*Национальный Исследовательский Университет – Высшая Школа Экономики*

Соловьев Кирилл Игоревич (студент 2к магистратуры)

Под руководством: Демидовой Ольги Анатольевны, Кафедра математической экономики и эконометрики.

**Практикум по прикладным эконометрическим исследованиям.**

**Введение**

Дорогой друг!

Данный практикум был составлен с единственной целью – помочь в процессе освоения дисциплины «Эконометрика» в части проведения практических исследований. Эконометрический инструментарий используется в большинстве научных статей; наличие исследовательской части в курсовой работе, дипломе является почти необходимым (но не достаточным!) условием успешной сдачи, поэтому желательно не только понимать теоретические основы курса «Эконометрика», но и знать, как технически провести исследование на реальных данных. Надеемся, что практикум поможет – в нем показан пошаговый процесс построения моделей и анализа данных. Результатом данного исследования стал препринт ВШЭ[[1]](#footnote-1). В случае возникновения каких-либо теоретических вопросов (о характере, источниках данных или результатах конкретных шагов в исследовании), можно попробовать найти их в препринте.

**Основная часть**

Исследование проходит в программном пакете “Stata”[[2]](#footnote-2). Уже вышла версия 13, но в открытых источниках можно найти ранние версии – радикальных различий в них нет. В данном практикуме используется версия 11.1 Устанавливаем пакет, открываем.

Рисунок Интерфейс STATA 11.1



Загружаем данные: File – Open и выбираем файл «Datafile.dta» (он прилагается к практикуму, необходимо скачать и сохранить).

Формат файла специальный, если в будущем данные будут в другом формате (Excel, SPSS и др.), то можно воспользоваться программой «Stat/Transfer»[[3]](#footnote-3) для перевода из одного формата в другой.

В файле «Datafile.dta» содержатся данные исследования Аналитического центра Юрия Левады, проведенного в 2010 году по заказу Высшей школы экономики для Института анализа предприятий и рынков в рамках проекта по исследованию взаимоотношений бизнеса и органов государственной власти и роли российских ассоциаций бизнеса как института координации этих отношений. Были собраны данные по 1013 предприятиям из 61 субъекта Российской Федерации. Информация, собранная в исследовании, содержит характеристики предприятия, данные о собственности, управлении, участии в бизнес - ассоциациях, оценку бизнес среды в регионе и взаимоотношений с властями, информацию о получении помощи от государства, а также о получении поддержки от властей. Выборка формировалась как репрезентативная в следующих разрезах:

1. Виды экономической деятельности (на уровне двузначных или трехзначных кодов ОКВЭД). В секторе промышленности рассматриваются следующие отрасли: машиностроение и металлообработка, легкая, деревообрабатывающая, химическая и металлургическая и пищевая промышленность. В секторе услуг анализируются грузовые перевозки, туризм, IT-услуги и розничная торговля.

2. Размер предприятия по укрупненным группам (крупные, средние, малые).

3. Месторасположение по типу поселения (столичные города – Москва и Санкт-Петербург, областные центры или населенные пункты за их пределами).

Поскольку в работе исследуется влияние региональных характеристик, использована база политических и социально-экономических показателей регионов, сформированная Международным центром изучения институтов и развития (переменные уже включены в «Datafile.dta»).

В качестве зависимой переменной взята бинарная переменная, характеризующая получение поддержки от властей. Предприятиям предложено было ответить на вопрос «оказывалась ли поддержка предприятию со стороны государства» в периоды 2006-2008 гг. и 2009-2010 гг. При этом компании предоставили информацию в разрезах поддержки от власти разных уровней: предоставлялась ли помощь на федеральном (Fed), региональном (Reg), местном (Loc), а также на любом (то есть, хотя бы на одном) уровне (FRL) (*Переменные «FedHelp0608», «RegHelp0910» и т.д., где Reg, Fed, Loc – уровень власти, 0608 – период с 2006 по 2008 гг., 0910 – период с 2009 по 2010 гг.*)

Из множества доступных характеристик предприятий были выбраны следующие[[4]](#footnote-4):

1. Количество работников. Переменная взята в натуральных логарифмах, поскольку абсолютные значения показателя, также как и разброс значений велики, и это может помешать анализу остальных факторов, которые являются преимущественно дамми-переменными (*lnSize10*)

2. Год основания. Было выделено 3 группы предприятий:

a) Основанные до 1992 года – условно, советские предприятия.

b) Основанные в период с 1992 г. по 1998 г. (*foundation92\_98*)

c) Основанные после 1998г. (*foundation98*)

3. Характеристики владельцев предприятий

a) Наличие государства в числе собственников. (*State\_Owner*)

b) Отказ от раскрытия собственников. (*Ownership\_no\_answer*)

c) Участие иностранцев в собственности. (*Foreign\_Stock*)

4. Участие в бизнес – ассоциации. (*BA*)

5. Контрольные переменные для типа населенного пункта. Проводится разделение между периферийными городами и центрами (областными/ краевыми/ республиканскими). В отдельную категорию выделена Москва. (*Moscow, PGT, Center*)

6. Участие (вхождение) в холдинг или группу компаний. (*Holding, Holdinghead)*

7. Оказание помощи государству в социальном развитии региона. (*FirmHelp, FirmHelpSuf)*

Первым делом вводим команду *set memory 50m* – эта команда увеличивает объем памяти для данных до 50 Мб (объем можно выбрать любой). Команда избавит от проблем при открытии файла, так как иногда загружаемый файл настолько большой, что объема памяти «по умолчанию» не хватает.

Затем рекомендуется ввести команду *set more off* – эта команда разрешает Стате показывать все результаты сразу (по умолчанию Стата показывает только ту часть, которая помещается на экран).

Перед построением моделей проводится статистический анализ данных, который позволяет понять характеристики выборки. Команда *sum variable* дает основные статистические данные о переменной. Например, *sum Moscow* выдает следующую таблицу:

******

Из таблицы видно, что это бинарная переменная, принимает значения от 0 до 1, у переменной есть 1013 наблюдения и в среднем в выборке 9,48% наблюдений (т.е. предприятий) относятся к Москве. Можно написать несколько переменных, и Стата выдаст результаты сразу для всех переменных или можно ввести, например, команду sum lnSize\*. Звездочка является заменой всех символов – Стата выдаст все переменные, название которых начинается с lnSize (прием со звездочкой работает и для остальных команд).



Если необходимо сузить описываемую выборку, то необходимо воспользоваться командой *if.* Например, *sum lnSize08 if Moscow==1 –* статистика будет дана только по тем предприятиям, которые находятся в Москве.



Обратите внимание на двойной знак равенства. Команда работает только так. **Важно!** Стата очень чувствительна к регистру и пунктуации, поэтому если какая-либо команда не работает, то необходимо проверить пунктуацию.

Для того, чтобы проверить независимость распределений двух переменных, используется команда *tab var1 var2, chi.* Например:

**

Нулевая гипотеза – переменные распределены независимо, и в данном случае гипотеза не подтверждается.

Для построения моделей бинарного выбора (в которых зависимая переменная принимает значения 0 и 1) используются команды *logit var1 .. varN* и *probit var1 .. varN.* Синтаксис команды и результат похож на команду для обычной линейной регрессии *reg var1 .. varN*

Например:



Стоит обратить внимание на сноску “sector10 omitted because of collinearity” – в регрессии 10 бинарных переменных (по принадлежности к сектору) и во избежание проблем с коллинеарностью и обратимостью матрицы оценок коэффициентов, одна из переменных взята за базу и опущена. В противном случае, у нас одна из переменных выражалась бы через другие, что являлось бы типичной коллинеарностью.

Регрессия строится для того, чтобы посмотреть значимость и знак коэффициентов. Для получения этих результатов в Стате можно сохранить данные (*est store name*) и вывести таблицу с уровнями значимости (*est tab, star*). *Star* и означает вывод на экран уровней значимости. Например:



Можно вывести для удобства в таблице сразу несколько результатов регрессий (*est tab name1 name2, star*). Дальше (с точки зрения исследования) анализируется значимость коэффициентов, знаки коэффициентов и делаются выводы о подтверждении (или его отсутствии) гипотез или, что бывает чаще, проводится корректировка модели – добавляются или убираются переменные, берутся логарифмы, квадраты и т.д. В этом процессе и заключается суть исследования.

В ходе конкретного исследования, о котором шла речь во введении, возникла ситуация, когда причинно–следственная связь между результирующей переменной и фактором(-ами) неоднозначна: как фактор может влиять на результирующую переменную, так и результирующая переменная на фактор. Поскольку оценка логит-модели с эндогенностью дает смещенные оценки коэффициентов использовалась оценка **систем двумерных пробит моделей** следующего вида:

$y\_{1}^{\*}=ay\_{2}+x\_{1}B\_{1}+e\_{1}$

$$y\_{2}^{\*}=x\_{2}B\_{2}+e\_{2}$$

$\left(e\_{1, }e\_{2}, \right|x\_{1}, x\_{2})\~N(0,0,1,1,p)$, где

yi=1 если yi\* >0 и yi=0 в противном случае, i=1,2

p – коэффициент корреляции ошибок e1 и e2.

Для качественной идентифицируемости параметров модели, векторы объясняющих переменных x1 и x2 не должны быть тождественны, во втором векторе должны быть уникальные факторы, не встречающиеся в первом. Команда в Стате: *mvprobit (Var = var1 var2) (var1 = var2 var3*), где Var – результирующая переменная, var1 – переменная(-ые), которая дает проблему эндогенности, var2 – независимые переменные, var3 – уникальная переменная(-ые). В исследовании, например, ожидалось, что на вероятность получения помощи от государства влияет помощь бизнеса государству:



В данном примере независимыми переменными были:

1. Наличие иностранных собственников

2. Административный статус населенного пункта

3. Участие в холдинге

Для того чтобы понять, какие именно переменные следует считать независимыми определяются коэффициенты корреляции Спирмена: *spearman var1 … varN, star(0.1).*

**Заключение**

В данном практикуме была сделана попытка рассказать техническую сторону проведения эконометрических исследований, и даны основные полезные команды статистической программы STATA на примере конкретного исследования, опубликованного в серии препринтов НИУ-ВШЭ. Надеемся, что читатель, независимо от опыта пользования Статой, смог почерпнуть для себя что-то полезное.

В случае если у Вас будут какие-либо комментарии, пожелания и т.д., с автором можно связаться по почте: kirill.solovyev@gmail.com

1. <http://www.hse.ru/data/2012/11/30/1302222458/WP1_2012_04_final.pdf> [↑](#footnote-ref-1)
2. <http://www.stata.com/> [↑](#footnote-ref-2)
3. <http://www.stattransfer.com/> [↑](#footnote-ref-3)
4. Выбор переменных и построение модели являются самыми важными частями исследования, но в данном Практикуме акцент сделан преимущественно на технической части, гипотезам и моделям посвящено не так много времени. [↑](#footnote-ref-4)