Использование программы SPSS для поиска зависимостей между переменными

Парадокс: экономисты пользуются *регрессиями,* но в экономической науке при этом наблюдается *прогресс.*
Сергей Гуриев

Ф.И. Хусаинов
кандидат экономических наук, доцент РОАТ МИИТ

Простая линейная регрессия

Одна из самых распространённых задач, которую приходится решать экономистам – поиск зависимости между переменными или влияние нескольких переменных на переменную, выбранную для анализа.

В самом простом случае зависимость одной переменной от другой описывается уравнением **простой линейной регрессии**. Первую переменную обычно называют «зависимой», т.к. в первоначальной гипотезе предполагается, что она зависит от другой (или нескольких) переменных. Последние называют «независимыми переменными».

В самом общем виде уравнение линейной регрессии имеет вид:

$$Y = A + B*X, \quad (1)$$

где A – коэффициент, характеризующий смещение по оси ординат;

B – коэффициент, характеризующий наклон кривой к оси абсцисс;

X – независимая переменная.

Соответственно, поиск значений коэффициентов A и B, на основе данных наблюдений (или иных эмпирических данных) – является основной задачей при определении уравнения простой линейной регрессии.

Предположим, что перед нами стоит задача, определить какие факторы влияют на объём перевозок грузов железнодорожным транспортом. При этом, у нас имеются следующие данные за 2004-2013 гг.:

– индекс промышленного производства (в процентах к прошлому году);

– объём перевозок (абсолютные значения, в млн. тонн),

– темпы прироста объёмов перевозок (индексы к прошлому году);

– коэффициенты индексации железнодорожных тарифов (индексы к прошлому году).

Покажем, каким образом можно использовать программу SPSS[1] для поиска зависимостей между указанными переменными.

Этот набор данных будет выглядеть так, как это представлено на рис. 1.

Прежде чем искать уравнение регрессии, оценим с помощью графиков, есть ли очевидные зависимости между этими показателями.

[1] *SPSS Statistics (аббревиатура англ. «Statistical Package for the Social Sciences» — «статистический пакет для социальных наук») — одна из популярных программ для статистической обработки данных*

данных2] - SPSS Statistics Data Editor

ые Преобразовать Анализ Графика Сервис Дополнения Окно Справка

4,0

ind_prom	per_mln_tonn	ind_perevoz	ind_tarif	пер
108,00	1220,9	1,0500	1,1200	
105,10	1273,1	1,0400	1,1000	
106,30	1311,3	1,0300	1,0800	
106,30	1344,2	1,0300	1,0800	
100,60	1303,7	0,9700	1,2100	
90,70	1108,2	0,8500	1,1100	
108,20	1205,8	1,0900	1,0900	
104,70	1241,5	1,0300	1,0800	

Рис. 1. Исходные данные

Необходимо проверить следующую гипотезу: «Индекс объёма перевозок железнодорожным транспортом в наибольшей степени зависит от индекса промышленного производства».

Если бы мы взяли данные не о перевозках в целом, а по отдельным грузам, то для каждого груза возможно выдвижение иных, более детализированных гипотез (например, объём перевозок зерна зависит от мировых цен на зерно и от собираемого урожая, объём перевозок нефти и нефтепродуктов зависит от мировых цен на нефть, от добычи нефти и от курса рубля к доллару и т.п.). Но в данном случае ограничимся только общим объёмом перевозок по стране.

Построение диаграммы рассеяния

Для определения наличия/отсутствия зависимости построим **диаграмму рассеяния**[2] для двух переменных: индекса промышленного производства и индекса объёмов перевозок.

Т.к. исследуемой переменной является в данном случае индекс объёмов перевозок, он будет зависимой переменной, которая традиционно откладывается по оси ординат (оси Y), а индекс промышленного производства – соответственно – независимой переменной, которую будем откладывать по оси абсцисс (ось X).

Для построения диаграммы рассеяния выберем в меню "Графика" опцию "Устаревшие диалоговые окна" и затем опцию "Рассеяние/точки", как это показано на рис. 2.

В появившемся окне "Рассеяние/Точки"(рис. 3) выберем опцию "Простая диаграмма рассеяния" и кликнем по кнопке "Задать".

Появится окно "Линейная регрессия" (рис. 4).

Здесь необходимо выбрать какую переменную по какой оси мы будем откладывать. Так как

[2] *Иногда её называют «диаграмма рассеивания»*

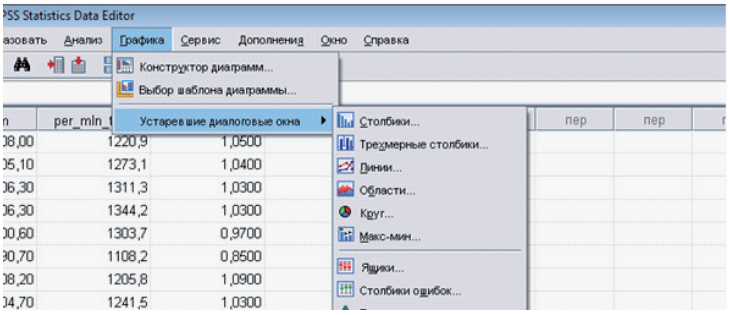

Рис. 2. Построение диаграммы рассеяния

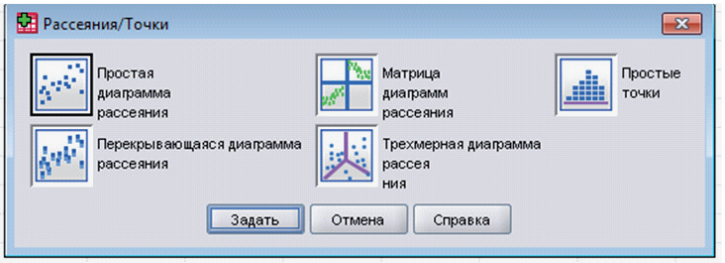

Рис. 3. Диалоговое окно "Рассеяние/Точки"

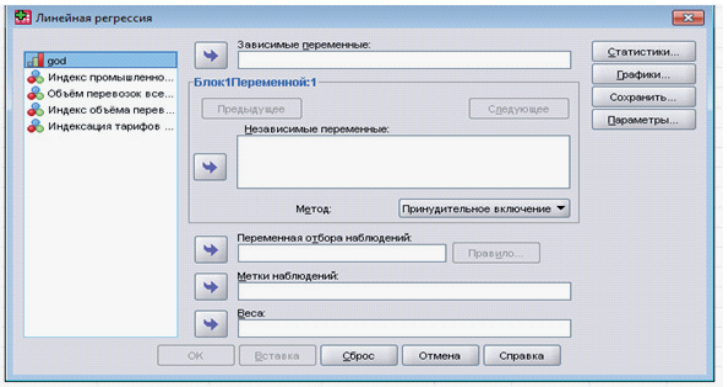

Рис. 4. Окно "Линейная регрессия"

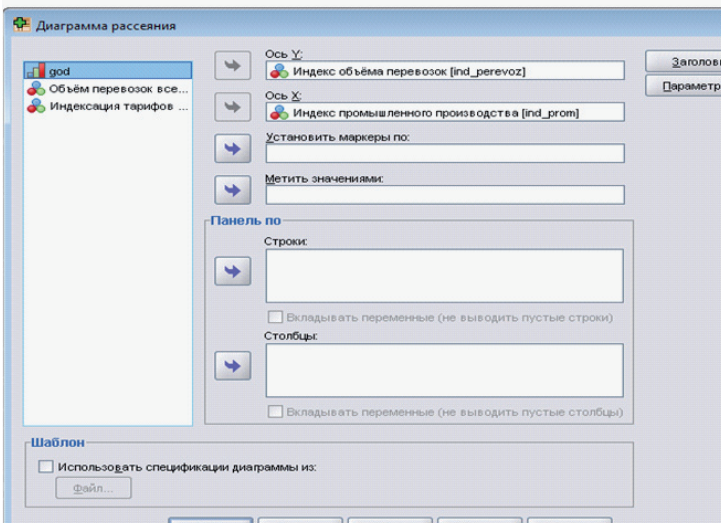

Рис. 5. Отнесение переменных к осям X и Y

выше мы уже определились, что зависимой переменной у нас будет *индекс объёма перевозок*[3], относим этот параметр к графе "ось Y", а *индекс промышленного производства*, как независимую переменную – к оси X (см. рис. 5).

Нажав кнопку "Ок" получаем диаграмму, приведённую на рис. 6, (индекс объёма перевозок в долях единицы) от промышленного производства (индекс промышленного производства в процентах к предыдущему году[4]) по данным за 2004-2013 гг.

В принципе, уже здесь (на рис. 6) видно, что зависимость между двумя переменными имеется.

Теперь построим линию аппроксимации. Для этого двойным щелчком мышки по полю диаграммы вызовем меню "Редактор диаграмм" и в меню "Элементы" выберем опцию "Аппроксимация для итога", как это показано на рис. 7. Линия аппроксимации в данном случае фактически является графиком уравнения регрессии. Линия аппроксимации для линейной регрессии – это такая прямая, для которой сумма квадратов вертикальных расстояний до соответствующих точек является минимальной (метод наименьших квадратов[5])[1].

Получим диаграмму в следующем виде (рис. 8), зависимость погрузки грузов железнодорожным транспортом (индекс объёма перевозок в долях единицы) от промышленного производства (индекс промышленного производства в процентах к предыдущему году) по данным за 2004-2013 гг.

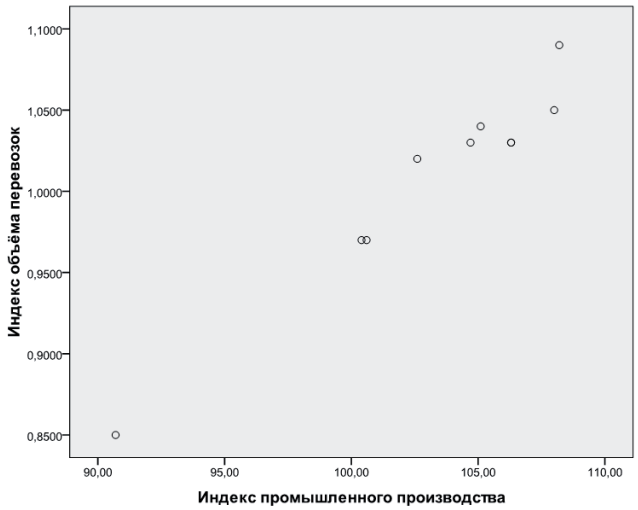

Рис. 6. Диаграмма рассеяния: зависимость погрузки грузов железнодорожным транспортом

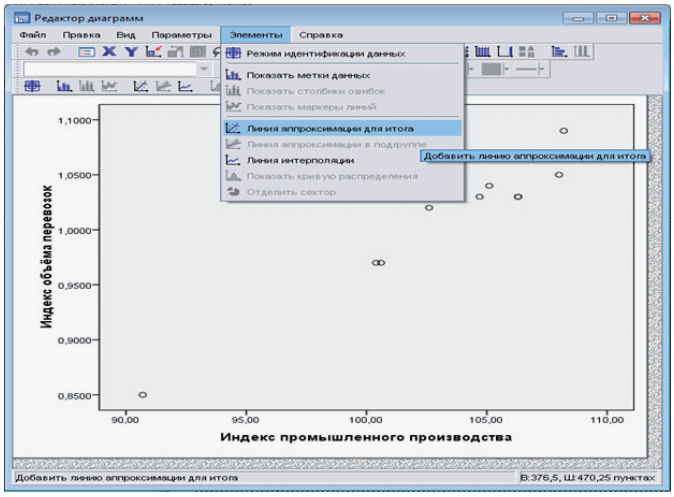

Рис. 7. Добавление линии аппроксимации на диаграмму рассеяния

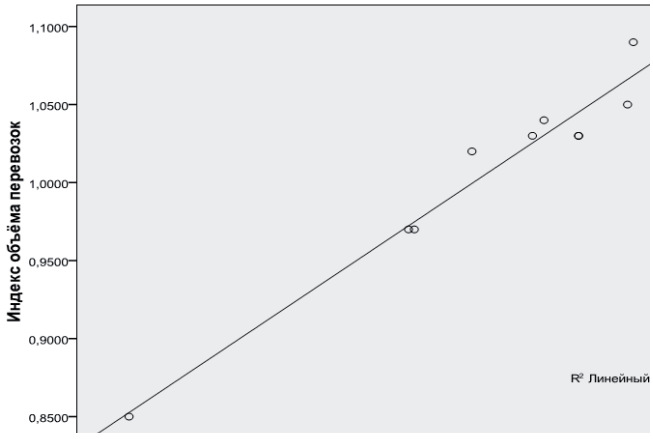

Рис. 8 Диаграмма рассеяния с линией аппроксимации

[3] Можно сделать зависимой переменной не только индекс объёма перевозок, но и абсолютное значение объёма перевозок (в млн. тонн), но зависимость будет не такой наглядной.

[4] В принципе можно перевести индексы в одинаковый вид, например оба выразить в долях единицы или оба – в процентах. В данном случае индексы приведены в разном виде, чтобы продемонстрировать, что это не влияет на результат. Главное не забыть, какой индекс в каком виде на этапе прогнозирования, на котором осуществляется подстановка данных в уравнение регрессии.

[5] Метод наименьших квадратов, являющийся частным случаем среднеквадратической аппроксимации, является самым распространённым, но не единственным методом аппроксимации данных.

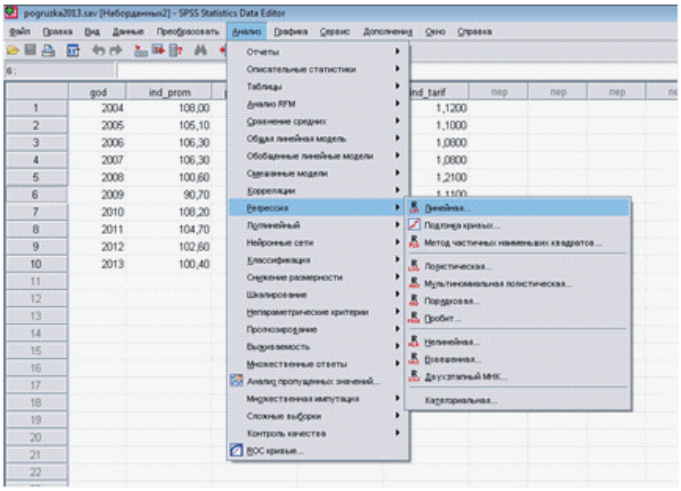

Рис. 9 Получение сводки для модели

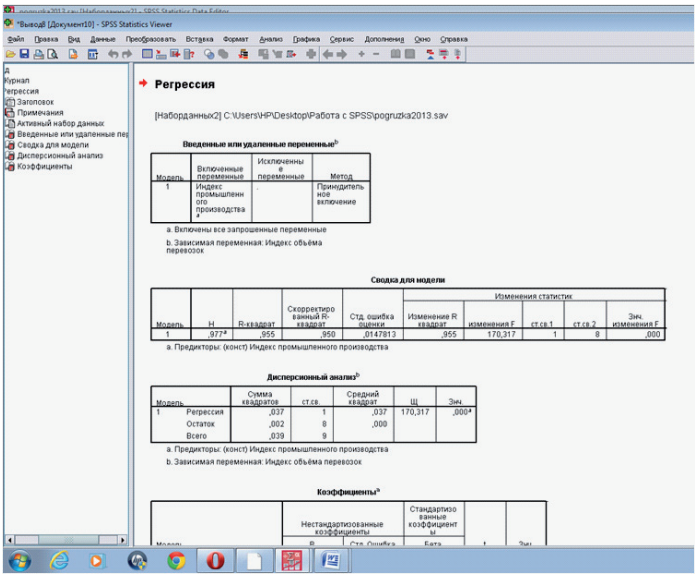

Рис. 10 Вывод данных

Таблица 1

Сводка для модели

Модель	H	R-квадрат	Скорректированный R-квадрат	Стд. ошибка оценки	Изменения статистик				
					Изменение R квадрат	изменения F	ст.св. 1	ст.св. 2	Знч. изменения F
1	,977[a]	,955	,950	,0147813	,955	170,317	1	8	,000

a. Предикторы: (конст) Индекс промышленного производства

Таблица 2

Дисперсионный анализ[b]

Модель		Сумма квадратов	ст.св.	Средний квадрат	Щ	Знч.
1	Регрессия	,037	1	,037	170,317	,000[a]
	Остаток	,002	8	,000		
	Всего	,039	9			

a. Предикторы: (конст) Индекс промышленного производства

b. Зависимая переменная: Индекс объёма перевозок

Обратим внимание, на высокое значение показателя R^2=0,955 (в SPSS он автоматически формируется в правой нижней части поля диаграммы). Этот показатель характеризует качество регрессионной прямой, т.е. степень соответствия между исходными данными и полученной регрессионной моделью [1].

Можно сказать, что этот показатель "показывает какая доля совокупной вариации в зависимой переменной описывается независимой переменной" [2].

Уравнение регрессии

Теперь мы должны построить уравнение регрессии и найти константы А и В для рассматриваемой зависимости.

Для этого в меню "Анализ" выберем опцию "Регрессия" и затем "Линейная", как это показано на рис. 9.

В появившемся окне "Линейная регрессия" (похожем на окно, приведённое выше на рис. 5) проверим указанные переменные (при желании, в этом окне можно выбрать другие переменные).

Нажав "Ок" получаем данные, приведённые на рис. 10.

Приведём эти данные в отдельных таблицах (1-3).

Подробное описание и смысл каждого из параметров, приводимых в табл. 1-3, генерируемых в окне вывода данных, приведено в работе [3].

Из таблицы 3 мы можем узнать коэффициенты А и В.

Соответственно, уравнение регрессии, полученное в результате будет таким:

$$Y = -0,267 + 0,012*X \qquad (2)$$

где Y – темп роста объёма грузовых перевозок,

X – темп роста промышленного производства.

С помощью данного уравнения регрессии можно попытаться спрогнозировать объём перевозок в будущем периоде.

Например, если объём производства на следующий год вырастет на 10 %, то индекс объёмов перевозок составит (помним, что Y – в долях

Таблица 3

Коэффициенты[a]

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.
		B	Стд. Ошибка	Бета		
1	(Константа)	-,267	,098		-2,729	,026
	Индекс промышленного производства	,012	,001	,977	13,051	,000

а. Зависимая переменная: Индекс объёма перевозок

единицы, а X – в процентах):

-0,267+0,012*110 = 1,053.

Или, иначе говоря, объём перевозок вырастет на 5,3 %.

Разумеется,таким образом можно прогнозировать лишь линейные зависимости, которыми описываются обычно лишь долгосрочные тренды.

При прогнозировании краткосрочных зависимостей с ярко выраженными сезонными циклами используются иные аналитические инструменты.

Литература

1. *Бююль А., Цёфель П.* SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей / Пер. с нем. – СПб.: ООО "ДиаСофтЮП", 2002. – 608 с.

2. *Моосмюллер Г., Ребик Н.Н.* Маркетинговые исследования с SPSS – М.: Инфра-М, 2007. – 160 с.

3. *Наследов А.* IBM SPSS 20 statistics и AMOS: профессиональный статистический анализ данных – СПб.: Питер, 2013. – 416 с.

Тахографам быть с 1 апреля

Штрафы за поездки без тахографов начнут взимать с 1 апреля 2015 года. Для граждан их размер будет варьироваться от 1 до 3 тысяч рублей, для должностных лиц от 5 до 10 тысяч. Теперь тахографы обязательны для грузовиков категории N2, не осуществляющих перевозку опасных грузов. Пользоваться тахографами обяжут грузовые автомобили независимо от марки, массой от 3,5 до 12 тонн, которые используются для коммерческих перевозок.

Европа уже сократила количество ДТП с участием транспорта, имеющего на борту тахографы. Цифры существенны – за десять лет использования контрольных устройств аварий стало меньше на 22%, а аварий со смертельным исходом на 55%.

После введения закона о тахографах в России, оптимистичнее стала и отечественная статистика. За 2014 год среди грузового транспорта количество ДТП снизилось на 0,1% (по сравнению с 2013 годом) – с 11,6 тысяч ДТП до 11,5 тысяч. При этом еще в 2012 году аварийность на дорогах РФ была куда выше – свыше 12,3 тысяч. За три года смертность на дорогах снизилась на 5,8%.

Конечно, по-прежнему, находятся те, кто не хочет ездить с тахографом, штрафы за это нежелание и неправильную установку в 2014 году получило 170 200 человек.

Для них мы и напомним основные возможности тахографа.

Тахограф позволяет контролировать режим отдыха и труда водителей и наказывать тех, кто его нарушает. Вспомните, сколько аварий происходит из-за того, что человек устал и просто уснул за рулем. Эта ситуация делает беспомощными всех, кто оказался с ним на одном отрезке пути.

Водитель, путешествующий с тахографом, защищен с правовой точки зрения гораздо эффективнее, чем водитель без тахографа, прибор документирует все данные на регистрационных картах, контрольных лентах, в памяти цифрового устройства.

Это устройство очень полезно тем, для кого машина – не роскошь и не средство передвижения, а работа. Если соблюдать предписанные тахографом режимы, то ресурс всех составляющих машины расходуется намного разумней и экономичней. Проблема объективности расчетов с заказчиками устраняется, с тахографом всегда понятен объем выполненных работ по перевозке грузов.

Как и все на свете, тахограф имеет две стороны. С его появлением в автомобиле, возникает риск того, что тахограф могут некачественно установить, есть вероятность возникновения проблем со специализированными автостоянками. Но ситуация не предполагает выбора «иметь или не иметь», 1 апреля штрафы станут реальностью.

Напоминаем, что личная карточка водителя, которая вставляется в тахограф, изготавливается 45 дней, до 1 апреля уже осталось меньше времени...

Специалисты экспертного центра "Движение без опасности"

РЖД готовы организовать движение пригородных поездов в соответствии с запросами регионов

РЖД совместно с пригородными железнодорожными перевозчиками готовы организовать движение пригородных поездов в соответствии с запросами субъектов РФ, сообщили РИА Новости в РЖД.

Раннее четыре региона - Белгородская, Пензенская, Смоленская и Ульяновская области – сообщили, что отказываются от ряда пригородных маршрутов, в том числе ранее восстановленных по поручению правительства.

В РЖД отметили, что предлагают различные альтернативные варианты, позволяющие повысить как доступность транспорта для населения, так и эффективность эксплуатации электричек и работы локомотивных бригад.

При этом департамент Смоленской области по промышленности и транспорту согласился включить в транспортный заказ только пригородные поезда сообщением Москва – Гагарин – Москва в четырехвагонном исполнении. По территории Москвы и Московской области будут курсировать электрички из 12 вагонов, а по территории Смоленской области – из четырех вагонов.

В феврале пригородные пассажирские компании восстановили все 312 электричек, отмененных в регионах РФ в январе. Сокращение электричек произошло из-за нежелания местных властей компенсировать разницу между комфортным для граждан социальным тарифом и экономически обоснованными затратами пригородных компаний.