Азы регрессионного анализа в социальных и демографических исследованиях для «чайников»

Синявская Оксана Вячеславовна
osinyavskaya@hse.ru

Общие вопросы

»

Зачем нам нужен регрессионный анализ?

- Доказательство взаимосвязи факторов «при прочих равных» (очищенное от других эффектов)
- Причинно-следственная связь
- Количественные оценки силы влияния
- Предсказание и прогноз

Регрессия – не панацея

- Любая статистическая модель – аппроксимация реальности
- Главное – теория в основе модели
- Модель – инструмент поиска ответа на исследовательский вопрос
  - Упрощение неизбежно
- Данные не идеальны
  - Компромисс между теорией, требованиями статистики и реальными данными
  - Недостаточное число наблюдений… мало значимых эффектов
- Неадекватно подобранная модель – ошибочные результаты
- Много моделей лучше, чем одна
- Нет единственно верного способа решения поставленной задачи
- Интерпретация результатов!

ОЧЕНЬ беглый список регрессионных моделей

- По типу зависимой переменной
  - Линейная регрессионная модель
  - Модели для категориальных данных
    - Бинарные (логит-, пробит-), порядковые (ordered/ordinal - логистическая, пробит), мультиномиальные, для счетных / дисретных (count) данных
  - Модели для цензурированных данных
    - Тобит-
    - Модели выживаемости / анализ наступления событий (event history, duration, survival)
- По типу данных
  - Модели для пространственных данных
  - Модели временных рядов
  - Модели для панельных данных
  - Иерархические модели
- Системы регрессионных уравнений

Использование моделей в эпоху продвинутых статистических пакетов

- Можно не знать формулы и не перепроверять расчеты, НО ВАЖНО ЗНАТЬ:
  - Базовые допущения относительно
    - Распределения зависимой переменной
    - Взаимосвязи между зависимой и независимыми переменными
    - Характера распределения случайных ошибок
  - Процедуры оценки качества модели и данных
  - Что обозначают коэффициенты, и как их интерпретировать
    - Коэффициенты, стандартные ошибки, t / z-статистики, значимость
      - Коэффициенты значимо отличаются от нуля?
      - Переменная существенно улучшает модель?
    - То, что получилось, вообще адекватно?!

Использование моделей в эпоху продвинутых статистических пакетов

- Можно не знать формулы и не перепроверять расчеты, НО ВАЖНО :
  - Не забывать про интерпретацию – в идеале – получить **понятные** оценки
    - Пр.: *увеличение стажа на 1 год дает прирост зарплаты в рублях…, увеличение возраста 1-го ребенка на 12 мес. приводит к тому, что 2-й ребенок рождается на … мес. быстрее; развод родителей увеличивает вероятность развода респондента на …% и т.п.*
  - Корреляция или причинно-следственная связь?
- Важен предварительный анализ данных
  - Простые и кросс-распределения, корреляции
  - Миссинги

Использование моделей в эпоху продвинутых статистических пакетов

- Много моделей лучше, чем одна
- Несколько спецификаций лучше, чем одна
  - Риски невключения переменных, или включения ненужных переменных
  - Теория в основе выбора переменных!
  - Тщательный отбор переменных
- Но в итоге оставляем одну, наилучшую модель
- Не уверен – выбирай простейшую модель
  - Потом можно усложнить
  - Про новые модели нужно вначале читать и смотреть примеры
- Времени мало, а память ненадежна – **синтаксис!**!!
  - В идеале – еще и log-файл (журнал), чтобы помнить, что проверяли, и что получили

Наиболее важное предположение регрессионного анализа?

- Данные имеют нормальное распределение
- Все ключевые независимые переменные включены в модель
- Все переменные измерены полностью и без ошибок
- Наблюдения независимы

Часто наблюдения оказываются связанными…

- Определенный дизайн проекта
  - Исследования в сфере образования: в выборку попадает несколько учащихся одного учебного заведения
  - Исследования в сфере здоровья: измерение динамики роста, веса и пр. у одних и тех же людей
  - Обследования домохозяйств (ОБДХ, РМЭЗ): ответы респондентов из одного домохозяйства будут похожи друг на друга больше, чем ответы респондентов из других домохозяйств
  - Панельные исследования: один и тот же респондент / домохозяйство опрашивается несколько раз
  - Стратифицированная выборка: ответы жителей одного населенного пункта / района будут более похожи…

- Некоторые взаимосвязи между индивидами существуют практически во всех социальных исследованиях (иерархические кластеры)

Что делать со связанными переменными..?

- Игнорировать этот факт?
  - Риск: ошибочно низкие стандартные ошибки, смещенные коэффициенты

- Корректировать стандартные ошибки
  - В статистических пакетах есть несколько процедур учета кластеров:
    - Stata: для всех типов регрессий: Variance estimators (vce) – default = oim (observed information matrix); Sandwich estimators: robust (Huber/White/sandwich estimator) или cluster [пр.: cluster(ID)]
    - SPSS: Robust estimator в Generalized Linear Model / Estimation
- Ввести дамми-переменные для каждой группы
- Добавить переменные более высокого уровня в модель (уровень безработицы, охвата ДОУ, и т.п.)

- Строить иерархическую модель

Иерархические модели

- Multilevel models

Семинар НУГ 18-апр-2014

Что такое иерархический регрессионный анализ (multilevel modelling)?

- Wikipedia (April 16, 2014):
  - **Multilevel models** are statistical models of parameters that vary at more than one level. These models can be seen as generalizations of linear models (in particular, linear regression), although they can also extend to non-linear models.

  - Иерархическая структура данных, данные сгруппированы
    - Развивается в основном в рамках социологии, исследований образования, медицинских, биостатистики
      - Экономисты – модели с фиксированными эффектами
    - В социальных науках: индивид – сообщество
    - Выборка: многоступенчатая, стратифицированная, кластерная

«Георгий Иванович, он же Гоша, он же Гога, он же Жора…»

- Иерархический регрессионный анализ
- Обобщенная линейная модель множественной (смешанной) регрессии

- Multi-level model
- Generalized Linear Mixed Models (GLMM)
- Hierarchical (linear) model
- Random coefficient (parameter, effects) model
- Variance component model
- Mixed (linear, logit…) model
- Mixed effects model
- Nested models
- Contextual analysis

…Хорошо забытое старое

- Новый метод?
- Расширение обобщенной линейной модели
- Усовершенствование дисперсионного анализа ANOVA (Analysis of variance and covariance), Repeated measures ANOVA, MANOVA…

Типы связанных данных

1. Индивиды внутри групп
   - Данные внутри группы будут иметь больше общего, чем данные случайно отобранных индивидов
   - Групповая динамика будет влиять на индивидов (пр.: ситуация на локальном рынке труда)
2. Повторяющиеся измерения одних и тех же индивидов
   - Фокус – изменение / рост
3. Повторяющиеся характеристики одних и тех же индивидов
   - Фокус – взаимосвязь переменных внутри одного индивида

Место иерархического анализа в традиционном анализе

Анализ микроданных (индивидуальный уровень, игнорируем группы)	Нарушение предположения о независимости наблюдений --> уменьшение стандартных ошибок
Иерархический анализ	- 2 и более уровней могут анализироваться одновременно - можно анализировать внутри- и межгрупповую вариацию
Анализ агрегированных данных (игнорируем индивидов)	Смысл переменных на индивидуальном уровне может отличаться от смысла переменных на уровне группы (пр.: уровень доходов семьи и населения города N, индивидуальный и региональный уровень безработицы, рождение детей в семье и на уровне города…)

Уровни. Типы зависимых переменных

- Уровни
  - Обычно 2
  - Реже 3
  - Больше – практически нет
- Типы (шкалы) зависимых переменных
  - Количественные → линейные модели
  - Бинарные → логистические
  - Категориальные (3+ категорий) / дисретные
  - Время до наступления события → event history

Основные компоненты иерархических моделей

- Спецификация независимых переменных с нескольких уровней (Фиксированные эффекты)
  - Какие переменные включать?
  - Ключевые взаимодействия (interactions)
- Спецификация корреляции между ответами из одного кластера (Случайные эффекты)
- Выбор должен быть основан на теории, исследовательском вопросе и эмпирических свидетельствах

Программное обеспечение

- Часть больших статистических пакетов
  - Stata: Multi-level fixed effects models + внутри типов регрессий, напр.: Binary outcomes / Multilevel logit (melogit), xtlogit…
  - SPSS (Mixed models – больше вариантов моделей, начиная с версии 19; но все равно пока в основном линейные)
  - SAS
  - R
- Специальные пакеты
  - HLM, AlM, MLA, MLwiN, etc.

Пример: многоуровневая регрессия в Stata

Пример синтаксиса:

```
ologit Rint3ord fagem fagem2 feduc_2 feduc_3 feduc_4                    ///
        nsibl0 nsibl2 nsibl3 chagey1m chagey1m2 union urban        ///
        lowrrarea_2 lowrrarea_3        yr2007 yr2011 if nomiss14f==1,        ///
        cluster(id2) nolog or
meologit Rint3ord fagem fagem2 feduc_2 feduc_3 feduc_4                  ///
        nsibl0 nsibl2 nsibl3 chagey1m chagey1m2 union urban        ///
        lowrrarea_2 lowrrarea_3        yr2007 yr2011 if nomiss14f==1 || id2:,   ///
        nolog or
meologit Rint3ord fagem fagem2 feduc_2 feduc_3 feduc_4                  ///
        nsibl0 nsibl2 nsibl3 chagey1m chagey1m2 union urban        ///
        lowrrarea_2 lowrrarea_3        yr2007 yr2011 if nomiss14f==1 || area: || id2:, ///
        nolog or
```

Пример: многоуровневая регрессия в Stata

	cluster (id2)				2-level; groups: id2				3-level: area <-- id2			
Rint3ord	OR	S.E.	z	P>z	OR	S.E.	z	P>z	OR	S.E.	z	P>z
fagem	0.95	0.01	-3.9	0.000	0.93	0.02	-3.8	0.000	0.93	0.02	-3.9	0.000
fagem2	0.99	0.00	-5.7	0.000	0.99	0.00	-5.5	0.000	0.99	0.00	-5.5	0.000
feduc_2	1.06	0.28	0.2	0.832	1.01	0.32	0.0	0.988	1.02	0.33	0.1	0.939
feduc_3	1.03	0.27	0.1	0.896	1.06	0.33	0.2	0.842	1.09	0.34	0.3	0.788
feduc_4	1.22	0.32	0.8	0.446	1.34	0.43	0.9	0.367	1.35	0.43	0.9	0.352
nsibl0	0.92	0.12	-0.7	0.496	0.94	0.17	-0.4	0.723	0.93	0.16	-0.4	0.683
nsibl2	0.93	0.12	-0.6	0.554	0.88	0.15	-0.7	0.457	0.87	0.15	-0.8	0.404
nsibl3	1.04	0.16	0.3	0.797	1.03	0.23	0.1	0.885	1.04	0.23	0.2	0.864
chagey1m	0.99	0.02	-0.7	0.508	0.99	0.02	-0.6	0.529	0.99	0.02	-0.6	0.542
chagey1m2	0.99	0.00	-2.6	0.009	0.99	0.00	-2.7	0.007	0.99	0.00	-2.7	0.007
union	1.86	0.19	6.2	0.000	2.27	0.33	5.6	0.000	2.24	0.33	5.5	0.000
urban	1.18	0.14	1.4	0.169	1.40	0.23	2.0	0.044	1.38	0.24	1.8	0.068
lowrrarea_2	1.25	0.20	1.4	0.155	1.33	0.27	1.4	0.158	1.25	0.36	0.8	0.430
lowrrarea_3	0.71	0.19	-1.3	0.201	0.61	0.23	-1.3	0.193	0.60	0.28	-1.1	0.271
yr2007	1.10	0.10	1.1	0.294	1.22	0.16	1.6	0.113	1.22	0.15	1.6	0.115
yr2011	1.37	0.14	3.0	0.003	1.73	0.25	3.8	0.000	1.74	0.25	3.9	0.000
/cut1	-0.44	0.30			-0.52	0.38	-1.4	0.167	-0.55	0.38	-1.4	0.152
/cut2	0.90	0.31			1.36	0.39	3.5	0.000	1.34	0.39	3.4	0.001
/cut3	2.77	0.32			3.87	0.42	9.1	0.000	3.84	0.43	9.0	0.000
									area			
var(_cons)									0.07	0.06		
					id2				area>id2			