Дисперсионный анализ

А.Р.Бессуднов
(bessudnov@gmail.com)

НИУ ВШЭ, факультет социологии

25 мая 2012

Дисперсионный анализ

- Analysis of variance (ANOVA)
- Метод изобретен Рональдом Фишером в 1920-е гг.
- Применяется, когда зависимая переменная является интервальной, а независимая (или несколько независимых) – категориальной (фактором)
- Задачей является установить, имеются ли статистически значимые различия в средних значениях переменной в группах, образованных одной или несколькими категориальными переменными
- Статистически эквивалентен регрессионному анализу с фиктивными переменными (который в большинстве случаев и применяется для решения этой задачи)
- Тем не менее понимание дисперсионного анализа важно как для более полного понимания регрессии, так и для других методов (таких как многоуровневое моделирование)

Пример: политическая идеология и партийная поддержка

- Предположим, мы хотим установить наличие зависимости между партийной поддержкой (демократы, независимые, республиканцы) и политической идеологией (измеренной по шкале от 1 “очень либеральная” до 7 “очень консервативная”)

партия	1	2	3	4	5	6	7	n	средняя	ст.откл.
демократы	9	20	17	36	4	5	0	91	3.23	1.28
независимые	7	11	17	48	12	11	5	111	3.9	1.43
консерваторы	0	2	7	23	23	17	2	74	4.7	1.1

- Источник: A.Agresti, B.Finlay, Statistical Methods for the Social Sciences

Варианты анализа

- Если мы рассматриваем обе переменные как категориальные, можно воспользоваться критерием χ^2
- Если рассматривать политическую идеологию как интервальную переменную и рассчитать средние по группам, мы можем выяснить, является ли разница между средними статистически значимой а) с помощью дисперсионного анализа или, эквивалентно, б) с помощью регрессии с фиктивными переменными

Принцип дисперсионного анализа

- Дисперсионный анализ сравнивает дисперсию зависимой переменной внутри групп и между групп, образованных фактором (-ами), и на основе этого сравнения делает вывод о значимости разницы между средними
- Чем больше дисперсия между группами по сравнению с дисперсией внутри групп, тем больше оснований считать, что между средними по группам есть значимая разница

Допущения дисперсионного анализа

- В каждой группе распределение зависимой переменной нормально (в генеральной совокупности)
- Стандартное отклонение зависимой переменной в каждой группе одинаково (в генеральной совокупности)
- Выборка является случайной и наблюдения в ней независимы друг от друга
- Как видим, эти допущения совпадают с допущениями МНК-регрессии

Оценка дисперсии внутри групп

- Для группы i, сумма квадратов отклонений наблюдений от среднего значения в группе равна $\sum(y - \bar{y}_i)^2$
- Сумма этих значений для каждой группы называется суммой квадратов внутри групп
- Число степеней свободы df для оценки дисперсии в группе i равно $n_i - 1$
- Дисперсия в этой группе равна $s_i^2 = \frac{\sum(y-\bar{y}_i)^2}{n_i-1}$
- Для все групп сразу дисперсия внутри групп равна $s^2 = \frac{\sum(n_i-1)s_i^2}{N-g}$, где N это общее число наблюдений, а g – число групп

Оценка дисперсии между группами

- Дисперсия между группами оценивается через разницу между средними по группам и общей средней в выборке
- Дисперсия между группами равна сумме квадратов между группами, деленной на число степеней свободы: $\frac{\sum n_i(\bar{y}_i - \bar{y})^2}{g-1}$, где g – это число групп
- Общая сумма квадратов TSS равна сумме квадратов внутри групп + сумме квадратов между группами:
  $TSS = \sum(y - \bar{y})^2 = \text{between-groups SS} + \text{within-groups SS}$
- Аналогично, общая дисперсия равна сумме дисперсии внутри групп и дисперсии между группами

Является ли разница между средними в группах случайной?

- Нулевая гипотеза в данном случае заключается в том, что средние во всех группах равны: $H_0 : \mu_1 = \mu_2 = ... = \mu_g$
- Альтернативная гипотеза: средние хотя бы в двух группах отличаются
- Для тестирования нулевой гипотезы используется F-тест
- $F = \frac{\text{оценка дисперсии между группами}}{\text{оценка дисперсии внутри групп}}$
- В случае если нулевая гипотеза верна, дисперсия между группами не больше дисперсии внутри групп, и $F \leq 1$
- Если F значительно больше 1, то между группами с большой вероятностью существует разница в средних
- Значимость F проверяется по F-распределению

Пример: партийная поддержка и политическая идеология

	сумма квадратов	df	квадрат ср.откл.	F	p
между партиями	88.43	2	44.21	26.3	<0.001
внутри партий	459.52	273	1.68		
общая	547.95	275			

t-тест и дисперсионный анализ

- Почему нельзя сравнить группы попарно с помощью t-теста?
- Если сравнений много (число попарных сравнений равно $g(g-1)/2$), на 95%-м уровне значимости можно ожидать статистически значимые отличия между средними в группах там, где они на самом деле отсутствуют
- Существуют методы коррекции стандартных ошибок при множественных сравнениях, которые дают более консервативную оценку (метод Бонферрони, метод Таки)

Регрессионная оценка

	коэф.	ст.ош.	p
константа	3.23	0.14	<0.001
Партии (баз. демократы)			
независимые	0.67	0.18	<0.001
республиканцы	1.47	0.20	<0.001
n	276		
R^2	0.16		

- идеология = 3.23 + 0.67*нез. + 1.47*респ.
- Статистические программы в регрессионном выводе также выдают таблицу результатов дисперсионного анализа (ANOVA table)

Двухфакторный дисперсионный анализ

- Используется, когда категориальных предикторов (факторов) два или более (многофакторный дисперсионный анализ)
- В этом случае модель может включать как основные эффекты, так и эффекты взаимодействия между факторами

Пример: зависимость политической идеологии от партийной поддержки и пола (без эффекта взаимодействия)

	сумма квадратов	df	квадрат ср.откл.	F	p
модель	86.7	3	28.9	17.3	<0.001
остатки	1569.6	939	1.7		
общая	1656.2	942			
Источник					
партия	84.3	2	42.1	25.2	0.001
пол	1.3	1	1.3	0.8	0.38

Зависимость политической идеологии от партийной поддержки и пола (с эффектом взаимодействия)

	сумма квадратов	df	квадрат ср.откл.	F	p
модель	90.3	5	18.1	10.8	<0.001
остатки	1565.9	937	1.7		
общая	1656.2	942			
Источник					
партия	87.8	2	43.9	26.3	0.001
пол	1.5	1	1.5	0.9	0.35
партия*пол	3.6	2	1.8	1.1	0.34

- Как и в регрессионных моделях, при наличии в модели эффектов взаимодействия основные эффекты не могут интерпретироваться непосредственно

Регрессионная модель без эффектов взаимодействия

	коэф.	ст.ош.	p
константа	3.8	0.08	<0.001
Партии (баз. демократы)			
независимые	0.17	0.1	0.1
республиканцы	0.71	0.1	<0.001
пол (баз. женщины)			
мужчины	0.08	0.09	0.38
n	943		
R^2	0.05		

- идеология = 3.8 + 0.17*нез. + 0.71*респ. + 0.08*муж.

Регрессионная модель с эффектами взаимодействия

	коэф.	ст.ош.	p
константа	3.8	0.09	<0.001
Партии (баз. демократы)			
независимые	0.11	0.13	0.42
республиканцы	0.59	0.13	<0.001
пол (баз. женщины)			
мужчины	-0.08	0.15	0.59
Эффекты взаимодействия			
мужчины*независимые	0.17	0.21	0.42
мужчины*республиканцы	0.31	0.21	0.14
n	943		
R^2	0.05		

- идеология = 3.8 + 0.11*нез. + 0.59*респ. - 0.08*муж. + 0.17*муж.*нез. + 0.31*муж.*респ.

Результаты дисперсионного анализа удобно представлять графически

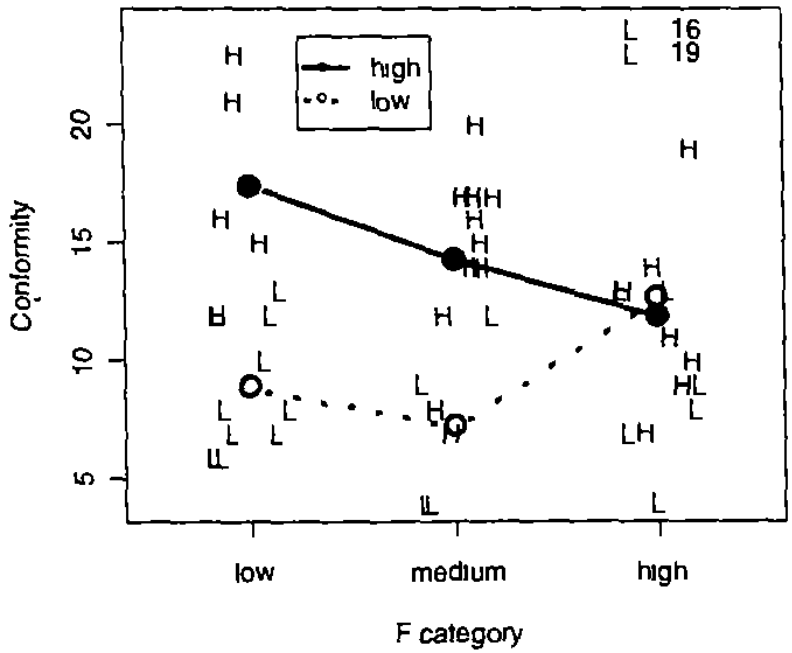

Figure 4.2 Conformity by partner's status and authoritarianism for Moore and Krupat's experiment. The points are jittered horizontally to avoid overplotting.

Источник: J.Fox, An R and S-Plus Companion to Applied Regression

Дисперсионный анализ с зависимыми наблюдениями

- В некоторых случаях наблюдения в выборке не являются независимыми
- Например, одни и те же люди отвечают на один и тот же вопрос несколько раз в разное время (повторяющиеся измерения)
- Другой пример: оценки учеников из нескольких школ (оценки учеников из одной школы могут быть более схожими)
- В общем случае это относится к данным с многоуровневой структурой
- В этом случае однофакторный дисперсионный анализ по сути становится двухфакторным (вторым фактором являются отдельные респонденты, школы, etc.)
- В регрессионном контексте это называется моделью с фиксированными эффектами: $E(y) = \alpha + \beta_j + \gamma_i$

Нарушение допущений дисперсионного анализа

- Допущение о том, что выборка является случайной и наблюдения независимы, является ключевым
- Нарушение допущения о том, что зависимая переменная в группах распределена нормально с одинаковым стандартным отклонением, не ведет к серьезным искажениям результатов
- Если отклонение от нормальности значительно, можно использовать непараметрический дисперсионный анализ Краскала-Уоллиса

Заключение

- Дисперсионный анализ – важный статистический метод. Необходимо понимать его принципы
- Но в практических задачах по анализу разницы средних в группах в большинстве случаев имеет смысл пользоваться регрессиоей с фиктивными переменными