Курсовая работа

Анализ российских домохозяйств по структуре потребления

товаров и услуг.

Факторы обеспеченности российских домохозяйств

товарами длительного пользования.

(на базе вторичного анализа данных RLMS)

Москва, 2008

**1. Введение**

В работе дается пример подхода к изучению зависимостей между доходами домохозяйств и различных факторов, влияющих на наличие в домохозяйствах товаров длительного пользования. На основе первичных данных опроса RLMS за 2004 г. построены и разобраны регрессионные модели потребления ТДП от дохода и различных социально-экономических факторов, таких как: число членов семьи, число источников дохода, местность проживания. Для исследования этой зависимости построено несколько вариантов статистических моделей линейной регрессии с различными объясняющими переменными и показана модель, которая будет наиболее точно отражать эту зависимость. Для построения статистической модели использованы многомерные статистические методы, в частности модель множественной линейной регрессии. В качестве зависимой переменной в модели будет использоваться взвешенная сумма наличия предметов длительного пользования, а в качестве объясняющих переменных – те показатели, которые я указала выше.

Для выполнения работы использовались вычислительные, графические возможности и возможности по преобразованию данных профессионального статистического пакета SPSS для Windows 14.0.

Изучение того, чем и как владеют российские домохозяйства, представляет собой актуальную задачу, так как непосредственно характеризует обеспеченность (благосостояние) семей, уровень их бытового комфорта и т.д. По данным российских статистических органов, в 2004 году российские домохозяйства тратили на такую статью расходов, как «Предметы домашнего обихода, бытовая техника, уход за домом», от 2,8 до 8,3% всех потребительских расходов (причем этот процент он тем больше, чем больше доходы у домохозяйства). Т.е. первая цифра соответствует 20% населения с наименьшими доходами, а вторая – 20% населения с наибольшими. То есть, с ростом благосостояния эта задача становится еще более актуальной.

В статистическом сборнике владение ТДП представлено в числе единиц на 100 домохозяйств (автомобили – в числе единиц на 1000 человек населения). Это разумно, т.к. позволяет сопоставлять между собой цифры, которые относятся к разным регионам страны и к различным временным периодам, то есть, характеризовать динамику владения ТДП.

Варианты постановки задачи об обеспеченности ТДП исключительно многообразны. Если посмотреть в региональном разрезе (см. сборник «Регионы России»), обеспеченность сильно варьируется по регионам страны. Так, например, в Москве на 100 домохозяйств в 2004 году приходилось 52 персональных компьютера (и это число заметно прогрессировало за последние 10 лет). Тогда как в Северной Осетии – только 5. То же самое, скажем, с автомобилями (учет которых ведется в штуках на 1000 чел. населения). В Москве – 224,2 шт, в Брянской области – 77,2 (данные того же 2004 года).

Но рассмотрение всех этих цифр не скажет ничего о том, от каких факторов *уровня домохозяйства* зависит владение ТДП. Это и составляет предмет моего исследования. Я выбрала несколько таких показателей домохозяйства и посмотрела, как они связаны с показателем обеспеченности ТДП. При этом дело не сводилось просто к загрузке данных в SPSS, т.к. перед этим они нуждались в преобразованиях.

Сложность моей задачи заключается в том, что наличие тех или иных товаров, вообще говоря, не обязательно зависит от материального положения. Например, холодильник, есть почти у всех, но он может быть старый. А телевизор сейчас и вовсе доступен многим (особенно – самые простые и дешевые модели). Поэтому, если я изучаю зависимость обладания ТДП от дохода и прочих переменных, имеет смысл сконструировать индекс ТДП так, чтобы он был связан с текущим положением домохозяйства (а не обозначал «процветание» этого домохозяйства в прошлом, что, собственно, и обозначает старый холодильник). Поэтому я попыталась в различных вариантах индекса сопоставить стоимость разных видов ТДП (то есть, приписать разные веса, например, телевизору и квартире), а также учесть срок давности покупки этих ТДП. Ведь новая квартира – не одно и то же, что старый автомобиль. Затем я посмотрела, как «объясняют» в регрессионной модели выбранные мной независимые переменные три различные варианта индекса и сделала соответствующие выводы.

Пользуясь синтаксисом, который приведен в приложении, любой желающий может повторить проведенный мной анализ при наличии данных. Там представлен и расчет показателей, и регрессионный анализ.

**2. Методы анализа данных**

Делая выбор метода анализа данных, я остановила выбор именно на регрессионном анализе, поскольку он способен объяснить взаимосвязь между многими переменными и показать, как один показатель зависит от остальных. Это именно то, что требуется в моем случае, поскольку требуется объяснить владение ТДП различными характеристиками домохозяйств. После ознакомления с литературой, описывающий данный метод, мне представляется возможным сделать следующие методические замечания, которые касаются моей задачи.

1. Регрессионный анализ предназначен для моделирования поведения одной количественной переменной от других. Следовательно, индекс обеспеченности ТДП, который я строю, должен быть количественным (а не качественным: например, высокая/средняя/низкая обеспеченность).

2. Регрессионный анализ предполагает также использование числовых переменных в качестве независимых (объясняющих). Некоторые показатели, которые есть в базе данных (например, доход) уже удовлетворяют этому требованию. Но, например, местность проживания, которую я тоже хочу учесть, так как городские домохозяйства обычно обеспечены лучше сельских, является качественной. Поэтому для нее требуется специальное преобразование, которое сделает эту переменную двоичной.

3. Регрессионный анализ является многомерным статистическим методом, то есть, учитывает больше, чем 1 взаимосвязь между признаками. Коэффициенты регрессионной модели должны интерпретироваться по принципу «при прочих равных условиях», а не каждый в отдельности. То есть, например (забегая вперед), нельзя говорить, что каждый дополнительный член семьи обеспечивает рост индекса ТДП на 0,148. Это утверждение верно лишь при прочих равных условиях, т.е. для семей с таким же доходом, таким же числом источников дохода и т.д.

4. Заложенный в SPSS регрессионный анализ является «линейным», что позволяет определить общие закономерности, но может быть недостаточно точным, если суть взаимосвязей между изучаемыми мной признаками нелинейная. Это надо тоже учесть при подготовке выводов. Но нелинейные модели, конечно, достаточно сложны. С другой стороны, если заглянуть в научные журналы, особенно зарубежные, линейный регрессионный анализ используется сплошь и рядом.

5. Качество модели регрессионного анализа определяют с помощью показателя R2 (R-квадрат). Он варьируется от 0 до 1. «0» означает абсолютно бесполезную модель, «1» - идеальную. Он же имеет интерпретацию в процентах объяснения поведения зависимой переменной. Например, R2=0,09 означает, что модель объясняет поведение зависимой переменной на 9%. Надо, забегая вперед, сказать, что качество моих моделей оказалось не очень высоким. Но это тоже важный результат. Я проверила и доказала, что индекс ТДП слабо зависит от тех переменных, которые я выбрала.

6. Имеет смысл обращать на значимость коэффициентов регрессии и значимость модели в целом (это графы Sig. в SPSS). Эти значения, наоборот, должны быть маленькими. Они как бы показывают, надежность результатов. Потому что, например, на маленькой выборке результаты могут быть не очень надежными.

7. Необходимо обратить внимание на наличии специальных кодов, которые могут содержаться в переменных. Например, в анкете RLMS если респондент отказывался отвечать на вопрос о доходе, там вбивался код «99999». Важно избавиться от этих кодов перед началом моделирования, иначе SPSS посчитает 999999 за величину дохода семьи респондента в рублях. Что, конечно, исказит результаты.

8. Перед началом моделирования необходимо изучить простые распределения переменных (т.н. Descriptive Statistics – описательная статистика), которая скажет, какие вообще есть значения у этих переменных, как часто они встречаются, какой там минимум и максимум и проч. Все это позволит проверить, подходят ли данные для анализа.

Что касается методической литературы по регрессионному анализу, то она в избытке имеется на российском рынке. Взять хотя бы книгу Э. Сигела, где регрессионный анализ рассматривается в одной из глав достаточно подробно. Этот вид анализа рассматривается в общих чертах и в книгах по общей теории статистики, например. Все это доказывает, что данный вид анализа очень важен и практически полезен.

3. Описание исходных (вторичных) данных

Задачи эссе предполагают использование данных уровня домохозяйства для построения моделей. Такие данные собираются Госкомстатом России в рамках выборочных обследований, а также в рамках специальных проектов (например, «Российский мониторинг экономики и здоровья», RLMS). Данное исследование будет построено на данных выборочного опроса RLMS. Выборка репрезентирует население России.

Объем выборки 4711 домохозяйств. Данные собирались по формализованной анкете (опроснику). Данные содержат богатую базу для различного моделирования.

База данных, которая будет использоваться для моделирования, содержит много переменных. Наиболее интересующими для нас являются переменные, которые позволяют выявить наличие тех или иных ТДП, такие как наличие ПК, стиральной машины и т.д., а также переменные, которые могут влиять/определять наличие ТДП у домохозяйств. Доходы, количество человек в семье, число источников домохозяйства и проч.

Не все переменные будут представлены в анализе так, как они представлены в исходной базе данных. Дело в том, что некоторые переменные нуждаются в преобразовании, а некоторые – в вычислении заново. Так, например, данные о месте проживания домохозяйства я преобразовала в двоичную переменную город=1/село=0, тогда как исходная переменная содержала 4 градации (в части анализа это показано). Это выгодно, т.к. мне требовалось включить эту переменную в регрессионную модель, а это допускается лишь для количественных, или для двоичных переменных. А число источников доходов я вообще считала по нескольким переменным, где респонденту предлагался целый спектр источников доходов и предлагалось согласиться или не согласиться что очередной источник доходов используется семьей респондента.

Всего в вычислениях задействовано более 30 переменных. Это отражено в синтаксисе (см. приложение).

Пример самой базы данных представлен в следующей таблице:

***Пример базы данных***

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Имя  перем. Номер семьи | iid\_h | Ic9.1a | Ic9.1b | … | indexTDP1 | … |
| 1 | 10101 | 1 | 8 | … | 0,51 | … |
| 2 | 10102 | 1 | 20 | … | 0,72 | … |
| 3 | 10103 | 1 | 24 | … | 0,54 | … |
| 4 | 10105 | 2 |  | … | 0 | … |
| 5 | 10107 | 1 | 6 | … | 1,11 | … |
| 6 | 10112 | 99 |  | … | 0 | … |
| … | … | … | … | … | … | … |

Это фрагмент данный в том формате, который отображается в SPSS. По горизонтали – семьи, по вертикали – переменные (показатели). Например, переменная iid\_h – код семьи в 13-й волне (2004 год), ic9.1a – показатель того, есть ли в домохозяйстве холодильник (код 1) , или нет (код 2), ic9.1b – содержит возраст холодильника, indexTDP1 – это уже рассчитанный мной индекс обеспеченности ТДП. Например, из тех семей, что есть в данном фрагменте, наиболее обеспеченной, вероятно, является семья №5.

Значения «99» для переменной ic9.1a не являются истинными ее значениями, а обозначают, что респондент отказался ответить на данный вопрос. Этот (и другие коды, означающие пропущенные значения), следует перед началом работы объявить пропущенными, чтобы программа исключала их из анализа.

**4. Анализ данных и интерпретация**

Для начала я проанализирую распределения тех переменных, которые я планирую использовать в регрессионных моделях. Это доход, число членов семьи, число источников дохода, городская/сельская местность, а также – показатели владения товарами длительного пользования.

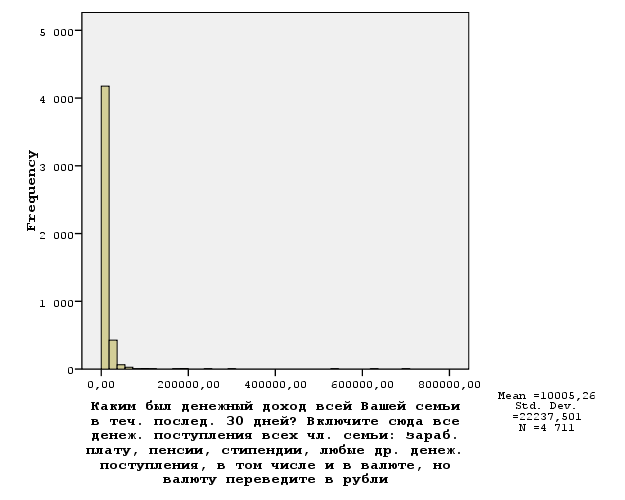
***Распределение домохозяйств по доходу***

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Число домохозяйств | Минимальный доход, руб. | Максимальный доход, руб. | Средний доход, руб. | Медиана дохода, руб. | Ст. откл. дохода, руб. |
| 4711 | 0 | 706964 | 10005 | 6400 | 22237 |

Графа «число домохозяйств» показывает, сколько домохозяйств согласились раскрыть свои доходы. Медиана меньше среднего, это означает, что на среднее значение дохода сильно повлияли семьи с большими доходами, резко отличающимися от доходов основной массы опрошенных.

Минимальный и максимальный (и даже средний) доходы еще далеко не все говорят о распределении переменной, поэтому лучшее представление о распределение дохода дает понять гистограмма распределения дохода.

***Гистограмма распределения респондентов по доходу***



Поскольку имеется незначительное число больших доходов, гистограмма не очень удачная получилась. Рассмотрим лучше квартили распределения дохода, которые скажут, с какой величины начинаются 25% минимальных доходов, медиану и с какой величины начинаются 25% самых крупных. Вот они:

|  |  |  |  |
| --- | --- | --- | --- |
| Квартили распределения дохода | I | II | III |
|  | 3100 | 6400 | 12000 |

Как видно, 50% срединных из всех наблюдаемых доходов находится в пределах от 3100 до 12000 руб.

Далее надо посмотреть, какие ТДП вообще есть в домохозяйствах. Вот их встречаемость:

***Встречаемость наличия ТДП***

|  |  |  |
| --- | --- | --- |
| **Вид ТДП** | **Количество семей** | **% от общего числа семей** |
| Холодильник | 4433 | 94,1 |
| Отдельная морозильная камера | 325 | 6,9 |
| Стиральная машина | 3800 | 80,7 |
| ч/б телевизор | 584 | 12,4 |
| цв. телевизор | 4180 | 88,7 |
| Видеомагнитофон/видеоплеер | 2225 | 47,2 |
| Компьютер | 823 | 17,5 |
| Легковой автомобиль | 1322 | 28,3 |
| Грузовой автомобиль | 89 | 1,9 |
| Мотоцикл, мотороллер, моторная лодка | 180 | 3,8 |
| Трактор/ минитрактор | 61 | 1,3 |
| Дача, другой дом, часть дома, садовый домик | 995 | 21,1 |
| Другая квартира / часть квартиры | 271 | 5,8 |

Как видно, наиболее распространенным ТДП из данного перечня является холодильник (он есть почтив каждом домохозяйстве), а наименее распространенными следует признать тракторы, грузовые автомобили, мотоциклы и мотороллеры. Суммарный процент, естественно, превосходит 100%, поскольку одно и то же домохозяйство может владеть несколькими ТДП одновременно. Эти цифры, в принципе, соответствуют тем, что предоставляет Госкомстат России, говоря об обеспеченности домашних хозяйств товарами длительного пользования. Но они рассчитывают обеспеченность в количестве единиц на 100 домохозяйств, поэтому представление данных немного другое. Например, в 2004 году на 100 домохозяйств в РФ приходилось: телевизоров – 135 шт., видеомагнитофонов/камер: 62 шт., персональных компьютеров: 28 шт., холодильников/морозильников: 114 шт., стиральных машин: 94 шт.

В модель также будет включена переменная, которая показывает, в каком типе населенного пункта проживает домохозяйство, т.к. как правило, в больших развитых городах, уровень благосостояния домохозяйств, проживающих там, значительно выше, что в свою очередь влияет на наличие ТПД в этих домохозяйствах.

***Распределение по типу населенного пункта, в котором проживает домохозяйство***

|  |  |  |  |
| --- | --- | --- | --- |
|  | | Частота | Процент |
|  | Областной центр | 2017 | 42,8 |
|  | Город | 1249 | 26,5 |
|  | ПГТ | 292 | 6,2 |
|  | Село | 1153 | 24,5 |
|  | Всего | 4711 | 100 |

Переменная типа населенного пункта где проживает домохозяйство, будет представлена в модели дихотомизированной, с двумя значениями - город (код 1, 69,3% опрошенных домохозяйств) и соответственно населенный пункт сельского типа (код 0, 30,7% опрошенных). Это нужно для того, чтобы включить данную качественную переменную в количественный анализ, т.е в регрессионную модель. Коэффициент перед данной переменной в модели будет показывать, как влияет проживание в городе на наличие ТДП.

Распределение по числу членов домохозяйства таково. Примерно 19% домохозяйств состоят из 1 человека, еще почти 28% - из 2-х человек, еще столько же – из 3-х, еще 16,5% - из 4-х. Все прочие (более крупные) домохозяйства дают, в целом, 10%.

Наконец, в модели будет присутствовать переменная, характеризующая число источников дохода. Минимальное значение данной переменной: 0, максимальное – 8. В среднем российское домохозяйство имеет 2,4 источника дохода из тех, что были представлены в анкете RLMS. Стандартное отклонение: 1,2. Если говорить в процентах, то не имеют источников дохода, примерно, 1,5% домохозяйств, 1 источник имеют 24,6%, 2 – 32%, 3 – 24,7%. Это наиболее распространенные показатели. Большее число источников имеют, в целом, не более 17% домохозяйств.

Чтобы посмотреть как влияют рассмотренные переменные на владение товарами длительного пользования, необходим показатель, который бы вобрал в себя эту информацию. В данном эссе я предлагаю 3 варианта вычисления такого показателя на основе сведений, которые имеются в анкете.

Первый вариант расчета показателя самый сложный. Допустим, у нас есть сведения о 13 ТДП (см. выше). Известно, есть тот или иной товар в семье, или нет, а также – известно, сколько лет этому товару. Мне кажется, имеет смысл не только подсчитывать суммарный индекс как число товаров, которое есть в домохозяйстве, но и попытаться учесть разную стоимость этих товаров и разный срок их службы. В самом деле, автомобиль, выпущенный в середине 1990-х гг и автомобиль, выпущенный в прошлом году – не одно и то же. Чтобы получить итоговый показатель, я суммирую следующие произведения отдельно по каждому виду ТДП:

«есть (1) или нет (0) товар в домохозяйстве» \* «вес товара» \*   
максимум из «0 и разности (10 – возраст товара)».

С первой частью ясно. Если товар есть, мы ставим вместо этой части 1, если нет – 0. Вес товара определяем так, что дополнительная квартира имеет вес, равный 1, автомобиль – вес, равный 0,1, компьютер – вес, равный 0,04 и т.д. Веса я сама предложила, исходя из примерного соотношения стоимостей этих ТДП. Таким образом, иметь, Например, компьютер и автомобиль – не одно и то же, что иметь дополнительную квартиру, в общем случае. Последняя составляющая нужна для того, чтобы учесть возраст товара, и приписать больший вес товарам, которые много моложе 10 лет. Если же товар 10 лет и старше, то разность может быть меньше нуля. И чтобы не делать индекс отрицательным, мы выбираем максимум из 0 и возможного отрицательного значения, то есть, зануляем слагаемое для данного товара. Так мы складываем эти результаты по всем товарам.

Мне кажется, это довольно трезвый способ расчета индекса обеспеченности ТДП, полностью учитывающий всю информацию о них, которая есть в анкете. Максимум этот индекс приобретает в том случае, если семья обеспечена всеми товарами, но в первую очередь – самыми дорогостоящими, и к тому же, если эти товары относительно новые.

Еще 2 варианта расчета показателя я предложила, честно говоря, после того, как эксперименты с первым показателем обнаружили не очень хорошую объясняющую способность регрессионных моделей.

Второй вариант заключается в том, что веса товаров принимаются за единицу (или не учитываются), но возраст продолжает учитываться. Т.е. мы изымаем вторую часть формулы, представленной выше.

Третий, самый простой способ расчета этого показателя заключается в том, все веса принимаются за единицу, а возраст товара не учитывается. Т.е. это просто – число товаров из списка вещей, которыми владеет семья.

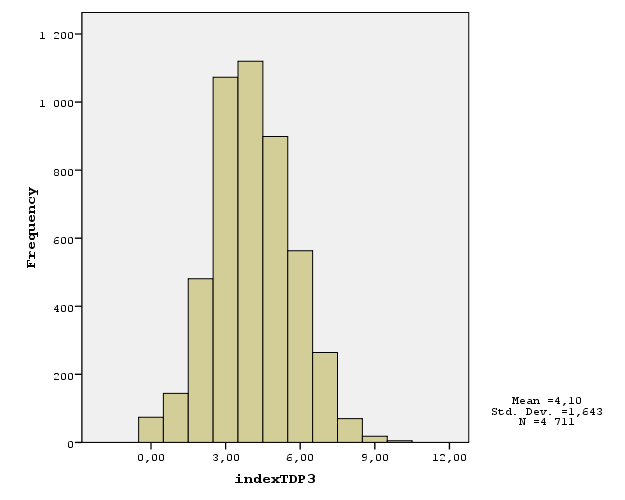
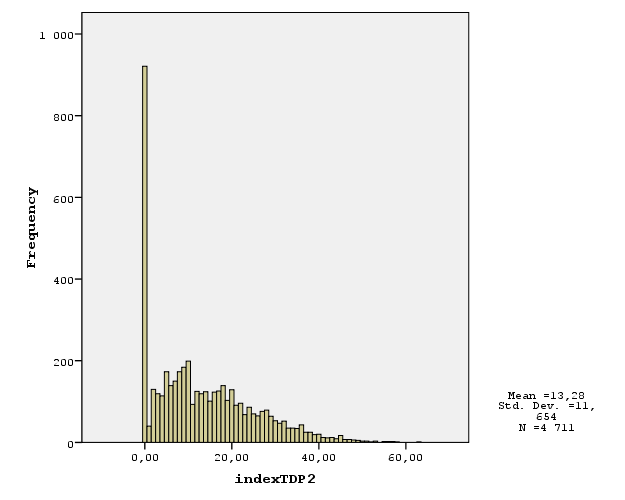
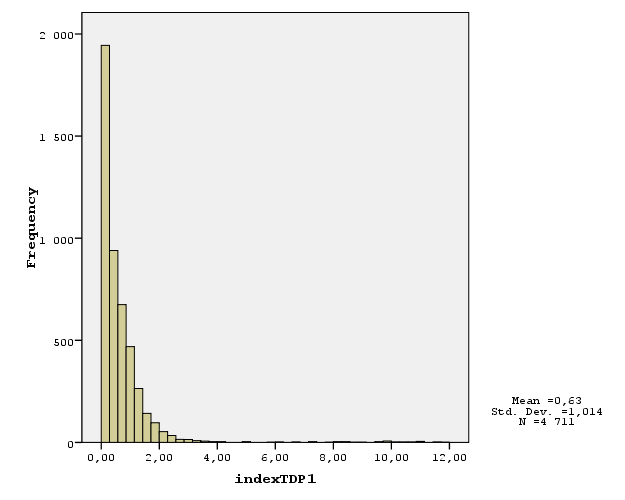
Показатели называются indexTDP1, indexTDP2 и indexTDP3, соответственно. Рассмотрим их распределения.

**Описательная статистика показателей обеспеченности ТДП**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | N | Minimum | Maximum | Mean | Std. Deviation |
| indexTDP1 | 4711 | .00 | 11.84 | .6333 | 1.01381 |
| indexTDP2 | 4711 | .00 | 63.00 | 13.2815 | 11.65379 |
| indexTDP3 | 4711 | .00 | 10.00 | 4.0964 | 1.64273 |

Есть семьи, где рассчитанный показатель наличия ТДП равен нулю, это говорит о том, что у этих домохозяйств вовсе отсутствуют перечисленные товары, либо они довольно старые (либо они просто не аккуратно ответили на данный вопрос анкеты).

Гистограммы распределения показателей следующие:



Как видно, для первого и второго показателей присутствует большая группа семей, для которых они равны 0 (или около того). Это происходит, скорее всего, потому, что вещи есть, но они достаточно старые (старше 9 лет), поэтому индексы их учитывают как будто бы их нет вовсе.

Наконец, для изучения зависимости между потреблением ТДП от дохода и других социально-экономических факторов строятся несколько регрессионных моделей.

Ниже отдельно для каждого из трех показателей приводится отдельная таблица, содержащая 3 или 4 модели. Первая модель всегда содержит только одну независимую переменную, вторая – две и так далее. Такое представление результатов позволяет показать, как добавление последующей объясняющей переменной отражается на качестве модели. Для добавления переменных в модель был использован алгоритм Stepwise программы SPSS. Значки Sig. показывают значимость модели в целом и коэффициентов по отдельности. Считается, что если Sig. близок к 0, то модель значима, то есть, имеет смысл.

Рассмотрим модели с первым показателем. Качество моделей (хотя они и значимы) очень низкое. Так как показатель R2 близок к 0. То есть, независимые переменные (доход, число источников дохода и проч.) плохо объясняют изменения переменной индекса владения ТДП. Максимум – на 7%. В первую очередь программа включила в модель число членов семьи, затем – двоичную переменную город/село, затем – суммарный доход и, наконец, число источников дохода. Интерпретация для 4-й модели, например, такая: каждый дополнительный член домохозяйства, при прочих равных условиях, увеличивает индекс потребления ТДП, в среднем, на 0,148 единиц. Иными словами, более крупные семьи, обычно, владеют большим набором ТДП. Что естественно. Аналогично, если домохозяйство из городской среды, то нужно увеличить прогноз индекса на 0,269. Каждый доп. рубль суммарного дохода домохозяйства увеличивает индекс на 0,00000514 единиц. А вот чем больше источников дохода, тем меньше как бы становится индекс. В части выводов причина этого обсуждается.

***Регрессионные модели потребления ТДП (первый вариант)***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Модель №** | **1** | **2** | **3** | **4** |
| Константа | 0,223 (Sig.=0,000) | 0,013 (Sig.=0,741) | 0,016 (Sig.=0,688) | 0,077 (Sig.=0,074) |
| Число членов семьи | 0,148 (Sig.=0,000) | 0,154 (Sig.=0,000) | 0,140 (Sig.=0,000) | 0,148 (Sig.=0,000) |
| Городская местность проживания | – | 0,279 (Sig.=0,000) | 0,254 (Sig.=0,000) | 0,269 (Sig.=0,000) |
| Суммарный доход домохозяйства за последние 30 дней | – | – | 5,10Е-006 (Sig.=0,000) | 5,14Е-006 (Sig.=0,000) |
| Число источников дохода за последнее время | – | – | – | –0,039 (Sig.=0,001) |
|  | **0,043** | **0,059** | **0,071** | **0,073** |
| **Значимость модели (Sig.)** | **0,000** | **0,000** | **0,000** | **0,000** |

Модель с использованием второго варианта индекса (без учета весов ТДП) кажется более удачной. Но и здесь объясняющая способность не превышает 16% для 4-х объясняющих переменных. Суть коэффициентов остается той же. Но они теперь чуть-чуть побольше, поскольку второй индекс имеет больший масштаб изменений.

***Регрессионные модели потребления ТДП (второй вариант)***

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Модель №** | **1** | **2** | **3** | **4** |
| Константа | 5,969 (Sig.=0,000) | 2,623 (Sig.=0,000) | 2,662 (Sig.=0,000) | 3,574 (Sig.=0,000) |
| Число членов семьи | 2,629 (Sig.=0,000) | 2,725 (Sig.=0,000) | 2,533 (Sig.=0,000) | 2,639 (Sig.=0,000) |
| Городская местность проживания | – | 4,443 (Sig.=0,000) | 4,087 (Sig.=0,000) | 4,314 (Sig.=0,000) |
| Суммарный доход домохозяйства за последние 30 дней | – | – | 7,41Е-005 (Sig.=0,000) | 7,46Е-005 (Sig.=0,000) |
| Число источников дохода за последнее время | – | – | – | –0,572 (Sig.=0,000) |
|  | **0,103** | **0,133** | **0,153** | **0,156** |
| **Значимость модели (Sig.)** | **0,000** | **0,000** | **0,000** | **0,000** |

Третий вариант индекса наилучшим образом подходит для моделирования (см. табл. ниже). Но и здесь 3-я модель объясняет лишь 18,7% колеблемости индекса. Число источников дохода здесь включать было излишне и SPSS не включил.

***Регрессионные модели потребления ТДП (третий вариант)***

|  |  |  |  |
| --- | --- | --- | --- |
| **Модель №** | **1** | **2** | **3** |
| Константа | 2,966 (Sig.=0,000) | 2,412 (Sig.=0,000) | 2,418 (Sig.=0,000) |
| Число членов семьи | 0,407 (Sig.=0,000) | 0,422 (Sig.=0,000) | 0,394 (Sig.=0,000) |
| Городская местность проживания | – | 0,735 (Sig.=0,000) | 0,683 (Sig.=0,000) |
| Суммарный доход домохозяйства за последние 30 дней | – | – | 1,09Е-005 (Sig.=0,000) |
|  | **0,124** | **0,1166** | **0,187** |
| **Значимость модели (Sig.)** | **0,000** | **0,000** | **0,000** |

Итак, поставленная цель была достигнута. Регрессионные модели построены. Они не очень хорошо объясняют все три варианта индекса, видимо, потому, что все-таки наличи в домохозяйстве ТДП объясняется, в основном, не этими переменными, а теми, что я не учла в анализе.

В следующем разделе обсуждаются некоторые выводы, которые можно из всего этого сделать.

**5. Выводы**

В ходе работы был проведен анализ данных опроса RLMS волны 2004 года. Основной целью работы было исследование зависимости наличия у домохозяйств товаров длительного пользования от доходов и различных социально-экономических факторов (числа членов домохозяйства, числа источников доходов, местности проживания).

Для учета наличия в домохозяйстве товаров длительного пользования было построено 3 варианта индикатора. Первый вариант, наиболее сложный, учитывал наличие ТДП по 13 позициям, веса различных ТДП (например, компьютер имеет значительно меньший вес, чем автомобиль), а также – количество лет, сколько данный товар уже используется в домохозяйстве. Логика данного индикатора такова, что чем больше дорогих товаров длительного пользования имеется в домохозяйстве и чем эти товары «моложе» (т.е. куплены относительно недавно), тем больше величина индикатора. Второй вариант упрощает логику первого индикатора, исключая из него веса ТДП. Т.е. теперь, например, дополнительная квартира и стиральная машинка имеют один и тот же вес. Но срок давности этих вещей по-прежнему учитывался. Этим я как бы проверяю обоснованность назначения весов товарам длительного пользования. Третий вариант индикатора еще проще. Он является простым пересчетом различных ТДП, без учета их возраста и весов. Это самый простой вариант.

Зависимости всех этих 3 индикаторов последовательно изучались в 3 блоках регрессионных моделей. Сначала изучалось влияние независимый переменных на 1-й индикатор, затем – на 2-й, затем – на 3-й. При этом в каждом блоке строилась не одна, а несколько моделей, т.е. сначала включалась одна независимая переменна, затем к ней добавлялась вторая и т.д. Для этого был использован метод анализа Stepwise пакета SPSS. Этот метод сам решает, нужно ли включить переменную в анализ, или нет. В итоге в 1 и 2 блоках были включены все независимые переменные, а в 3 блоке – все за исключением числа источников дохода. С добавлением каждой из независимых переменных в модель, объясняющая способность модели возрастала, при этом построенная в итоге модель множественной линейной регрессии была значимой (значения Sig. в таблице ANOVA были малы).

К сожалению, ни в одном блоке не удалось добиться высокого показателя качества регрессионной модели R2. Он был далек от единицы во всех случаях. Хуже всего поддавался моделированию 1-й показатель (самый сложный). Включением в модель всех 4 независимых переменных удалось добиться «объяснения» показателя лишь на R2=0,073 (т.е. на 7%). Это, конечно, мало. Второй показатель показал себя лучше. Он объяснялся максимум на 15,6%, что, хотя, тоже немного. Третий показатель показал себя чуть лучше второго. Он объяснялся на 18,7%.

Таким образом, основной вывод, который мы можем сделать – это тот, что отчасти показатель наличия, давности приобретения и веса ТДП в домохозяйстве объясняются текущими показателями дохода, числа источников дохода, а также – числом членов семьи и местностью проживания, но, вообще, зависимость от всех этих переменных – довольно слабая. Во всяком случае, не превышает 20%.

Другой вывод, который можно сделать – это тот, что SPSS включал переменные в модель всегда в одном и том же порядке. Сначала число членов семьи, затем – город, затем – доход, затем – число источников дохода (кроме блока 3). Наверное, это логично, поскольку разнообразие товаров длительного пользования, конечно, во многом зависит от размера семьи. В большой семье сложно обойтись без основных вещей. Наличие ТДП, конечно, зависит и от местности проживания, поскольку городские жители все же пока лучше, чем сельские обеспечены самым необходимым. Кроме этого, сельские жители редко владеют, скажем, дачами, т.е. у них показатель ТДП часто оказывается заниженным. С другой стороны, городские жители, например, реже, чем сельские, владеют тракторами. То, что доход домохозяйства за последние 30 дней находился далеко не на 1-м месте, значит, наверное, то, что, хотя мы и пытались учесть срок давности приобретения ТДП, но все же это товары длительного пользования, а, значит, их наличие лишь в очень небольшой степени объясняется доходом за последний месяц.

Замечу, что для всех независимых переменных коэффициенты были положительными, за исключением числа источников дохода. Получается, что чем больше у домохозяйства источников дохода, чем меньше у него индекс ТДП. Конечно, эта переменная влияет на индекс слабее остальных, но все же может показаться странным, что большой спектр источников дохода оборачивается малым количеством (или большой давностью ТДП). Я думаю, этот «парадокс» объясняется довольно просто. При подсчете числа источников доходов мы учитывали и такие источники, как пенсия, субсидии, помощь от государственных и негосударственных организаций, помощь родственников и других людей (в том числе – не только деньгами, но и вещами). Получается, что большое число источников дохода – не показатель благополучия домохозяйства, а, скорее, наоборот - обозначение того, что семья вынуждена прибегать к помощи со стороны. Тогда как состоятельные семьи часто существуют, в основном, на зарплату и, может быть, проценты от акций и т.д. Я считаю это довольно интересным выводом.

Довольно тяжело объяснить, почему аккуратный учет располагаемых ТДП в домохозяйстве выражающийся индексом №1, оказался хуже, чем остальные индексы, которые не учитывают, ни вес ТДП, ни их возраст. Может быть, это от того, что c увеличением дохода потребление различных товаров длительного пользования изменяется в разной степени независимо от их цен. А может быть мы просто подобрали такие веса, которые не точно соответствуют соотношениям цен на товары. Может, сложность заключается еще в том, что у нас как бы смешались ТДП, которые есть почти в каждой семье (холодильник, телевизор) и товары, которые есть лишь у некоторых (автомобиль, компьютер, дополнительная квартира). Возможно, проблема состоит еще и в том, что, если учитывать возраст вещей и не учитывать вещи, которые старше 10 лет (как это было сделано в индексах №1 и 2), то около 20% домохозяйств имеют индекс ТДП, равный 0, т.е. вовсе не имеют вещей, которые нас интересуют. А для третьего индекса таких домохозяйств только 1%.

Итак, в результате проведенных исследований мы выяснили, что зависимость потребления ТДП от дохода и других социально-экономических факторов можно описать с помощью множественной линейной регрессии, но далеко не полностью.

**6. Литература**

1. Салин В.Н., Шпаковская Е.П. Социально-экономическая статистика: Учебник. – М.: Юристъ, 2001. – 461 с.
2. Социальная статистика: Учебник / Под ред. чл.-кор. РАН И.И. Елисеевой. – 3-е изд., перераб и доп. – М. Финансы и статистика, 2002. – 480 с.
3. Социальное положение и уровень жизни населения России: Стат. сб. / Госкомстат России. – М., 2001. – 463 с.
4. SPSS Base 14.0 Руководство пользователя. – SPSS Inc, 2005. – 814 с.
5. Российский статистический ежегодник. 2005: Стат. сб. / Росстат. – М., 2006. – 819 с.
6. Сигел, Эндрю. Практическая бизнес-статистика. : Пер. с англ. – М. : Издательский дом «Вильямс», 2002. – 1056 с.

7. Приложения

Командный синтаксис SPSS-15 для построения моделей. В приложении приводится перечень команд трансформации и статистического анализа в SPSS, выполнение которых позволяет при наличии исходных данных получить расчетные показатели, а также таблицы с результатами моделирования. Синтаксис позволяет при необходимости быстро воспроизвести ход процесса моделирования, а также допускает легкую модификацию для построения аналогичных моделей на других данных, имеющих схожую структуру (либо на этих же данных, но по подгруппам респондентов). Дополнительно о синтаксисе SPSS можно прочитать на сайте www.spsstools.ru, или в руководстве пользователя по синтаксису (см. выше).

Внимание! Перед запуском синтаксиса необходимо определить пропущенные значения по всем переменным, чтобы они исключались из анализа и не искажали результатов расчета.

Вычисление показателей.

Вычисление числа членов семьи (это присутствует либо в переменной i1.o, либо в i1.n).

COMPUTE nfam=SUM(i1.o,i1.n).

Вычисление двоичной переменной «город» (если код 1 или 2, то это – город).

COMPUTE gorod=status<3.

Вычисление числа источников дохода (если код в этих переменных равен 1, значит респондент согласился, что у него есть такой источник дохода).

COMPUTE ndohod=SUM(0, if3=1, if6.1=1, if6.2=1,

if9.1a=1,

if9.2a=1,

if9.3a=1,

if9.4a=1,

if9.5a=1,

if9.6a=1,

if9.7a=1,

if9.8a=1,

if9.9a=1,

if9.91a=1,

if9.10a=1,

if10=1,

if11.3=1,

if12.1a=1,

if12.2a=1,

if12.3a=1,

if12.4a=1,

if12.5a=1,

if12.6aa=1,

if12.6ba=1,

if12.7a=1,

if12.8a=1,

if12.9a=1,

if1210ba=1).

Вычисление первого варианта индекса (наличие предмета (код=1), т.е. 1 мы умножаем на вес и на максимум из 0 или 10-«возраст». Если возраст предмета больше или равен 10 годам, мы берем не отрицательное значение, а 0, т.е. не учитываем данный предмет).

COMPUTE indexTDP=SUM( 0, (ic9.1a=1)\*0.04\*Max(0,10-ic9.1b),

(ic9.2a=1)\*0.04\*Max(0,10-ic9.2b),

(ic9.3a=1)\*0.04\*Max(0,10-ic9.3b),

(ic9.4a=1)\*0.01\*Max(0,10-ic9.4b),

(ic9.5a=1)\*0.03\*Max(0,10-ic9.5b),

(ic9.6a=1)\*0.03\*Max(0,10-ic9.6b),

(ic9.6.2a=1)\*0.04\*Max(0,10-ic9.6.2b),

(ic9.7a=1)\*0.1\*Max(0,10-ic9.7b),

(ic9.7.1a=1)\*0.1\*Max(0,10-ic9.7.1b),

(ic9.8a=1)\*0.05\*Max(0,10-ic9.8b),

(ic9.9a=1)\*0.1\*Max(0,10-ic9.9b),

(ic9.101a=1)\*0.2\*Max(0,10-ic9.101b),

(ic9.12a=1)\*1\*Max(0,10-ic9.12b)).

Вычисление второго варианта индекса ТДП. Тут мы не учитываем вес предмета.

COMPUTE indexTDP1=SUM( 0, (ic9.1a=1)\*Max(0,10-ic9.1b),

(ic9.2a=1)\*Max(0,10-ic9.2b),

(ic9.3a=1)\*Max(0,10-ic9.3b),

(ic9.4a=1)\*Max(0,10-ic9.4b),

(ic9.5a=1)\*Max(0,10-ic9.5b),

(ic9.6a=1)\*Max(0,10-ic9.6b),

(ic9.6.2a=1)\*Max(0,10-ic9.6.2b),

(ic9.7a=1)\*Max(0,10-ic9.7b),

(ic9.7.1a=1)\*Max(0,10-ic9.7.1b),

(ic9.8a=1)\*Max(0,10-ic9.8b),

(ic9.9a=1)\*Max(0,10-ic9.9b),

(ic9.101a=1)\*Max(0,10-ic9.101b),

(ic9.12a=1)\*Max(0,10-ic9.12b)).

Вычисление третьего варианта индекса ТДП. Здесь мы учитываем только наличие предметов.

COMPUTE indexTDP2=SUM( 0, (ic9.1a=1),

(ic9.2a=1),

(ic9.3a=1),

(ic9.4a=1),

(ic9.5a=1),

(ic9.6a=1),

(ic9.6.2a=1),

(ic9.7a=1),

(ic9.7.1a=1),

(ic9.8a=1),

(ic9.9a=1),

(ic9.101a=1),

(ic9.12a=1)).

Описательная статистика рассчитанных и присутствующих в базе данных показателей, которые будут использованы для моделирования:

FREQUENCIES

VARIABLES=nfam gorod ndohod

/ORDER=ANALYSIS .

GRAPH

/HISTOGRAM=if14 indexTDP1 indexTDP2 indexTDP3.

EXAMINE

VARIABLES=ndohod nfam if14 indexTDP1 indexTDP2 indexTDP3

/PLOT BOXPLOT STEMLEAF

/COMPARE GROUP

/PERCENTILES(5,10,25,50,75,90,95) HAVERAGE

/STATISTICS DESCRIPTIVES

/CINTERVAL 95

/MISSING LISTWISE

/NOTOTAL.

Построение 3-х блоков регрессионных моделей, по одному на каждый вариант индекса.

REGRESSION

/MISSING LISTWISE

/STATISTICS COEFF OUTS R ANOVA

/CRITERIA=PIN(.05) POUT(.10)

/NOORIGIN

/DEPENDENT indexTDP1

/METHOD=STEPWISE nfam gorod ndohod if14.

REGRESSION

/MISSING LISTWISE

/STATISTICS COEFF OUTS R ANOVA

/CRITERIA=PIN(.05) POUT(.10)

/NOORIGIN

/DEPENDENT indexTDP2

/METHOD=STEPWISE nfam gorod ndohod if14.

REGRESSION

/MISSING LISTWISE

/STATISTICS COEFF OUTS R ANOVA

/CRITERIA=PIN(.05) POUT(.10)

/NOORIGIN

/DEPENDENT indexTDP3

/METHOD=STEPWISE nfam gorod ndohod if14.