Выбор контрольных переменных. Backdoor criterion

Формализуем задачу выбора подходящих контрольных переменных для исследования (к примеру, можно непосредственно включить в регрессионную модель или реализовать мэтчинг на основе релевантных «третьих» факторов). Что мы уже знаем о контрольных переменных?

- Мы включаем их в регрессионную модель наряду с ключевыми предикторами для того, чтобы уменьшить смещение в оценках. Так, они играют вспомогательную роль, относительно контрольных переменных мы не формулируем гипотез. В частности, мы предполагаем, что их эффект на зависимую переменную однородный, контрольные характеристики задают некоторый общий контекст.

- Контрольные переменные мы, как правило, находим в литературе – предшествующих исследованиях, изучающих, какие факторы влияют на зависимую переменную, интересующую и нас.

- Контрольные переменные должны быть экзогенны не только по отношению к зависимой переменной, но и по отношению к ключевым предикторам. В противном случае можно «своими руками» привнести post-treatment bias.

- Ошибочно думать, что контрольные характеристики не должны быть никак скоррелированы с другими объясняющими переменными в модели. Для того, чтобы контрольные переменные выполняли свою задачу уменьшения смещения в оценках коэффициентов при ключевых объясняющих переменных, у контролей и ключевых предикторов, разумеется, должна быть совместная изменчивость.

- Даже если эффект каких-то отдельных контрольных переменных оказался незначимым, убирать их из модели не стоит. Они в комбинации с другими контролями могут корректировать эффект ключевых предикторов.

Когда Вы задумываете количественное исследование, прежде чем переходить к реализации эмпирической части, постарайтесь наглядно изобразить, как работает механизм связи между ключевыми переменными, и обозначить, какие вспомогательные переменные можно выделить, так или иначе связанные с ключевыми предикторами и зависимой переменной (примеры таких визуализаций в виде графов – см. ниже в текущем файле). Далее на основе полученного графа можно отобрать подходящие контрольные переменные. Прежде чем сформулировать правило такого отбора, разграничим переменные-confounders и переменные-colliders.

Переменные-confounders – переменные, которые влияют и на предиктор, и на зависимую переменную (см. ниже Рис. 1), их также можно называть «вмешивающимися», или «мешающими» переменными. Это явный претендент на включение в модель в качестве контрольной переменной. Дело в том, что выявление истинного каузального эффекта переменной-treatment на зависимую переменную осложняется ложной корреляцией, возникающей между ними из-за наличия «третьей» переменной, влияющей и на X, и на Y. К примеру, вспомним распространенный пример: мы наблюдаем, что при росте количества продаж мороженого растет и количество нападений акул на людей. Из этого, конечно, было бы ошибочно сделать вывод, что рост продаж мороженого положительно влияет на количество нападений акул. Просто есть третий фактор – летний период, в который вместе с повышением температуры увеличиваются продажи мороженого и купания людей, в том числе, и в океане, где есть опасность встретиться с акулами.

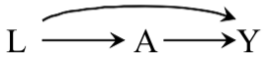
Переменные-colliders – переменные, которые являются следствием и предиктора, и зависимой переменной (или переменных, непосредственно связанных с ними). Соответствующий граф с переменной L в качестве коллайдера изображен ниже на Рис. 2. Можно представить себе мишень, в которую входят стрелы. Контролировать такую переменную нельзя, так как это приведет к ложной связи между переменными. К примеру, рассматриваем, связаны ли рост человека и умение позировать на камеру. Изначально связи между этими переменными нет, однако если проконтролируем факт, работает ли человек моделью (высокий рост и умение позировать могут выступить предпосылками, факторами, определившими выбор профессии), то получим ложную связь между ростом и умением позировать.

Теперь введем **«критерий черного хода»** в качестве правила, на основании которого можно по графу определить, какие переменные стоит контролировать, а какие - наоборот, было бы опасно. Для возможности интерпретировать результаты в терминах причинно-следственной связи и во избежание ложного вывода (к примеру, изначально связь отсутствует, а мы делаем вывод об отрицательном или положительном значимом эффекте) необходимо «заблокировать» все дополнительные пути (будем называть их «черным ходом») кроме основного, через которые от переменной treatment можно дойти до зависимой переменной. В свою очередь, «заблокировать» можно, проконтролировав любую переменную, появляющуюся как промежуточное звено на дополнительном пути – в «черном ходе», при условии, что данная переменная НЕ является переменной-collider или post-collider (то есть, следствием коллайдера).

Перед нами стоит следующая задача: определить, стоит ли включать ту или иную обозначенную на графе переменную как контрольную в модель. В качестве treatment на последующих графиках используется A / X, в качестве outcome – всегда Y.

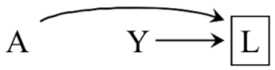
Пройдемся последовательно по DAGs (directed acyclic graphs – название говорит о том, что связи однонаправленны, при этом в графе отсутствуют циклы).

Рис. 1: Стоит ли контролировать L?



Комментарий: это классический случай контрольной переменной: L влияет и на Y, и на X - confounder. От A можно дойти к Y через «черный ход»: A – L – Y. При этом L – промежуточное звено на этом дополнительном пути. Вывод: надо контролировать L.

Рис. 2: Стоит ли контролировать L?

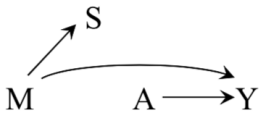


Комментарий: это классический случай переменной-коллайдера: изначально между A и Y нет связи, и переменная treatment, и outcome влияют на L. Если бы мы проконтролировали L, то между A и Y могла бы появиться ложная связь. Поэтому L контролировать не надо.

Рис. 3: Стоит ли контролировать M?

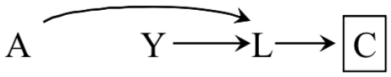
Комментарий: «черного хода» от переменной A к Y нет, поэтому M можно не проконтролировать. Между M и A нет совместной изменчивости, поэтому включение M никак не повлияет на оценку коэффициента при A. Можно проконтролировать, если хотим увеличить объяснительную силу модели. Однако на каузальный вывод о связи A и Y это никак не повлияет.

Рис. 4: Стоит ли контролировать S?



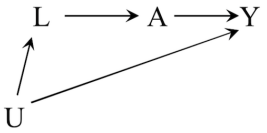
Комментарий: схожая ситуация, что и в предыдущем графе. Нейтрально, если нужно улучшить предсказание Y, а S тесно связан с M, то можно S использовать вместо M. Однако на каузальный вывод о связи A и Y это никак не повлияет.

Рис. 5: Стоит ли контролировать C?



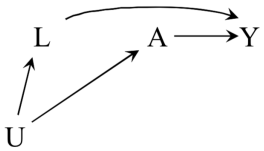
Комментарий: здесь C является post-collider, то есть, следствием коллайдера. Контролировать не надо, иначе появится ложная связь между A и Y.

Рис. 6: U - латентная переменная. Стоит ли контролировать L?



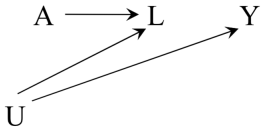
Комментарий: «черный ход» – A – L – U – Y. Блокируем его через L – контролируем переменную. Гипотетически можно было бы также заблокировать через U, но она латентная, у нас нет ее в массиве данных.

Рис. 7: U – латентная переменная. Стоит ли контролировать L?



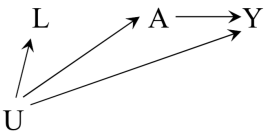
Комментарий: схожая ситуация с предыдущим графом, «черный ход» – A – U – L – Y. Блокируем его через L – контролируем переменную. Гипотетически можно было бы также заблокировать через U, но она латентная, у нас нет ее в массиве данных.

Рис. 8: U – латентная переменная. Стоит ли контролировать L?



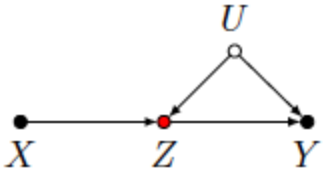
Комментарий: схожая ситуация с предыдущим графом, L – коллайдер. Если проконтролируем, то откроем «черный ход»: A – U – Y. Поэтому контролировать L не надо.

Рис. 9: U – латентная переменная. Стоит ли контролировать L?



Комментарий: на этом графе явный претендент на контроль – это переменная U. Однако U – латентная, и в массиве у нас ее нет, следовательно, как замещающую переменную можно использовать L, которая имеет совместную изменчивость с U.

Рис. 10: Стоит ли контролировать Z?



Комментарий: если проконтролировать Z, то частично учитывается U, а значит откроется backdoor X – U – Y. Это приведет к смещению каузального эффекта, поэтому контролировать не надо.