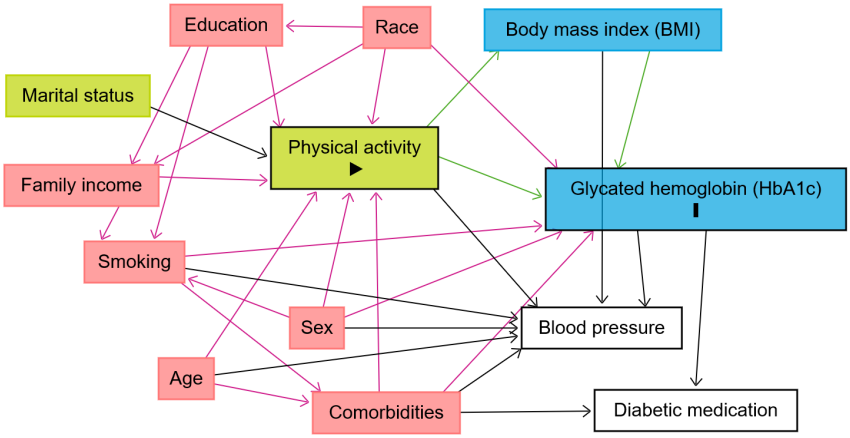
Домашнее задание по регрессионному анализу

Дедлайн - 22.12.2024

Вы изучаете **ассоциацию между физической активностью и уровнем гликированного гемоглобина (Hb1Ac)**. Согласно Википедии, гликированный гемоглобин отражает процент гемоглобина крови, необратимо соединенного с молекулами глюкозы, и является интегральным показателем гликемии за три месяца: чем выше уровень гликированного гемоглобина, тем выше была гликемия за последние три месяца и, соответственно, больше риск развития осложнений сахарного диабета.

Вы сделали обзор литературы по теме, пообщались с клиницистами -- и в результате у вас родился следующий **направленный ациклический граф (DAG)**[1]:



Для его отрисовки вы использовали приложение DAGitty, о котором узнали из лекции по причинно-следственному выводу на программе по биостатистике и анализу медицинских данных от Института биоинформатики. Вот этот граф онлайн, вы можете с ним работать (чтобы он выглядел совсем как на картинке выше, можно в Diagram style слева выбрать SEM-like).

Если вы подзабыли, что такое DAG'и, какие роли, с точки зрения причинно-следственного вывода, в нем играют разные показатели и как его использовать для отбора ковариат, на которые нужно adjust'ить эффект при интересующей вас переменной (exposure), и для определения тех, на которые adjust'ить не следует, вы можете вкратце почитать об этом здесь, здесь и здесь или посмотреть здесь и здесь. Конечно, DAGitty уже дает вам подсказки в виде вариантов тех наборов ковариат (в левом верхнем углу, в разделе *Causal effect identification*), коррекция на которые необходима, но для полного понимания того, почему он предлагает именно такие наборы, рекомендую хотя бы посмотреть указанные видео, а также попробовать отметить все переменные из соответствующего набора как adjusted (слева в разделе *Variable*), потом добавить к adjusted что-что, что в этот набор не вошло -- и объяснить для себя, почему в каких-то случаях коррекция на эти доп.ковариаты будет допустимой (справа сверху будет указание *Correctly adjusted*), а в каких-то -- нет (*Incorrectly adjusted*).

[1] На самом деле основа для этого графа взята из этой статьи -- заглянуть в нее не возбраняется и может быть даже чем-то полезным, но не является обязательным для выполнения задания

Ваша задача -- оценить общий (total) эффект физической активности в отношении гликированного гемоглобина.

Для этого вы провели **кросс-секционное исследование**, полученные данные (уже после подготовки и чистки, без пропусков) содержатся в файле ``HW_data.xlsx"[2], спецификация к ним (лейблы для переменных и категорий) -- в разбивке на разделы (на нескольких листах) -- в файле ``HW_codebook.xlsx". Обратите, пожалуйста, внимание на то, что все данные представлены в виде numeric переменных и не забудьте перевести нужные вам категориальные показатели в факторы.

Вы должны сами принять решение, каким образом вы будете измерять физическую активность по имеющимся данным (подойдет любой обоснованный способ). Аналогично -- с теми ковариатами, которые вы решите включить в свою модель. Это могут быть количественные или категориальные показатели, для категориальных показателей вы можете использовать как исходную (как в данных) категоризацию, так и объединять какие-то категории в одну (с соответствующим обоснованием) -- на ваше усмотрение.

Зависимую переменную (гликированный гемоглобин) категоризовать не нужно!

Задачи и вопросы:

1. Каким образом вы будете оценивать физическую активность респондентов? Есть ли у вас предварительные предположения относительно того, каким образом выбранный вами показатель может быть ассоциирован с гликированным гемоглобином?
2. Ковариаты для каких показателей вы включите в модель для коррекции эффекта физической активности в отношении гликированного гемоглобина? Каким образом вы будете их оценивать по имеющимся данным?

- *Бонусное задание*: для представленного DAG'a укажите роль каждого показателя по отношению к изучаемой ассоциации между физической активностью и гликированным гемоглобином (конфаундеры (в том числе proxy конфаундеры), коллайдеры, медиаторы)

3. Проведите необходимый эксплораторный анализ перед оценкой модели.
4. Оцените модель для зависимости гликированного гемоглобина от выбранного вами показателя физической активности без ковариат и с ними. Проведите необходимую диагностику этих моделей -- требует ли что-либо коррекции и почему? В случае необходимости коррекции по результатам диагностики сделайте ее.
5. Представьте результаты оценки модели без ковариат и с ковариатами в виде точечной и интервальной оценки эффекта физической активности. Дайте им словесную интерпретацию. Какие выводы мы можем сделать, исходя из точечной оценки? А из интервальной? Как вы думаете, можно ли считать эффект клинически значимым? Если затрудняетесь с ответом, что бы вам помогло дать ответ на этот вопрос?[3]
6. Проверьте гипотезу об отсутствии ассоциации между физической активностью и гликированным гемоглобином. Сделайте выводы по полученным результатам.
7. Является ли пол модификатором эффекта физической активности в отношении гликированного гемоглобина? Если да, каков эффект для мужчин и женщин и насколько он отличается между ними?

[2] В действительности эти данные -- кусочек NHANES (National Health and Nutrition Examination Survey) за 2013-2014 гг., но для выполнения задания это несущественная информация -- просто для понимания того, почему спецификация на английском :)

[3] Возможно, эта замечательная статья или эта не менее замечательная будут полезны при ответе на эти вопросы или в будущем для интерпретации результатов своих и чужих исследований

8. Соответствуют ли полученные вами результаты вашему исходному предположению? Как меняется оценка эффекта физической активности при добавлении ковариат в модель и почему?

- *Бонусное задание*: оцените прямой (direct) эффект физической активности на гликированный гемоглобин (со всей необходимой диагностикой и коррекциями). Как он отличается от общего (total) эффекта? В чем причина/ механизм этих различий?

Формат сдачи задания:

1) .rmd файл с пояснениями хода решения задач и ответами на поставленные вопросы, а также чанками с кодом для соответствующего анализа и вывода результатов
2) .html файл с отчетом, сформированным по .rmd из п.1. Отчет должен выглядеть так, как будто вы предоставляете его внешнему заказчику, а не преподавателю ИБ (чанки с кодом должны быть скрыты, графики и таблички с результатами аккуратно ``причесаны", никаких названий переменных -- только понятные лейблы на русском или английском языке, также должна быть видна логика ваших рассуждений при ответе на поставленные вопросы, выводы и интерпретация по полученным результатам).

P.S. Бонусные задания не являются обязательными, выполняются по желанию, дают возможность заработать дополнительные баллы (если вдруг вы их потеряете на каких-то других этапах выполнения задания), **в случае их невыполнения вы ничего не теряете**