

A.V. Kurochkin

WHAT REGRESSION MODEL REALLY TELLS

Alexey Kurochkin – candidate for a Master’s Degree in Labor Economics, St. Petersburg State University of Economics, St. Petersburg; e-mail: kurochalex@yandex.ru.

We consider restrictions and claims against the use of regression models in economics. The constraints of regression models are looked at; we assess probabilities from the point of view of normal distribution and limitations of forecasts, homoscedasticity requirements, and the break in the trend line based on the case of discrepancy between the wages and productivity of labor in the USA.

Keywords: regression model; linear regression; normal distribution; homoscedasticity; trend line break; productivity of labor.

A.V. Курочкин

ЧТО НА САМОМ ДЕЛЕ ПОКАЗЫВАЕТ РЕГРЕССИОННАЯ МОДЕЛЬ

Алексей Викторович Курочкин – магистрант кафедры экономики труда, Санкт-Петербургский государственный экономический университет, г. Санкт-Петербург; e-mail: kurochalex@yandex.ru.

В статье разбираются претензии к использованию регрессионных моделей в экономической науке. Рассматриваются ограничения регрессионных моделей: оценка вероятностей исходя из нормального распределения и ограничение прогнозов, требование гомоскедастичности, слом тренда на основе кейса расхождения между заработной платой и производительностью труда в США.

Ключевые слова: регрессионная модель; линейная регрессия; нормальное распределение; гомоскедастичность; слом тренда; производительность труда.

Экономисты стремятся найти некие соотношения между экономическими параметрами и на основе тренда предсказать будущие значения показателей. Многочисленные примеры устойчивых соотношений наталкивают на мысль, что они являются закономерностями. Экономика как бы функционирует в устоявшемся режиме, и мы можем вывести какие-то устойчивые связи и даже попробовать сделать предсказание переменных по тренду.

Если исходить из тех предсказаний, которые можно было сделать по тренду, то вылетающие иногда «чёрные лебеди» оказываются чрезвычайно невероятными согласно этим моделям, тем не менее они вылетают с завидной регулярностью, на что указывает и сам автор концепции «чёрных лебедей» Нассим Талеб. Автор акцентирует внимание на том, что фактическая вероятность оказывается гораздо выше теоретической, и объясняет это тем, что рассматриваемые явления не относятся к областям, в которых действует нормальное распределение. Он говорит, что некорректно рассматривать, допустим, заработки писателей как величину того же рода, что рост и вес людей, потому что эти величины имеют разную природу, и поэтому первые показатели описываются, скорее, степенными функциями, нежели линейными, в них действует эффект масштабирования [2]. При этом Нассим Талеб не рассматривает «чёрных лебедей» как исключительно негативные явления, это может быть сильное любое отклонение от «предсказанных» значений.

Если отдельные аномальные значения и короткие периоды кризисов мы можем пытаться рассматривать как выбросы и подлежащие исключению аномалии (которые «портят» нам модель), то в данной работе мы рассмотрим смену тренда и попробуем интерпретировать её. Очевидно, что математический инструментарий экономистов требует правильной интерпретации, если они стремятся продолжать использовать его с той или иной долей успеха в глазах общества.

Само требование гомоскедастичности должно исключать увеличение дисперсии ошибок вне зависимости от того, насколько выросла зависимая величина. Однако мы понимаем, что в разных диапазонах и разных порядках величин должны действовать разные законы [1]. В противном случае получается, что относительная величина требования к размеру ошибки постоянно повышается. Огромный заработка футболиста должен быть исчислен с такой же абсолютной точностью, как и скромный заработка мигранта-подёнщика. Мы, таким образом, напрочь исключаем действие на показатель каких-либо иных непостоянных факторов, а также качественных скачков при превышении определённых показателей.

Если речь идёт о временных рядах, то говорим не о разных законах, действующих на разных уровнях, а о сломах трендов. Показательный пример этого – расхождение между ростом заработных плат и производительностью труда, которое произошло в США в 1970-х гг. Этот факт приобрёл спекулятивное значение, и авторы некоторых статей переводят это в социальную плоскость. Таблица с данными приводится в одной из статей, которая в любое время доступна в сети «Интернет», поэтому нет необходимости приводить саму таблицу [5]. Если до 1970-х гг. заработные платы росли практически параллельно с производительностью труда (приблизительно в то же число раз за определённый период), то после этого времени рост заработных плат значительно отстаёт от роста производительности труда. Мы позаимствуем у авторов статьи сырье данные, чтобы рассмотреть реальную картину и возьмем данные о средних заработных plataх и о росте производительности труда.

Мы не рассматриваем медианные заработные платы, которые также приводятся в статье, поскольку нас интересует именно взаимосвязь выработки и заработных плат в целом, а не социальная справедливость как таковая. Кроме того, эти данные приведены за период лишь с 1973 г., тогда как данные о росте средних заработных плат в экономике США и о росте производительности труда в ней же приведены с 1949 по 2014 г.

Предвосхищая замечание о том, что производительность труда зависит и от использования капитала, обратим внимание лишь на тот факт, что в расчётах в США повышение производительности труда приписывают, наряду с интенсивностью использования капитала и составом рабочей силы, так называемой «многофакторной производительности», т.е. технологическому прогрессу в интерпретации ещё Солоу [4]. Всё влияние человеческого капитала ограничивается параметром «состав рабочей силы», который определяется тем, сколько зарабатывают люди обоих полов, с различным образованием, различного возраста и т.п. [3]. Будем пока считать, что это в какой-то мере объясняет то, что мы увидим далее в регрессиях. Стоит лишь заметить, что в соответствии с этим до 1970-х гг. человеческий капитал, отражаемый в соответствующих заработных plataх, был определяющим фактором роста производительности труда, а потом перестал быть таковым, что представляется довольно парадоксальным выводом в эпоху, когда влияние человеческого капитала так активно обсуждается. Ясно, что, во-первых, затраты на человеческий капитал не могут ограничиваться лишь расходами на формальное образование и обучение на предприятии; во-вторых, человеческий капитал влияет не только на доходы собственно рабочей силы, но и на прибыль предприятия, что очень часто упускается из вида; в-третьих, сама взаимосвязь между расходами на образование и обучение и производительностью труда предприятия в такой модели затуманена, хотя сама

концепция учёта именно расходов в перспективе принципиально даёт рычаги влияния на производительность труда.

Понятно желание авторов такой концепции расчётов свести человеческий капитал к расходам на образование и некоторым характеристикам типа гендерных. Такая модель неправильна уже потому, что заработные платы в результате объясняются ими же самими, и это довольно поверхностный взгляд на вещи. Кроме того, получается, что многофакторная производительность растёт непредсказуемым образом, как ей благорассудится, и объясняет всё, что невозможно объяснить использованием практически абстрактного по Марксу (и демографически индивидуализированного) труда и столь же абстрактного (в итоге облагороженного многофакторной производительностью) капитала. Ещё предстоит сделать исследование, выявляющее реальные факторы производительности труда и величины заработных плат, мы же пока ограничимся интерпретацией регрессий.

Для наглядности, чтобы соотнести расчёты со спекулятивными объяснениями, мы ставим в качестве независимой переменной отражение «человеческого капитала», т.е. заработную плату, а в качестве зависимой переменной – производительность труда. Мы не имеем в виду, что здесь реально присутствует такая зависимость, она, скорее, должна быть обратной, но так модели и их коэффициенты будут более наглядны для интерпретации.

На всей длине выборки модель не строится удовлетворительным образом. Оказывается, что кубическая модель лучше, и даже она не даёт удовлетворительные показатели. Что касается линейной модели, то в ней средняя ошибка аппроксимации равна 75%, тогда как максимальный порог ошибки аппроксимации равен 15%.

Если мы учитываем предоставленные нам данные о неожиданном «необъяснимом расхождении» между ростом заработных плат и ростом производительности труда, то, используя диаграмму рассеяния, мы можем разделить выборку на три части: 1949–1972 гг., 1973–1991 гг. и 1992–2014 гг. (см. рисунок).

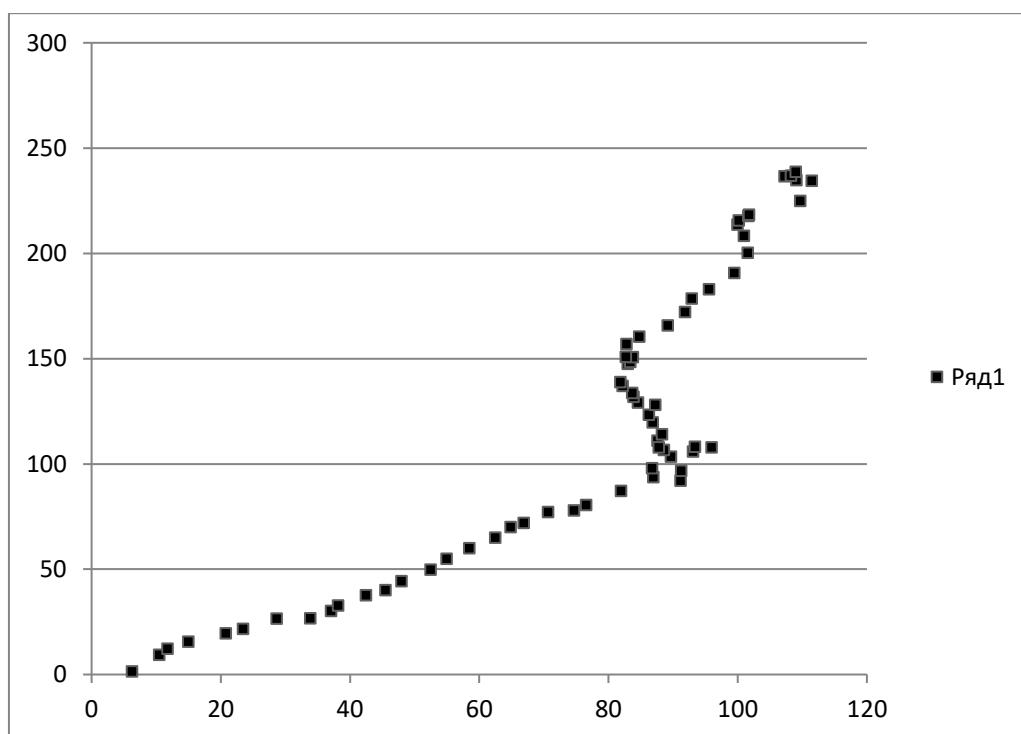


Диаграмма рассеяния для всего периода 1949–2014 гг.:
уравнение регрессии $y = -48,584 + 2,192x$ недостоверно

Уравнение регрессии для первого периода таково:
 $y = 1,0985x - 5,1336$.

Коэффициент детерминации для этой модели составляет 0,985, а средняя ошибка аппроксимации – 10,1517%.

$t_b = 37,96$. Поскольку $37,96 > 2,074$, то статистическая значимость коэффициента регрессии b подтверждается (отвергаем гипотезу о равенстве нулю этого коэффициента).

Расчетное значение RS-критерия (3,32) попадает в интервал (2,7-3,7), следовательно, выполняется свойство нормального распределения. Таким образом, модель адекватна по нормальному распределению остаточной компоненты.

$F = 1,53$. Поскольку $F < F_{kp} = 5,32$, то гипотеза об отсутствии гетероскедастичности принимается.

По тесту Дарбина-Уотсона статистика $DW = 0,44$. Поскольку $1,5 > 0,44 < 2,5$, то автокорреляция остатков присутствует. Это не делает модель полностью не валидной. Это значит, что в модель необходимо внести дополнительную переменную, что логично, поскольку такие сложные экономические взаимосвязи не могут ограничиваться одной переменной. Пока не будет изучен сам механизм этих взаимосвязей, построить полностью валидную регрессию не представляется возможным, но, в данном случае, это не мешает изучать основные моменты в интерпретации регрессионных моделей.

Второй период – это некий переход, который, как кажется, имеет отрицательный наклон на диаграмме рассеяния. Более или менее удовлетворительно он описывается только квадратичной регрессией вида

$$y = 0,374x^2 - 68,901x + 3276,6606.$$

Коэффициент детерминации для квадратичного уравнения составляет 0,6843, а средняя ошибка аппроксимации – 5,1924%.

Мы не будем оценивать линейную модель, поскольку в ней недостаточен коэффициент детерминации. Как можно видеть из диаграммы рассеивания, это объясняется не столько плохими данными в рамках этого периода, сколько их хаотичностью в начале перехода в этот период и при выходе из него, но факт остаётся фактом.

Третий период описывается линейным уравнением вида

$$y = 3,2323x - 118,515.$$

Коэффициент детерминации составляет 0,9954, а средняя ошибка аппроксимации – 3,0301%.

$t_b = 21,21$. Поскольку $21,21 > 2,08$, то статистическая значимость коэффициента регрессии b подтверждается (отвергаем гипотезу о равенстве нулю этого коэффициента).

Расчетное значение RS-критерия (3,228) попадает в интервал (2,7-3,7), следовательно, выполняется свойство нормального распределения. Таким образом, модель адекватна по нормальному распределению остаточной компоненты.

По тесту Голдфелда-Квандта $F = 4,49$. Поскольку $F < F_{kp} = 5,32$, то гипотеза об отсутствии гетероскедастичности принимается.

По тесту Дарбина-Уотсона статистика $DW = 0,82$. По таблице Дарбина-Уотсона для $n=23$ и $k=1$ (уровень значимости 5%) находим: $d_1 = 1,26$; $d_2 = 1,44$. Поскольку $1,26 > 0,82$ и $1,44 > 0,82 < 4 - 1,44$, то автокорреляция остатков присутствует. Это не делает модель полностью не валидной и говорит о необходимости ввести дополнительные переменные.

Видно, что в третьем периоде показатели следуют линейной регрессии, просто вид уравнения изменился. Что касается коэффициента b , то он возрос практически точно в 3 раза, что представляется довольно примечательным.

Отстранив всякие классовые интерпретации вопроса, мы можем прийти к выводу,

что после фазового перехода система перешла в некое иное состояние. Оно, конечно, судя по всему, не является неизменным, но в рамках рассматриваемого временного периода является строго заданным и вызванным теми или иными фактами жизни, и можно предсказывать сохранение такого положения до нового фазового перехода. Объяснением такого положения вещей может быть рост доли капитала в производительности труда, необходимость реинвестирования в капитал и в разработки, понимание капиталистом роли рабочих и т.д. и т.п. Этот вопрос предстоит изучить. Однако мы можем видеть, что неизменное положение дел весьма хорошо описывается регрессионным соотношением.

Стабильный коэффициент b указывает на то, что в экономике существует некая сложившаяся заданность, текущее расхождение в росте заработных плат и производительности труда не является произвольным в смысле неупорядоченного своеvolutionия предпринимателей, а неким устойчивым фактом.

Конечно, предсказательная сила модели ограничена тем, что, по сути, производительность и заработные платы могут колебаться любым образом, оставаясь в рамках сложившихся соотношений. Кроме того, это не позволяет предсказать фазовый переход и наступление этапа «чёрных лебедей» в талебовском смысле. Возможно, что такой этап уже наступил. Однако регрессионная модель позволяет описывать некие экономические взаимоотношения в устойчивой системе и постигать их суть. Итак, мы можем сделать следующие выводы:

1. Регрессионная модель позволяет описывать некие устойчивые в рамках данной фазы соотношения и вычленять реальные факторы, оказывающие влияние на некоторые важные для нас параметры на текущем этапе и, вероятно, в рамках построенной общей модели, – на всех этапах.
2. Предсказательная сила даже хорошо построенной регрессии в случае параметров рынка труда является ограниченной не тем вероятностным распределением, которое проистекает из заложенных данных, а возможностью неожиданного фазового перехода.
3. В частном кейсе расхождения между ростом производительности труда и заработной платы следует не поддаваться эмоциям, а попытаться выяснить, почему так происходит и что можно изменить, чтобы улучшить положение людей без пустой риторики.

ЛИТЕРАТУРА

1. Гомоскедастичность. URL: <https://wiki.loginom.ru/articles/homoscedastic-regression.html> (дата обращения: 19.04.2020).
2. Талеб Н.Н. Чёрный лебедь. Под знаком непредсказуемости. URL: <https://100i1kniga.ru/read/?book=15784> (дата обращения: 19.04.2020).
3. Changes in the Composition of Labor for BLS Multifactor Productivity Measures, 2014 // Пресс-релиз на сайте Бюро трудовой статистики США от 24 марта 2016 г. Режим доступа на территории РФ – только по архивной копии от 18.02.2020. URL: <https://web.archive.org/web/20200218033523/https://www.bls.gov/mfp/mprlabor.pdf> (дата обращения: 19.04.2020).
4. Multifactor Productivity Trends – 2019 // Пресс-релиз на сайте Бюро трудовой статистики США, график 3 (Chart 3). Режим доступа на территории РФ – только по архивной копии от 24.03.2020. URL: <https://web.archive.org/web/20200324192645/https://www.bls.gov/news.release/pdf/prod3.pdf> (дата обращения: 19.04.2020).
5. Productivity and a Typical Worker's Pay Why It Matters and Why It's Real. URL: <https://www.epi.org/publication/understanding-the-historic-divergence-between-productivity-and-a-typical-workers-pay-why-it-matters-and-whats-real/> (дата обращения: 19.04.2020).