Измерение и наглядное представление практической значимости регрессионных связей

К.К. Фурманов,
кафедра математической экономики и эконометрики НИУ ВШЭ

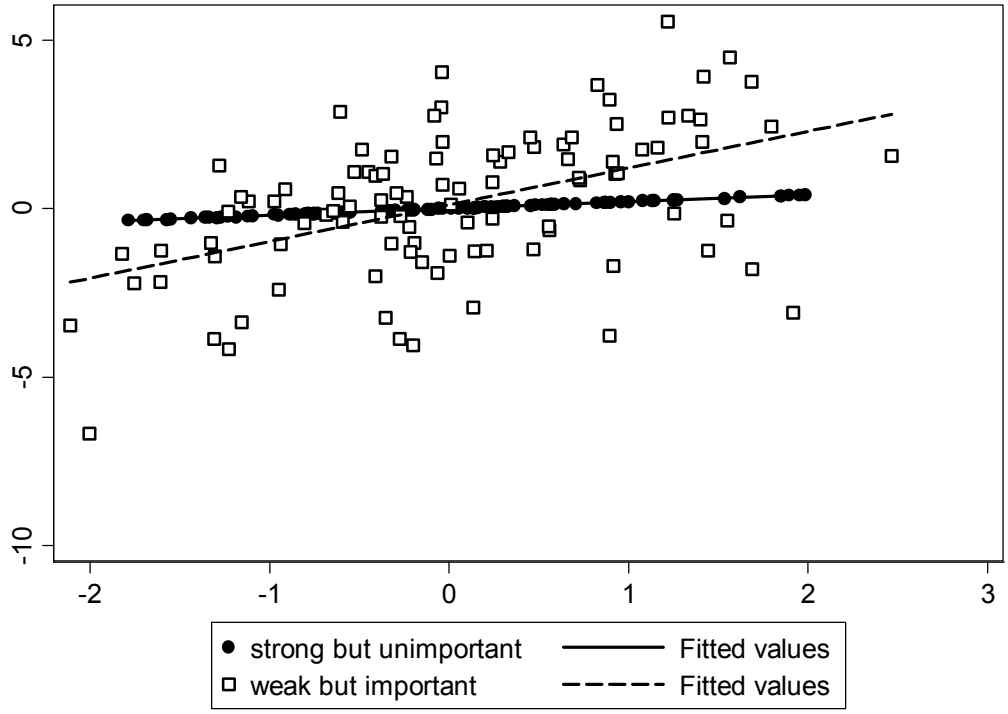
"Over time we learn about and use fancier and more abstract regression models… The utility of these fancier models diminishes if we have greater difficulty interpreting and visualizing the results"[1]

Среди экономических исследований значительную долю составляют работы, направленные на определение детерминант какой-либо интересующей исследователя величины (заработной платы, продолжительности периода безработицы и т.п.). Основным инструментом анализа в таких случаях оказывается, как правило, модель множественной регрессии. Интерпретация результатов регрессионного анализа сводится, по большей части, к двум аспектам: истолкованию оценок коэффициентов регрессии и их статистической значимости. Во многих моделях (особенно нелинейных) оценки с трудом поддаются интерпретации, что побуждает обращаться к дополнительным средствам: расчёту предельных эффектов, сравнению прогнозных значений объясняемой переменной при различных значениях детерминант, построению графиков функции отклика. Однако функции отклика – характеристика исключительно динамических моделей, а предельных эффекты и различия в прогнозах зависят от того, при каких уровнях объясняющих переменных их рассчитывать. Результаты таких расчетов сложно систематизировать, свести в одну ясную картину связи переменной отклика с регрессорами, так что основная цель моделирования – сведение большого объёма информации к небольшому числу интерпретируемых параметров – не достигается. Что касается значимости оценок, то значительное число недоразумений и вольных истолкований (среди которых – подмена статистической значимостью практической важности) позволяют утверждать, что неверная интерпретация статистической значимости – одна из наиболее распространённых ошибок в статистике (Good, Hardin, 2012). В действительности, применение аппарата проверки статистических гипотез даёт экономисту весьма скудную информацию о практической важности связи по многим причинам:

- «однобокость» вывода (можно обнаружить связь, но не её отсутствие),
- невозможность ранжировки объясняющих переменных по значимости их вклада,
- возможность высокой статистической значимости совершенно незначительных с практической точки зрения коэффициентов,
- сомнительная применимость вероятностных методов к не экспериментальным данным.

В настоящем докладе предлагается подход к измерению и наглядному представлению важности вклада объясняющих признаков в разброс отклика, позволяющий существенно дополнить традиционные способы представления оценок регрессионных моделей и приблизить исследователя к оцениванию практической значимости статистических связей. Прежде чем перейти к этому подходу, рассмотрим ещё одну опасность в интерпретации статистических оценок - подмену практической важности связи её теснотой. Рисунок ниже даёт пример связи тесной, при которой одна из величин почти не изменяется при изменении другой (чёрные кружки, линия регрессии со слабым наклоном), и пример связи куда менее тесной, при которой, однако, средний уровень одной из величин заметно зависит от значений другой (квадраты с белой заливкой, линия регрессии с большим наклоном):

[1] Michael N. Mitchell. Interpreting and Visualizing Regression Models Using Stata. Stata Press, 2012.



После рассмотрения этого рисунка может возникнут мысль, что практическую значимость отражает величина коэффициента регрессии, однако обратим внимание на два факта:

- Коэффициент регрессии зависит от единиц измерения регрессора. Всегда можно подобрать такие единицы измерения, чтобы какой-либо регрессор имел наибольший коэффициент и, таким образом, представлялся самым важным. Единицы измерения отклика тоже играют роль, но не мешают сравнению вклада объясняющих переменных. Хотя, пожалуй, стоит сделать уточнение: сравнивая два облака на рисунке, мы предполагаем, что речь идёт о связи одних и тех же статистических признаков, единицы измерения которых для обоих случаев совпадают.
- Большой коэффициент может стоять при почти не варьирующемся признаке, так что вклад этого признака в разброс объясняемой величины будет невелик.

Решение этих проблем ивзестно: чтобы ранжировать регрессоры по уровню важности их вклада в разброс объясняемой переменной можно рассчитать стандартизованные коэффициенты регрессии[2]. Стандартизованный коэффициент (иногда используется термин «бета-коэффициент») определяется так:

$$\beta_j^* = \beta_j \frac{\sigma_{x_j}}{\sigma_y}$$

Умножение на стандартное отклонение соответствующего регрессора и деление на стандартное отклонение отклика приводят к тому, что стандартизованный коэффициент не зависит от единиц измерения переменных модели. Кроме того, если коэффициенты при двух объясняющих переменных совпадают, стандартизованный коэффициент окажется больше у регрессора с большей дисперсией.

Интересный факт: стандартизованные коэффициенты почти игнорируется экономистами. О частоте использования каждый может получить представление сам, зайдя, например, на сайт repec.org и запустив поиск словосочетания «standardized coefficients» или «standardized regression coefficients». Современные учебники по эконометрике Грина, Вербика, Баума, Камерона и Триведи либо не уделяют внимания стандартизованным коэффициентам, либо упоминают о них вкратце как о способе получить оценки, не зависящие от единиц

[2] Существует множество других мер относительной важности, которые изложены в обзорах (Kruskal, 1984), (Nathans et al., 2012), (Johnson, LeBreton, 2004), (Soofi, 2000) но они, по большей части, основаны скорее на измерении тесноты связи, способности объясняющей переменной снизить необъяснённую часть разброса отклика.

измерения. Это замечание не относится к книгам по неэкономическим приложениям статистических методов, а также к учебнику (Johnston, DiNardo, 1997).

Хотя стандартизация полезна при изучении практической значимости регрессоров, потому что позволяет сравнивать их вклад в разброс объясняемого признака, стандартизованные коэффициенты всё же имеют существенные недостатки:

- Их неудобно интерпретировать. Мы можем сказать, что увеличение регрессора x_j на одно стандартное отклонение сопряжено с ожидаемым увеличением объясняемой величины на β_j^* стандартных отклонений при прочих равных условиях. Это толкование вряд ли удовлетворительно, потому что стандартное отклонение – неудобная для интерпретации характеристика.
- Нормировка на σ_y имеет нежелательный эффект: большой стандартизированный коэффициент может наблюдаться в случае, когда объясняемая переменная почти не варьируется при изменении регрессора. С практической точки зрения нам скорее важно оценивать изменение отклика в натуральных единицах.

Второй недостаток легко решаем – достаточно перейти к полустандартизованным (semistandardized) коэффициентам $\beta_j^{**} = \beta_j \sigma_{x_j}$, однако проблема неинтерпретируемости остаётся. Так как проблема эта вызвана недостатком стандартного отклонения как меры разброса, разумным решением будет использование другой меры. Отдадим предпочтение мерам, основанным на квантилях – квантильному размаху (разности двух квантилей случайной величины) и квантильному коэффициенту (отношению двух квантилей). Переход к квантильным мерам – не единственный способ превратить полустандартизованный коэффициент в нечто интерпретируемое, но именно этот способ будет нам удобен для графического представления.

Измерение вклада объясняющей переменной в случае линейной зависимости. Рассмотрим линейное уравнение $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k + \varepsilon$. Обозначим за $Q_j(p)$ функцию квантилей[3] регрессора x_j. Квантильный размах порядка γ равен $Q_j((1+\gamma)/2) - Q_j((1-\gamma)/2)$ - это длина отрезка, включающего долю γ всех наблюдений за величиной x_j. Домножив эту величину на коэффициент β_j, получаем удобную характеристику важности объясняющей переменной:

$$CS_j = \beta_j \left(Q_j((1+\gamma)/2) - Q_j((1-\gamma)/2) \right)$$

Буквы *CS* – аббревиатура для contribution spread (размах вклада). Интерпретация: если бы наблюдения в нашей выборке отличались только значениями регрессора x_j, а остальные объясняющие переменные и случайная ошибка не менялись бы от наблюдения к наблюдению, то, согласно нашей модели, квантильный размах значений объясняемой величины составил бы CS_j единиц. Иначе говоря, в средних $\gamma \times 100\%$ наблюдений наибольшее различие между величиной отклика составило бы CS_j.

Наиболее ясной такая интерпретация выглядит при использовании полного размаха – разности между наибольшим и наименьшим значением признака, но 100% размах чувствителен к выбросам. Кроме того, использование 100% размаха создаёт проблемы при распространении результатов на генеральную совокупность: наиболее часто применяемые вероятностные распределения имеют бесконечный размах. Поэтому разумнее сосредоточится на центральных 90% или 95% наблюдений – или любой другой доли по желанию исследователя.

Пример 1. Модель участия женщин в рабочей силе.

По данным о 50 штатах США[4] оценивалось уравнение:

$$LFP_i = \beta_1 + \beta_2 Income_i + \beta_3 Educ_i + \beta_4 UR_i + \varepsilon_i,$$

где LFP_i - уровень участия женщин в рабочей силе в штате *i* (%),

[3] Точнее, выборочную функцию квантилей. В дальнейшем везде речь идёт именно о выборочных характеристиках.
[4] Взяты из книги (Newbold, 2007)

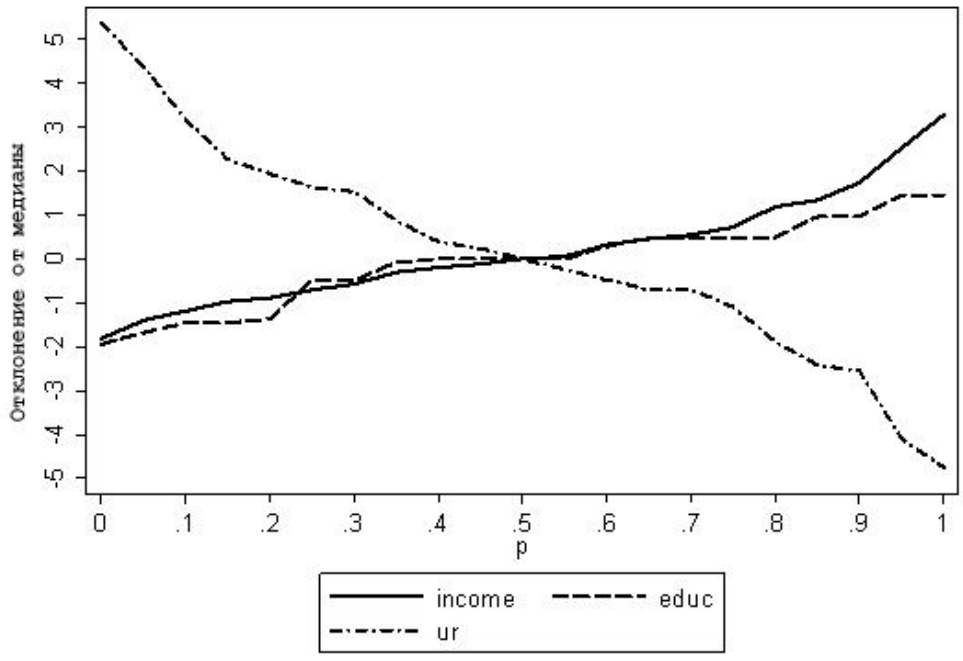
$Income_i$ - медианный доход домохозяйства (тыс. долл.),
*Educ*i - средняя продолжительность обучения среди женщин (годы),
UR_i - уровень безработицы среди женщин (%).

Таблица ниже содержит оценки коэффициентов и важности вклада каждого из объясняющих признаков:

признак	*коэфф.*	*станд. коэфф.*	*размах вклада*	*90% размах вклада*
Доход	0.406	0.257	5.090	3.915
Обучение	4.842	0.209	3.389	3.123
Безработица	-1.554	-0.510	-10.101[5]	-8.454

Сравнив стандартизированные коэффициенты регрессии, мы придём к выводу, что в рамках нашей модели наибольший вклад в различия между штатами по уровню участия женщин в рабочей силе даёт уровень безработицы. К тому же заключению приведут нас и следующие столбцы таблицы, однако приведённые в них значения позволят описать и величину этого вклада. Так, если бы все штаты различались только уровнем безработицы, то наибольшее различие между штатами в уровне участия женщин в рабочей силе составило бы 10.1%, а при отбрасывании 5% штатов с самим высоким уровнем безработицы и 5% штатов с низкой безработицей этот разрыв сократился бы до 8.5%. Различия в длительности образования при прочих равных условиях соответствовали бы размаху уровня участия женщин в рабочей силе в 3.4% для всех штатов и 3.1% для «средних» 90% штатов.

Конечно, возникает вопрос, как именно выбирать используемые квантили. Проблема легко решается графически: можно построить график, в котором по горизонтальной оси откладывались бы порядки квантилей p, а по вертикальной оси - величина $\hat{Q}_j(p)-\hat{Q}_j(0.5)$. Сдвиг квантильной функции на выборочную медиану $\hat{Q}_j(0.5)$ удобен при сопоставлении вкладов разных признаков и делает положение графика нечувствительным к выбросам. Можно отразить и направление связи, если для признаков, коэффициент перед которыми отрицателен, откладывать на графике величину $\hat{Q}_j(0.5)-\hat{Q}_j(p)$, чтобы соответствующая линия имела отрицательный наклон. Приведём такой график для оценённого уравнения участия в рабочей силе:



Квантильный размах вклада объясняющего признака отражён на этом графике как величина прироста или спада соответствующей линии между нужными квантилями. Такой график позволяет более полно представить себе важность объясняющих переменных.

[5] Размах, конечно, не может быть отрицательным. Знак «минус» добавлен, чтобы отражать направление связи.

Например, из него видно, что доход и длительность обучения имеют схожий по величине вклад в переменную отклика. Вклад дохода имеет больший размах только за счёт нескольких штатов с высоким уровнем благосостояния – об этом свидетельствует близость линий «educ» и «income» на графике и ускоренный рост линии «income» в правой части (около девятой децили и правее).

Отметим, что линии на графике – результат сдвига и пропорционального растяжения квантильных функций объясняющих признаков и потому несут в себе всю информацию о частном распределении регрессоров, так как частное распределение однозначно задаётся функцией квантилей. Поэтому графики такого рода дают наглядную поддержку не только регрессионным оценкам, но и описательной статистике.

Общий случай. Рассмотренный подход к измерению и наглядному представлению вклада может быть применён и для нелинейных зависимостей. Пусть объясняемая переменная связана с набором регрессоров и случайной составляющей следующим образом:

$$y = f(x_1,...,x_{j-1},x_{j+1},...,x_k,\varepsilon) + g(x_j), \qquad (*)$$

где $f(), g()$ - функции, на которые мы не накладываем ограничений[6]. Тогда мы можем оценить важность признака x_j по квантильному размаху значений оценённой функции $g(x_j)$ в имеющихся наблюдениях. При этом теряется возможность отразить направление связи – это, однако, не проблема способа измерения важности объясняющего признака, а просто следствие того, что связь может иметь разные направления на разных участках значений объясняющей переменной. При этом в уравнение регрессии признак может быть включён с помощью нескольких переменных – например, линейным и квадратичным членами, либо набором двоичных величин.

Не возникает трудностей и при анализе логарифмических зависимостей вида:

$$\ln y = f(x_1,...,x_{j-1},x_{j+1},...,x_k,\varepsilon) + g(x_j). \qquad (**)$$

Различие от предыдущего случая в том, что здесь можно измерить относительный вклад вместо абсолютного – оценить, *во сколько раз* изменяется величина отклика при изменениях признака x_j. Показателем важности признака будет квантильный коэффициент $Q_G((1+\gamma)/2)/Q_G((1-\gamma)/2)$ для величины $G = \exp(g(x_j))$. В силу инвариантности квантилей к монотонным преобразованиям этот коэффициент равен потенцированному квантильному размаху значений функции $g(x_j)$ - вклада признака x_j в логарифм переменной отклика. Рассмотрим этот случай на ещё одном примере из области статистических исследований рынка труда.

Пример 2. Уравнение заработной платы.

По индивидуальным данным[7] обследования рабочих в Бельгии (1994 г.) оценивалось уравнение:

$$\ln W_i = \beta_1 + \beta_2 Male_i + \beta_3 Exp_i + \beta_4 Exp_i^2 + \beta_5 E2_i + \beta_6 E3_i + \beta_7 E4_i + \beta_8 E5_i + \varepsilon_i,$$

где W_i – заработная плата *i*-го рабочего в выборке (бельгийские франки),
$Male_i$ – пол (1 – мужчина, 0 – женщина),
Exp_i – опыт работы (годы),
$E2_i$, …, $E5_i$ – дамми-переменные для уровня образования ($E5$=1 для наивысшего уровня, базовая категория - самый низкий уровень образования).

Отметим, что коэффициенты β_3, β_4 практически не поддаются интерпретации (остальные интерпретируемы в потенцированном виде) и что ни один из стандартизированных или полустандартизированных коэффициентов не имеет смысла. Для двоичных переменных бессмысленно рассматривать изменение на одно стандартное отклонение, для переменных *Exp* и Exp^2 не может идти речь о «прочих равных условиях» - абсурдно рассматривать изменение

---

[6] Здесь можно порассуждать на тему «а что если эти функции не могут быть случайными величинами», но ценность таких рассуждений сомнительна. В любом случае, выборочные квантили будут существовать, даже если нет теоретических.

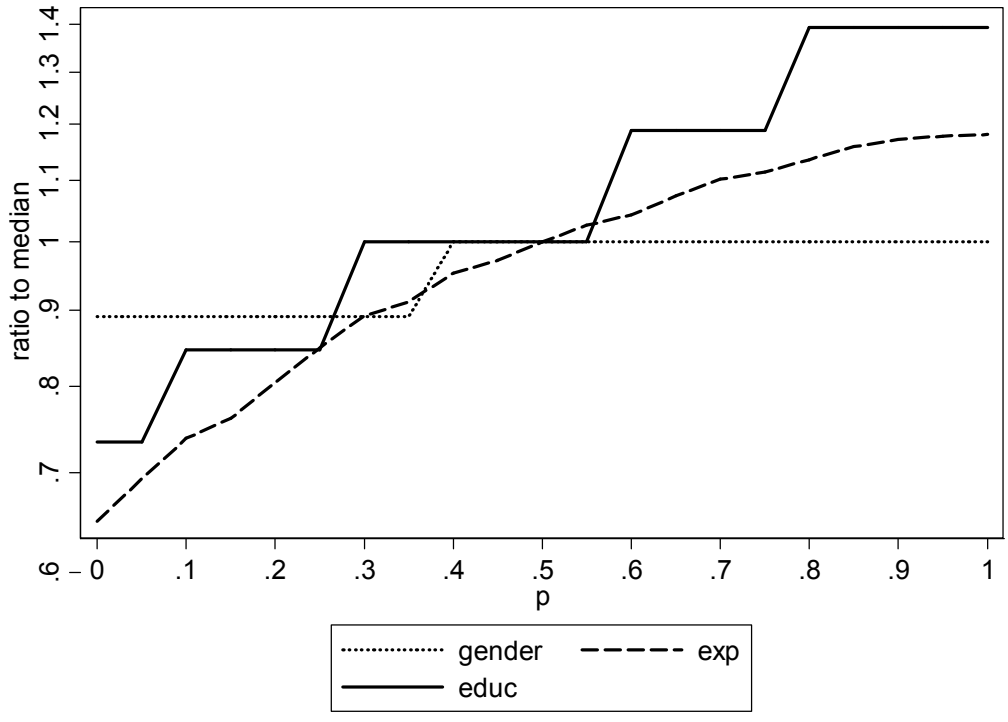
[7] взятым отсюда: http://www.wiley.com/legacy/wileychi/verbeek2ed/datasets.html - здесь выложены файлы с данными для примеров из учебника М. Вербика по эконометрике.

одной из них при постоянстве другой. Тем не менее, оценки уравнения регрессии могут быть сопровождены осмысленными мерами вкладов каждого из трёх признаков.

Признак	Коэфф.	Оценка	Потенцированная оценка	Вклад, относительный размах[8]	Вклад, квантильный коэффициент (Q(0.95)/Q(0.05))
Пол	β_2	0.115	1.122	1.122	1.122
Опыт работы	β_3	0.034	1.035	1.821	1.700
	β_4	-0.0005	0.9995		
Образование	β_5	0.141	1.152	1.898	1.898
	β_6	0.308	1.362		
	β_7	0.481	1.618		
	β_8	0.641	1.898		

Согласно полученным оценкам, наиболее существенный вклад в уровень зарплаты приходится на уровень образования: различия по этому признаку объясняют расхождение заработной платы в 1.898 раз. Схожую по величине роль играет и общий стаж. Если бы наблюдения в выборке отличались только по числу отработанных лет, то модельные значения заработной платы отличались бы не более чем в 1.821 раз во всей выборке и не более чем в 1.7 раза в средних 90% наблюдений.

Наглядное изображение важности вклада признаков даёт график отношения квантилей вкладов к медианному вкладу, в которому по горизонтальной оси откладывается порядок квантили p, а по вертикальной – величина $Q_G(p)/Q_G(0.5)$:



Так как по вертикальной оси откладывается отношение, график предпочтительно изображать в логарифмической шкале. Отражая схожий во величине вклад образования и опыта работы, график подчёркивает асимметричность: при прочих равных условиях у людей без опыта работы заработная плата отклоняется от медианной сильнее, чем у наиболее опытных.

Критика, ответ и опять критика. В отношении предложенного способа измерения практической важности вклада можно выразить ту же критику, что неоднократно высказывалась против стандартизированных коэффициентов: оговорка «при прочих равных»

[8] Под относительным размахом здесь понимается отношение наибольшего значения к наименьшему.

выглядит слишком отдаляющей от действительности. Для экономических приложений естественна связь между объясняющими переменными: изменению одной из них должны сопутствовать изменения в других. Само по себе такое замечание неоспоримо, однако стоит иметь в виду, что цель множественного регрессионного анализа – именно выделение связей, изолированных от постороннего влияния. То есть, критика относится скорее к самому подходу, при котором основой анализа становится одна модель множественной регрессии без попыток изучения опосредованных связей (mediation). При желании «освободить» какой-либо из регрессоров можно просто исключить его из модели и сравнить результаты, полученные до и после исключения – это только увеличит объём возможно полезных сведений в копилке исследователя.

Впрочем, не будем отрицать, что оговорка «при прочих равных» выглядит особенно неестественной при рассмотрении значительных изменений объясняющей переменной, соответствующих размаху выборки.

Возможные расширения применимости. Формулы (*) и (**) не стоит рассматривать как строгие границы, вне которых для измерения важности объясняющих переменных придётся искать существенно иные подходы. Заметим, что в качестве объясняемой величины y не обязательно должен фигурировать именно статистический признак – это может быть какая-либо интерпретируемая характеристика распределения этого признака, например шансы (odds) какого-либо события при моделировании бинарного выбора или функция риска (hazard function) при моделировании времени жизни.

Дополнительные возможности для иллюстрации предлагаемого в докладе подхода даёт использование панельных данных: разбиение вклада признаков на межгрупповой (between) и внутригрупповой (within) разброс, сравнение вклада наблюдаемой разнородности (отдельных переменных или всей совокупности регрессоров) с ненаблюдаемой (индивидуальным эффектом).

Уточнение в конце, которое, быть может, стоило поместить в начало. В этом докладе речь идёт о таком аспекте практической значимости, как вклад разброса объясняющего признака в разброс регрессанта при прочих равных условиях (“dispersion importance”, говоря словами К.Эйкена – см. (Achen, 1982, стр. 73-77)). Исследование практической значимости, конечно, не сводится только к этому аспекту. Это вообще не то, что можно было бы полностью формализовать и свести к математическам мерам. Цель доклада скромнее – предложить такой способ численно и графически описать связь между признаками, который был бы полезным подспорьем при выяснении практической значимости.

Основа этого текста - доклад на семинаре кафедры математической экономики и эконометрики НИУ ВШЭ, сделанный в мае 2013 года. Автор благодарит Г.Г. Канторовича, Э.Б. Ершова, Б.Б. Демешева и А.А. Пересецкого за обсуждение.

В тексте были ссылки на источники:

C.H. Achen (1982). Interpreting and using regression. Sage University Paper series on Quantitative Applications in the Social Sciences, 07-029. Beverly Hills and London: Sage Pubns.

P.I. Good, J.W. Hardin (2012). Common Errors in Statistics (and How to Avoid Them), 4th edition. Wiley.

J.W. Johnson, J.M. LeBreton (2004). History and Use of Relative Importance Indices in Organizational Research // Organizational Research Methods, Vol. 7 No. 3, pp. 238-257.

J. Johnston, J. DiNardo (1997). Econometric Methods, 4th edition. McGraw-Hill.

W. Kruskal (1984). Concepts of Relative Importance // Qüestiió, Vol. 8 No. 1, pp. 39-45.

M. N. Mitchell (2012). Interpreting and Visualizing Regression Models Using Stata. Stata Press.

L.L. Nathans, F.L. Oswald, K. Nimon (2012). Interpreting Multiple Regression: A Guidebook of Variable Importance // Practical Assessment, Research & Evaluation, Vol. 17 No.9. http://pareonline.net/getvn.asp?v=17&n=9

P. Newbold. Statistics for Business and Economics. – London, Prentice-Hall, 2007.

E.S. Soofi, J.J. Retzer, M. Yasai-Ardekani (2000). A Framework for Measuring the Importance of Variables with Applications to Management Research and Decision Models // Decision Sciences, Vol. 23 No. 3, pp. 595-625.