Как регрессии могут нас обманывать

Насколько оборот вагона зависит от скорости доставки грузов и к каким рискам может приводить формальный подход к использованию статистических инструментов анализа без более глубокого учета событий, происходящих в исследуемой сфере?

Фарид Хусаинов,
к. э. н., эксперт НИУ ВШЭ

Где зарыта собака?

Корреляционно-регрессионный анализ является одним из самых популярных инструментов, которыми пользуются исследователи.

Но у такого инструмента, как регрессия, при всей его эффективности есть довольно важный недостаток, о котором всегда следует помнить. Его полезно проиллюстрировать с помощью данных, привычных для тех, кто занимается анализом тех или иных процессов, происходящих на железнодорожном транспорте.

Предположим, мы хотим оценить, существует ли какая-то устойчивая связь между показателем оборота вагона рабочего парка и скоростью доставки одной грузовой отправки. Интуитивно мы понимаем, что такая связь должна быть, да и данные, приведенные на рис. 1, свидетельствуют: на фоне понижающейся динамики средней скорости доставки груза в 2010–2012 гг. (с 274 до 219 км/сут.) оборот вагона соответственно увеличивался – с 13,4 до 15,5 сут.

Затем, с 2013 года по 2018-й, скорость доставки грузов увеличивается с 223 до 370,1 км/сут. (Здесь использован показатель средней скорости, включающий в себя и груженые, и порожние грузовые отправки. Скорость груженых отправок без учета порожних несколько выше, например, в 2018 г. она составила, если верить данным ОАО «РЖД», 389,7 км/сут.) И на этом фоне несколько снижается оборот вагона (хотя масштаб изменения скорости существенно больше).

В такой ситуации для подтверждения или опровержения гипотезы о том, что эти показатели связаны, обычно используют корреляционно-регрессионный анализ.

Если мы возьмем данные о скорости доставки и обороте вагона за период 2010–2018 гг. и построим диаграмму рассеяния, то получим примерно то, что изображено на рис. 2. Линия аппроксимации всех точек облака лежит почти горизонтально, то есть, формально говоря, никакой связи – ни положительной, ни отрицательной – между показателями нет. Во всяком случае получившееся уравнение регрессии и близкий к нулю показатель «R-квадрат» (коэффициент детерминации) свидетельствуют в пользу того, что связи между скоростью доставки груза и оборотом вагона не существует.

Но если посмотреть на рис. 2 более внимательно, то можно заметить, что это неоднородное облако, внутри него можно различить два разных облака – и что-то подсказывает, что именно здесь и зарыта собака.

Если посмотреть на ту же диаграмму еще чуть-чуть внимательнее, то получится то, что изображено на рис. 3.

Определенно, здесь два разных облака, причем в первом (слева) находятся точки, соответствующие данным за 2010–2013 гг., а во втором (справа) – точки, соответствующие данным за 2014–2018 гг.

И здесь аналитик, который изучает этот вопрос, в дополнение к режиму «ученый» должен включить режим «эксперт». Под ученым обычно понимают человека, который делает исследование с помощью каких-то научных методов, а под экспертом – человека, который знает что-то про тот рынок, который мы исследуем (иногда этими свойствами обладает один человек, иногда это два разных специалиста).

Так вот, как только режим эксперта в голове будет активирован, сразу возникнет вопрос: что такого произошло на стыке между двумя указанными выше периодами?

Манипуляция скоростями

Напомню, что именно с 2014 года ОАО «РЖД» изменило методику учета общей скорости доставки груза, после чего скорость доставки резко подскочила вверх. Если в 2013-м она (для всех видов отправок) составляла 223 км/сут., то в 2014-м уже равнялась 299,2 км/сут. Причем здесь важно понимать, что не только физическая скорость перемещения объектов привела к этому росту, но и особенности учета скоростей и различных элементов времени, которые попадают в знаменатель формулы скорости. Если вагоно-километры в числителе оставить те же самые, а вагоно-часы в знаменателе уменьшить, например, исключив оттуда часть вагоно-часов простоя в брошенных поездах, то фактически скорость не изменится, но формально, в отчетности РЖД, она возрастет. Кстати, в подобных условиях зачастую у потребителей услуг железнодорожного транспорта возника-

Рис. 1. Оборот вагона и средняя скорость доставки грузовой отправки на сети РЖД в 2010–2018 гг.

Зависимость оборота вагона от средней скорости доставки одной отправки

Рис. 2

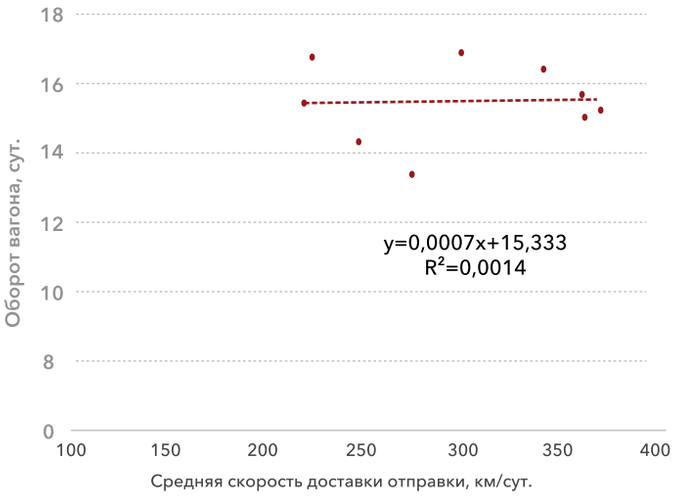


Рис. 3

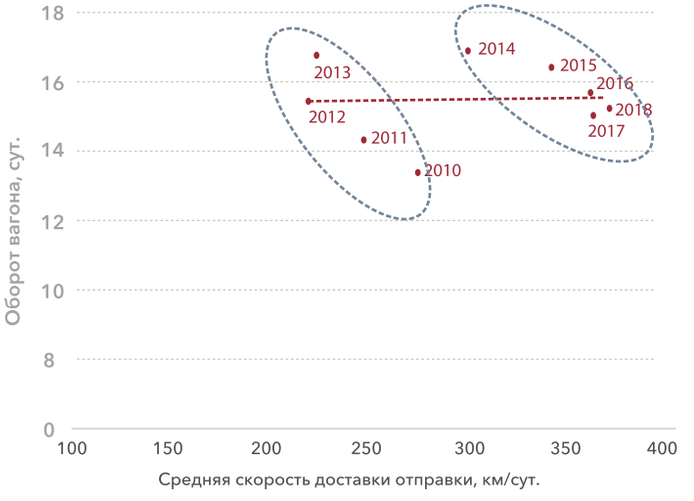
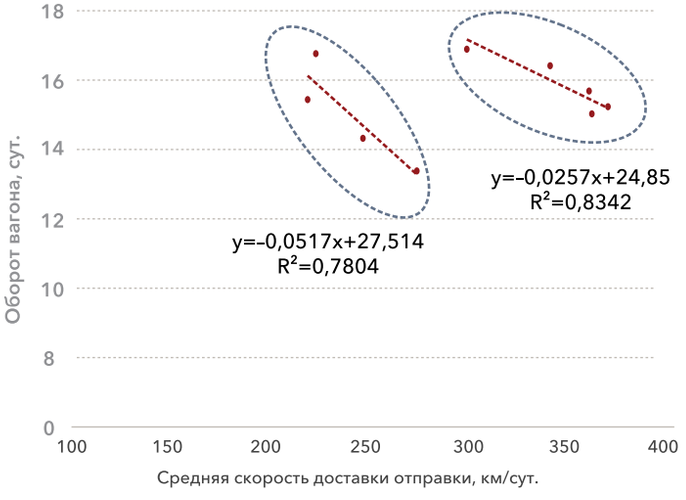


Рис. 4



ет вопрос: а какова настоящая скорость? Но об этом – ниже.

Вернемся к диаграмме рассеяния. В данном случае правильнее рассматривать эту совокупность точек не как единое облако, а как две самостоятельные совокупности – так, как это сделано на рис. 4.

В этом случае мы видим, что между оборотом вагона и скоростью доставки груза есть устойчивая отрицательная связь (линия аппроксимации имеет отрицательный наклон). И эта отрицательная связь, судя по высокому значению коэффициента детерминации («R-квадрат»), который для данных за 2010–2013 гг. составляет 0,78, а для данных за 2014–2017 гг. – 0,83, не случайна.

Таким образом, вместо первоначального вывода «связи нет» мы приходим к выводу, что, во-первых, связь есть, во-вторых, она отрицательная, и, в-третьих, зависимый показатель (а в данном случае зависимая переменная – это оборот вагона, а независимая – скорость) в довольно большой степени (примерно на 80%) обусловлен скоростью.

Так сильно могут меняться выводы, если, помимо формальных аналитических инструментов, мы используем экспертное знание.

Было бы неправильно делать здесь вывод о том, что остальные показатели, влияющие на оборот вагона, влияют в сумме только на 20%. Речь идет лишь о том, что при прочих равных условиях (то есть в ситуации, когда все остальные показатели – рабочий парк, время простоя под грузовыми операциями, на технических и промежуточных станциях и т. п. – не изменились) зависимый показатель не только связан с независимым, но и детерминирован им примерно на 80%. (Если взять квадратный корень из коэффициента детерминации 0,80, то получим коэффициент корреляции, он составит 0,89, и это тоже свидетельствует о тесной связи между этими параметрами.)

Разумеется, само наличие корреляции между двумя показателями совершенно необязательно должно свидетельствовать о наличии причинно-следственной связи. На два фактора может влиять третий (или группа факторов), и поэтому между факторами может наблюдаться корреляция. А там, где речь идет об обороте вагона, существует много факторов, которые сложным образом влияют друг на друга. Например, если один из элементов оборота вагона – простой под грузовыми операциями – не зависит от скорости доставки, то время в пути следования зависит от участковой скорости, а она, в свою очередь, находится в некоторой корреляционной зависимости от общей скорости доставки (хотя и не тождественна ей и, более того, в разные годы степень этой зависимости существенно различается, что можно наблюдать, сопоставляя динамику этих двух видов скоростей). Поэтому в данном случае пример со скоростью доставки и оборотом вагона призван лишь проиллюстрировать ситуацию, при которой формально правильный вывод, сделанный с помощью математических методов, может быть ошибочным.

Делайте верные выводы

Аналогично можно рассмотреть и влияние других факторов на оборот вагона, таких как динамика времени простоя на технических или промежуточных станциях, участковая скорость.

Анализируя связь между скоростью доставки груза и оборотом вагона, можно сделать некоторый побочный вывод.

Если посмотреть на эти два облака на диаграмме рассеяния, можно предположить, что если бы методика учета скорости не изменилась, то правое облако располагалось бы примерно там же, где и левое, и весьма вероятно, что настоящая скорость доставки грузовых отправок железнодорожным транспортом, очищенная от улучшений, произведенных в процессе изменения методики учета скорости, была бы примерно на 76–115 км/сут., или на 34–46%, ниже той, что сейчас показывают отчеты РЖД. Но это так, к слову.

Подводя итог, отметим, что роль статистических аналитических инструментов при анализе работы железнодорожного транспорта очень велика, так как именно они позволяют избежать субъективности, свойственной отдельным участникам рынка, в силу того что каждый из них видит лишь небольшую часть картины. Но вместе с тем формальное отношение к результатам статистического анализа без понимания сущностных аспектов исследуемого явления, без погружения в отраслевую специфику (условия регулирования, нормативно-правовая среда, общеэкономический контекст и т. п.) может привести к поверхностным и необоснованным выводам.