УДК 303.72

ШУМЕТОВ В.Г.,
ЛЯСКОВСКАЯ О.В.

Опыт применения методов дисперсионного анализа в прикладных социологических исследованиях

Статья посвящена проблеме применения методов дисперсионного анализа данных в практике прикладных социологических исследований.

Ключевые слова: прикладные социологические исследования, дисперсионный анализ данных.

В нашей предыдущей публикации[1] обсуждалась проблема применения методов количественного анализа данных в практике прикладных социологических исследований; отмечалось, что объективно эта проблема обусловлена низким уровнем шкал измерения социологических данных – в шкалах наименований и в лучшем случае в ранговых шкалах[2]. Напомним, что номинальная шкала (шкала наименований, классификационная шкала) является наиболее «слабой» качественной шкалой, по которой объектам дается некоторый признак. Этот тип шкал соответствует простейшему виду измерения, при котором шкальные значения используются лишь как имена объектов. Более информативной является ранговая шкала (шкала порядка), когда допускается не только номинальное различение объектов, но и их упорядочение по измеряемым свойствам. Отмечалось также, что наиболее информативными являются абсолютные шкалы, которые характеризуются единственностью измерения и применяются, в частности, для измерения вероятности событий. Применительно к социологическим данным таковыми являются частости различных вариантов ответов респондентов на вопросы анкет массовых опросов, измеряемые в процентах от общего числа ответов и интерпретируемые как вероятности отношения изучаемых групп населения к исследуемым социальным явлениям или процессам.

Техника перехода от частот различных вариантов ответов респондентов на вопросы анкет массовых опросов нами была продемонстрирована и на примере разработки модели здоровья населения Орловского региона. В качестве эмпирической базы моделирования при этом использовались результаты социологических исследований «Изучение отношения населения Орловской области к своему здоровью», которые охватывали население города Орла и Орловской области, а в качестве инструмента количественного представления данных и моделирования – пакет анализа данных общественных наук SPSS Base. Общий объем выборки составил 793 респондента, из них 492 – проживающих в г. Орле, 301 – в Орловской области. При моделировании факторов здоровья особое внимание уделялось гендерному аспекту данной проблемы, и в этой связи модели разрабатывались отдельно для мужского и женского населения Орловского региона.

В контексте данного исследования речь может идти о количественных моделях следующего вида:

1) модели дисперсионного анализа, которые отражают связь вероятности хорошего здоровья населения с уровнями номинальных статусных признаков;

2) регрессионные модели, отражающие связь вероятности хорошего здоровья населения с иными количественными факторами в виде уравнения множественной линейной регрессии;

3) нейросетевые модели, отражающие связь вероятности хорошего здоровья населения со значениями статусных признаков в виде нейросети.

В данной публикации мы осветим вопросы разработки наиболее простых из них – моделей дисперсионного анализа – на примере модели, отражающей связь между количественной переменной – вероятностью хорошего здоровья мужского населения Орловского региона – со значениями двух статусных признаков, измеренными по номинальным шкалам, – групп возраста в интервале от 18–30 лет до 51–60 лет и групп уровня доходов в интервале от категории «До 2000 руб. в месяц» до категории «Более 5000 руб. в месяц». С учетом требований к надежности статистических выводов число групп обоих признаков было уменьшено – вместо восьми групп возраста рассматривается пять групп (от группы «18–30 лет» до группы «51–60 лет»), вместо шести уровней доходов – пять групп (от категории «До 2000 руб. в месяц» до категории «Более 5000 руб. в месяц»). Исходные данные для построения модели приведены в работе[3].

Продемонстрируем технологию построения модели дисперсионного анализа с помощью процедуры «General Linear Model» (общая линейная модель) пакета анализа данных общественных наук SPSS Base[4]. Предполагается, что модель может быть представлена в виде линейной функции вида

$$Y_{ij} = \mu_0 + \alpha_i + \beta_j + \varepsilon_{ij}, \qquad (1)$$

где Y_{ij} – наблюдаемое значение выходной переменной Y (вероятности хорошего здоровья мужского населения) на i-м уровне одного (уровень дохода) и j-м уровне другого (группа возраста) фактора; μ_0 – оценка свободного коэффициента модели; α_i и β_j – оценки главных эффектов на i-м и j-м уровнях; ε_{ij} – случайная ошибка.

Полагаем, что статусные признаки – фиксированные, при этом их взаимодействие в модель не включаем с тем, чтобы иметь возможность использовать дисперсию, отвечающую взаимодействию признаков, для проверки статистической значимости их влияния на результативную переменную, а также статистической значимости параметров модели[5]. Последовательность построения модели дисперсионного анализа такова:

1) предварительно по эмпирической базе, полученной по технологии, приведенной в работе[6], строятся графики, иллюстрирующие связь результативной количественной переменной – вероятности хорошего здоровья мужского населения Орловского региона – с номинальными статусными признаками – градациями уровня доходов и возраста, формулируются выводы о характере связи;

2) средствами процедуры «General Linear Model» пакета анализа данных общественных наук SPSS Base производится построение общей линейной модели дисперсионного анализа, описывающей связь вероятности хорошего здоровья мужского населения Орловского региона с градациями уровня доходов и возраста;

3) выполняется интерпретация результатов дисперсионного анализа (оценка межобъектных эффектов модели вероятности хорошего здоровья мужского населения Орловского региона);

4) выполняется интерпретация параметров общей линейной модели вероятности хорошего здоровья мужского населения Орловского региона;

5) строятся графики, отражающие связь расчетных значений вероятности хорошего здоровья мужского населения Орловского региона с изучаемыми признаками, их анализ;

6) строятся графики, отражающие качество расчетов значений вероятности хорошего здоровья мужского населения Орловского региона по линейной модели, формулируются выводы;

7) строятся графики, отражающие адекватность линейной модели вероятности хорошего здоровья мужского населения Орловского региона, формулируются выводы;

8) формулируются общие выводы по результатам моделирования.

Ниже приводятся результаты реализации каждого из перечисленных этапов моделирования.

Первый этап моделирования – визуальный анализ связи вероятности хорошего здоровья мужского населения Орловского региона с группами уровня доходов и возраста (рис. 1, 2).

Из характера графиков на рис. 1 и 2 видно, что вероятность хорошего здоровья мужского населения Орловского региона с ростом доходов увеличивается, а

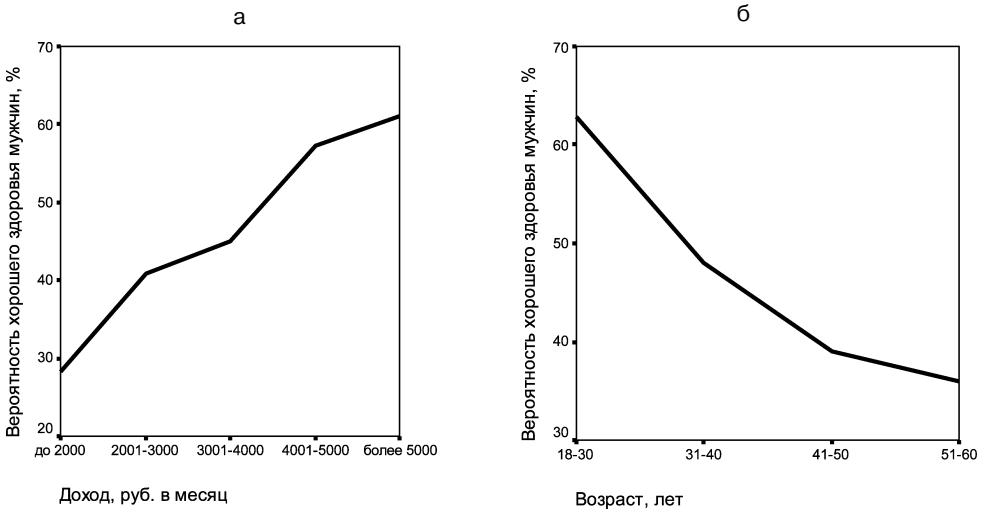


Рис. 1. Зависимость вероятности хорошего здоровья мужского населения Орловского региона от статусных факторов: *а* – градаций уровня дохода; *б* – градаций возраста (результаты опроса; средние значения)

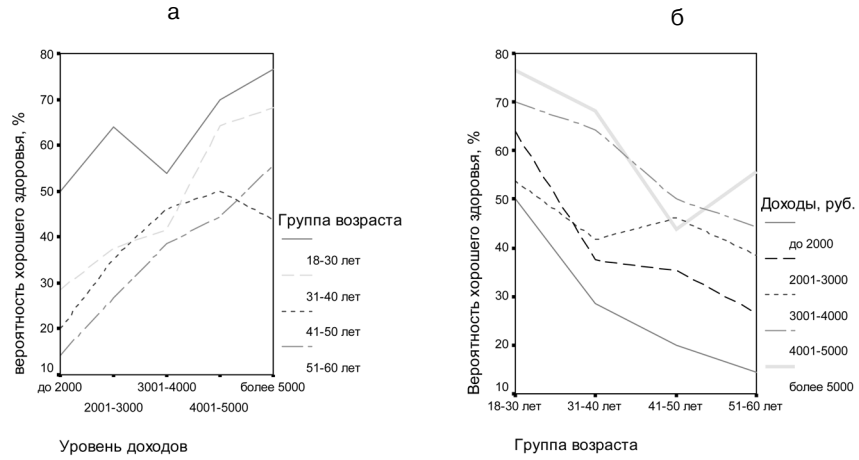


Рис. 2. Зависимость вероятности хорошего здоровья мужского населения Орловского региона от статусных факторов: *а* – градаций уровня дохода для разных возрастных групп; *б* – градаций уровня возраста для разных групп по уровню дохода (результаты опроса)

с возрастом – уменьшается. Эти зависимости – нелинейные, предположительно носят гиперболический характер. В то же время из рис. 2 следует, что эмпирические данные существенно «зашумлены» из-за наложения на средние тенденции случайной компоненты. Это подтверждает нецелесообразность построения более сложной модели, чем линейная.

Результаты моделирования представлены в табл. 1 и 2.

Таблица 1 – Таблица дисперсионного анализа (критерии межобъектных эффектов модели здоровья мужского населения Орловского региона)

Источник изменчивости	Сумма квадратов	Степени свободы *df*	Средний квадрат	Критерий Фишера F	Уровень значимости
Corrected Model	4959,548	7	708,507	16,867	0,000
Intercept	43189,218	1	43189,2-18	1028,164	0,000
ДОХОД	2770,582	4	692,645	16,489	0,000
ВОЗРАСТ	2188,966	3	729,655	17,370	0,000
Ошибка	504,074	12	42,006		
Всего	48652,840	20			
Corrected Total	5463,622	19			

Из табл. 1 следует, что вероятность хорошего здоровья мужского населения Орловского региона определяется примерно в равной степени как возрастом – критерий Фишера для этого фактора (отношение среднего квадрата фактора «ВОЗРАСТ» к среднему квадрату ошибки) равен F=17,37, так и фактором «ДОХОД» (критерий Фишера для этого фактора 16,49).

По модели (1) можно проводить расчет значений вероятности хорошего здоровья мужского населения Орловского региона по МНК-оценкам ее параметров, представленных в табл. 2. В целом она достаточно хорошо описывает зависимость вероятности хорошего здоровья населения Орловского региона от возраста и уровня доходов – коэффициент детерминации $R2$=0,908, при этом параметры модели, кроме параметра b_3 (линейный эффект уровня «ВОЗРАСТ=41–50» фактора «ВОЗРАСТ»), статистически значимы: p-уровень не хуже критического значения 0,05, а 95%-й доверительные интервалы не включают нуль. Для параметра b_3 критерий Стьюдента t=0,771 статистически не значим: p-уровень больше 0,05, а 95%-й доверительный интервал для этого параметра включает нуль.

Графики, отражающие связь вероятности хорошего здоровья мужского населения Орловского региона с изучаемыми признаками, представлены на рис. 3 и 4.

Легко убедиться, что расчетные диаграммы вероятности хорошего здоровья мужского населения для уровней статусных факторов, приведенные на рис 3, полностью идентичны эмпирическим данным, в графическом виде представленным на рис. 1.

Таблица 2 – МНК-оценки параметров модели здоровья мужского населения Орловского региона

Параметр	Обозначение	Коэф. *B*	Ст. ошибка	t-критерий	P-уровень	95% доверительный интервал	
						нижний уровень	верхний уровень
Intercept	μ_0	17,655	4,099	4,307	0,001	8,724	26,586
[ДОХОД=2001-3000]	α_2	12,650	4,583	2,760	0,017	2,665	22,635
[ДОХОД=3001-4000]	α_3	16,825	4,583	3,671	0,003	6,840	26,810
[ДОХОД=4001-5000]	α_4	28,950	4,583	6,317	0,000	18,965	38,935
[ДОХОД=более 5000]	α_5	32,800	4,583	7,157	0,000	22,815	42,785
[ДОХОД=до 2000]	α_1	0	,	,	,	,	,
[ВОЗРАСТ=18-30]	β_1	26,960	4,099	6,577	0,000	18,029	35,891
[ВОЗРАСТ=31-40]	β_2	12,160	4,099	2,967	0,012	3,229	21,091
[ВОЗРАСТ=41-50]	β_3	3,160	4,099	0,771	0,456	-5,771	12,091
[ВОЗРАСТ=51-60]	β_4	0	,	,	,	,	,

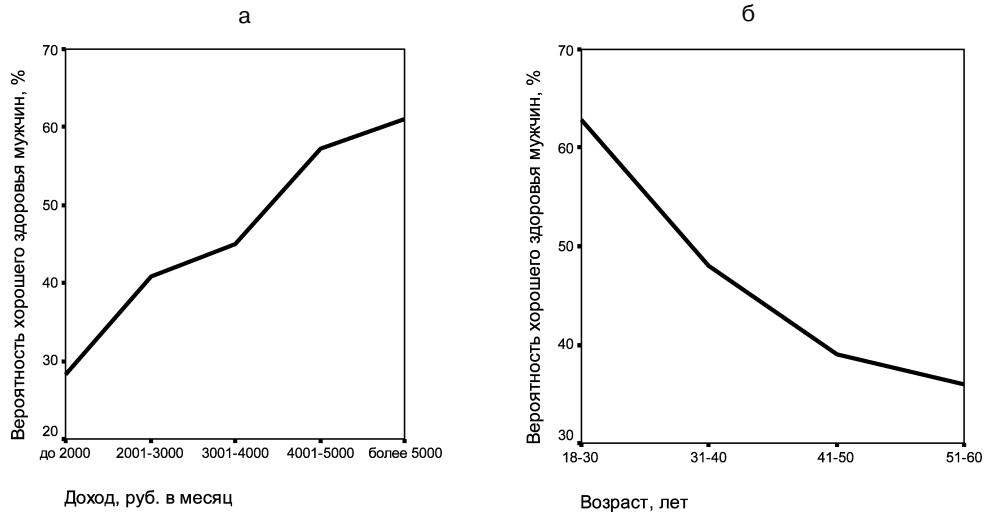


Рис. 3. Диаграмма вероятности хорошего здоровья мужского населения для уровней статусных факторов: *а* – дохода; *б* – возраста (расчет по модели (1))

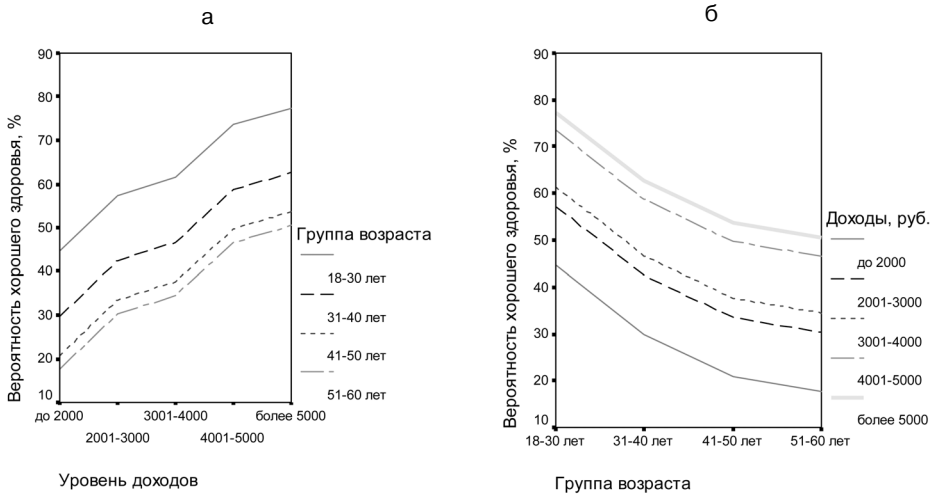


Рис. 4. Диаграмма вероятности хорошего здоровья мужского населения для уровней каждого из статусных факторов при различных значениях другого фактора: *а* – дохода; *б* – возраста (расчет)

Иная картина наблюдается при переходе к диаграммам, отражающим связь расчетных значений вероятности хорошего здоровья мужского населения Орловского региона от уровней каждого из статусных факторов при различных значениях другого фактора: в отличие от эмпирических данных, они сглаженные, симбатны друг другу – рис. 4. Это следствие линейности модели (1).

О характере сглаживания можно судить по рис. 5, на котором представлены графики зависимости вероятности хорошего здоровья мужского населения (фактические и расчетные значения) от уровней дохода для двух возрастных групп (18–30 и 31–40 лет). Видно, что графики для расчетных значений результативного признака лишены локальных максимумов и минимумов, проходят более гладко.

О качестве линейной модели дисперсионного анализа (1) можно судить по графикам рис. 6.

Из графика последовательности (рис. 6 *а*) и корреляционной диаграммы (рис. 6 *б*) видно, что в целом качество модели является удовлетворительным: на графике рис. 6 *а* расчетные значения повторяют эмпирические наблюдения, а на графике рис. 6 *б* эмпирические наблюдения достаточно близки к теоретической прямой регрессии. Наибольшие отклонения расчетных и фактических значений наблюдаются для групп респондентов 13 и 15. Это группы

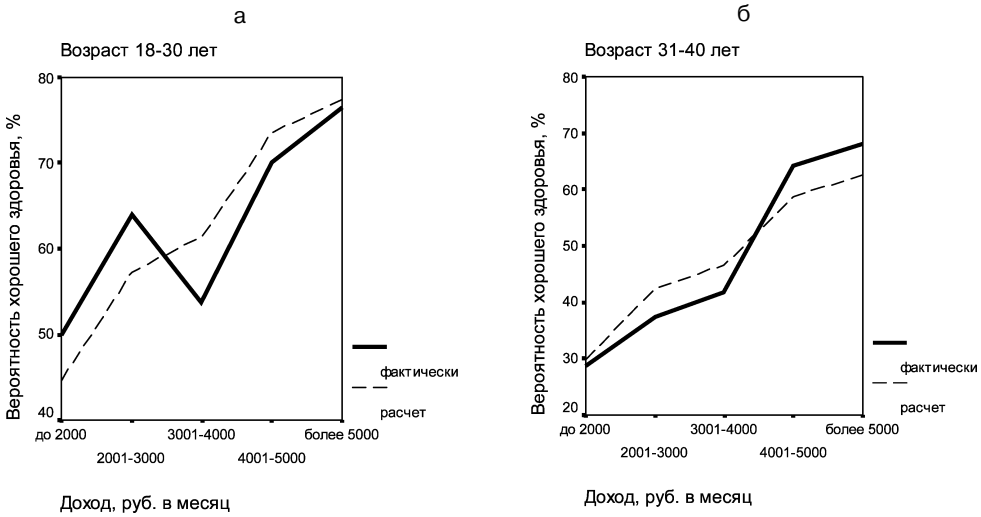


Рис. 5. Зависимость вероятности хорошего здоровья мужского населения Орловского региона (фактические и расчетные значения) от уровней дохода для возрастных групп: *а* – 18–30 лет; *б* – 31–40 лет

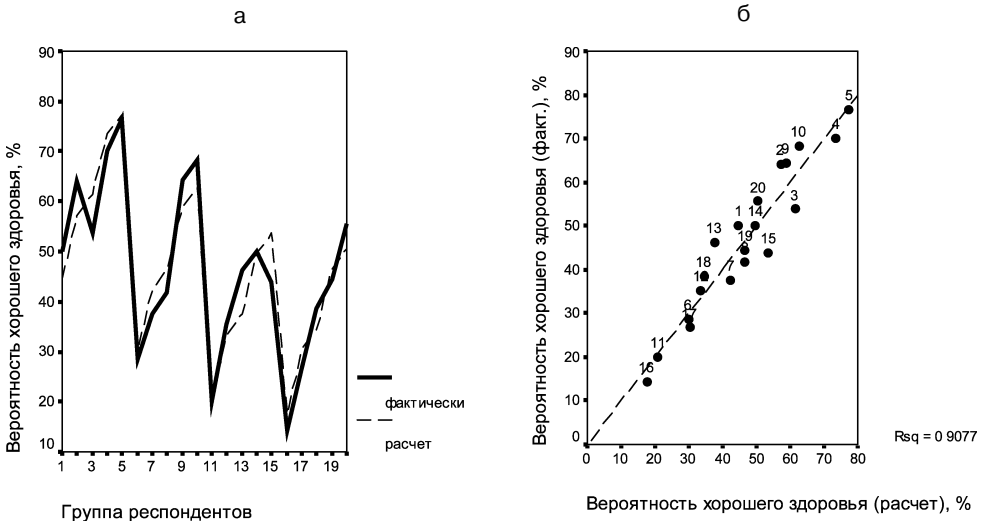


Рис. 6. Оценка качества модели вероятности хорошего здоровья мужского населения Орловского региона: *а* – график последовательности; *б* – корреляция расчетных и фактических значений

мужчин возраста от 41 до 50 лет, располагающих доходами 3001–4000 и более 5000 руб. в месяц соответственно. Именно для этой возрастной группы соответствующий параметр модели (1) оказался статистически незначимым.

Линейная модель дисперсионного анализа (1) адекватна эмпирическим данным. Об этом можно судить по графикам рис. 7 – графику остатков и гистограмме их распределения: как следует из графика рис. 7 *а*, остатки равномерно распределены в полосе вдоль нулевого значения, что и позволяет принять гипотезу об адекватности линейной модели дисперсионного анализа (1). Распределение остатков близко к нормальному (рис. 7 *б*), хотя и наблюдаются повышенные значения остатков, отвечающие группам респондентов 13, 2, 15, 3.

По результатам выполненного моделирования можно сделать следующий вывод: линейная модель дисперсионного анализа (1) адекватна, обладает достаточно высоким качеством (объясняет более 90% общей дисперсии) и может быть использована для расчета вероятности хорошего здоровья мужского населения Орловского региона в зависимости от их возраста и уровня доходов.

Определенным недостатком модели (1) является исключение из эмпирической базы для ее построения важной возрастной группы мужского населения – респондентов старше 60 лет. Для устранения этого недостатка следует применить часто используемый в статистических исследованиях прием – замену ненадежных данных средними значениями. Тогда вместо данных, приведенных нами ранее в публикации[7], мы получаем новую эмпирическую базу, в которой частости, отмеченные курсивом, заменены средними значениями (табл. 3).

Таблица 3 – Таблица частости категории «здоровье хорошее» для признаков «Возраст» * «Уровень доходов» – мужское население Орловского региона (проценты «по строкам»)

Доход, руб. в месяц	Возраст, лет				
	18-30	31-40	41-50	51-60	более 60
До 2000	32,8 (50,0)	28,6	20,0	14,3	7,7
2001-3000	42,7 (64,0)	37,5	37,5	26,7	27,3
3001-4000	53,8	45,5	46,2	38,5	28,6
4001-5000	70,0	75,0	55,0	44,4	40,0
Более 5000	76,5	68,2	43,8	55,6	63,1 (100,0)

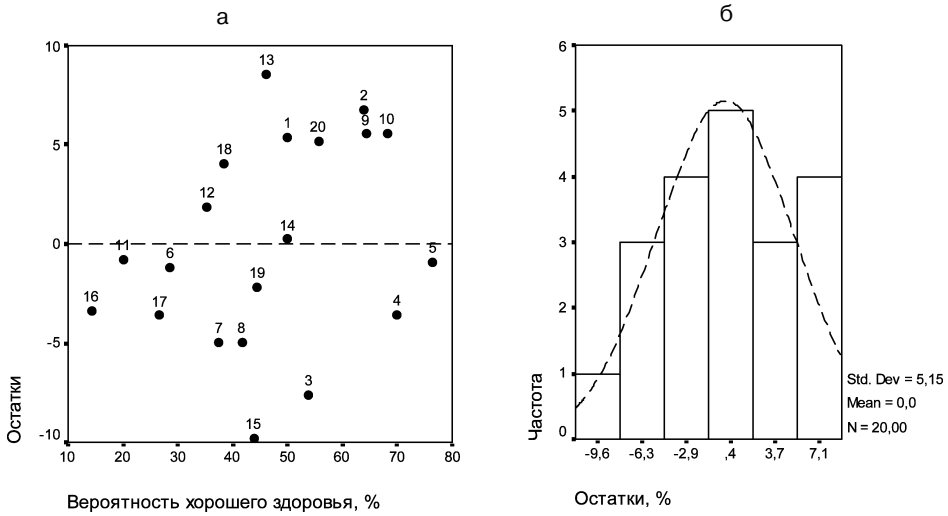


Рис. 7. Оценка адекватности модели вероятности хорошего здоровья мужского населения Орловского региона: *а* – график остатков; *б* – гистограмма распределения остатков

Таблица 4 – МНК-оценки параметров модели (2) (вероятность хорошего здоровья мужского населения Орловского региона)

Параметр	Обозначение	Коэф. B	Ст. ошибка	t-критерий	P-уровень	95% доверительный интервал	
						нижний уровень	верхний уровень
Intercept	μ_0	10,848	4,071	2,665	0,017	2,218	19,478
[ДОХОД=2001-3000]	α_2	13,660	4,291	3,183	0,006	4,563	22,757
[ДОХОД=3001-4000]	α_3	21,840	4,291	5,089	0,000	12,743	30,937
[ДОХОД=4001-5000]	α_4	36,200	4,291	8,436	0,000	27,103	45,297
[ДОХОД=более 5000]	α_5	40,760	4,291	9,498	0,000	31,663	49,857
[ДОХОД=до 2000]	α_1	0	,	,	,	,	,
[ВОЗРАСТ=18-30]	β_1	21,820	4,291	5,085	0,000	12,723	30,917
[ВОЗРАСТ=31-40]	β_2	17,620	4,291	4,106	0,001	8,523	26,717
[ВОЗРАСТ=41-50]	β_3	7,160	4,291	1,669	0,115	-1,937	16,257
[ВОЗРАСТ=51-60]	β_4	2,560	4,291	,597	0,559	-6,537	11,657
[ВОЗРАСТ=более 60]	β_5	0	,	,	,	,	,

В результате получена модель (2), математически идентичная модели (1), но с другими значениями входящих в нее параметров – табл. 4.

В содержательном плане модель (2) дает результаты, несколько отличающиеся от полученных ранее по модели (1) – табл 5.

Так, если по модели (1) вероятность хорошего здоровья мужского населения Орловского региона в равной мере определяют возраст и доход, то по модели (2) это соотношение меняется в пользу фактора «доход»: из табл. 5 следует, что с учетом возрастной группы «более 60 лет» результативный признак в большей мере определяет доход – критерий Фишера для этого фактора, т.е. отношение среднего квадрата фактора «ДОХОД» к среднему квадрату ошибки равен F=30,03, а критерий Фишера для фактора «ВОЗРАСТ» меньше (F=9,80), т.е. этот фактор менее значим. Несколько изменился и ход соответствующих зависимостей.

Рис. 8 иллюстрируют сказанное (для наглядности на графиках принят одинаковый масштаб оси «вероятность хорошего здоровья»).

Применив аналогичный прием к массиву социологических данных для женского населения Орловского региона, получаем табл. 6, где частости, отмеченные курсивом, заменены средними значениями.

Таблица 5 – Критерии межобъектных эффектов модели (3.2) (вероятность хорошего здоровья мужского населения Орловского региона)

Источник изменнивости	Сумма квадратов	Степени свободы *df*	Средний квадрат	Критерий Фишера F	Уровень значимости
Corrected Model	7335,017	8	916,877	19,916	0,000
Intercept	46595,540	1	46595,540	1012,130	0,000
ДОХОД	5529,742	4	1382,436	30,029	0,000
ВОЗРАСТ	1805,274	4	451,319	9,803	0,000
Ошибка	736,594	16	46,037		
Всего	54667,150	25			
Corrected Total	8071,610	24			

Таблица 6 – Таблица частости категории «здоровье хорошее» для признаков «Возраст» * «Уровень доходов» – женское население Орловского региона

Доход, руб. в месяц	Возраст, лет				
	18-30	31-40	41-50	51-60	более 60
До 2000	61,1	27,3	11,5	20,0	9,1
2001-3000	70,6	34,6	15,4	20,8	5,0
3001-4000	61,9	52,6	31,6	37,5	7,7
4001-5000	75,0	36,4	37,5	23,1	46,0 (57,1)
Более 5000	80,0	43,5	42,4 (9,1)	25,0	42,4 (0,0)

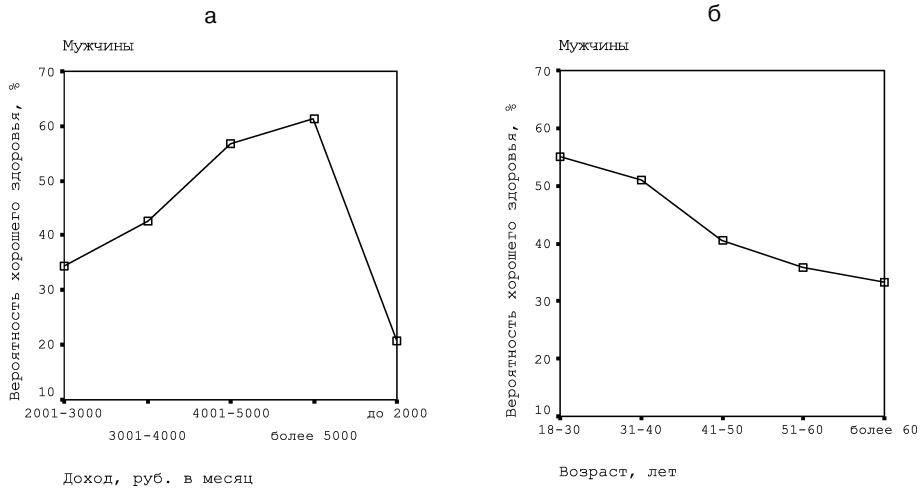


Рис. 8. Диаграмма вероятности хорошего здоровья мужского населения Орловского региона для уровней статусных факторов: *а* – дохода; *б* – возраста. Расчет по модели (2)

Таблица 7 – МНК-оценки параметров модели (3.3) (вероятность хорошего здоровья женского населения Орловского региона)

Параметр	Обозначение	Коэф. B	Ст. ошибка	t-критерий	P-уровень	95% доверительный интервал	
						нижний уровень	верхний уровень
Intercept	μ_0	10,848	4,071	2,665	0,017	2,218	19,478
[ДОХОД=2001-3000]	α_2	13,660	4,291	3,183	0,006	4,563	22,757
[ДОХОД=3001-4000]	α_3	21,840	4,291	5,089	0,000	12,743	30,937
[ДОХОД=4001-5000]	α_4	36,200	4,291	8,436	0,000	27,103	45,297
[ДОХОД=более 5000]	α_5	40,760	4,291	9,498	0,000	31,663	49,857
[ДОХОД=до 2000]	α_1	0	,	,	,	,	,
[ВОЗРАСТ=18-30]	β_1	21,820	4,291	5,085	0,000	12,723	30,917
[ВОЗРАСТ=31-40]	β_2	17,620	4,291	4,106	0,001	8,523	26,717
[ВОЗРАСТ=41-50]	β_3	7,160	4,291	1,669	0,115	-1,937	16,257
[ВОЗРАСТ=51-60]	β_4	2,560	4,291	,597	0,559	-6,537	11,657
[ВОЗРАСТ=более 60]	β_5	0	,	,	,	,	,

В результате реализации процедуры «Общая линейная модель» получаем МНК-оценки параметров модели (3), связывающей зависимость вероятности хорошего здоровья женского населения Орловского региона с градациями возраста и здоровья (табл. 7), и оценки степени влияния статусных факторов (табл. 8).

Из табл. 8 следует, что возраст женщин больше, чем доход, влияет на вероятность хорошего здоровья. Этот вывод противоположен полученному выше результату для мужчин: вероятность

Таблица 8 – Критерии межобъектных эффектов модели (3.3) (вероятность хорошего здоровья женского населения Орловского региона)

Источник изменчивости	Сумма квадратов	Степени свободы *df*	Средний квадрат	Критерий Фишера F	Уровень значимости
Corrected Model	9224,364	8	1153,045	11,405	0,000
Intercept	33708,960	1	33708,960	333,419	0,000
ДОХОД	1615,548	4	403,887	3,995	0,020
ВОЗРАСТ	7608,816	4	1902,204	18,815	0,000
Ошибка	1617,616	16	101,101		
Всего	44550,940	25			
Corrected Total	10841,980	24			

хорошего здоровья мужского населения Орловского региона в большей мере определяет доход.

Статистический вывод о различии степени влияния факторов на исследуемый показатель иллюстрируют графики на рис. 9 и 10 (порядок следования уровней дохода процедурой «общая линейная модель» изменен: уровень «до 2000» следует за уровнем «более 5000»).

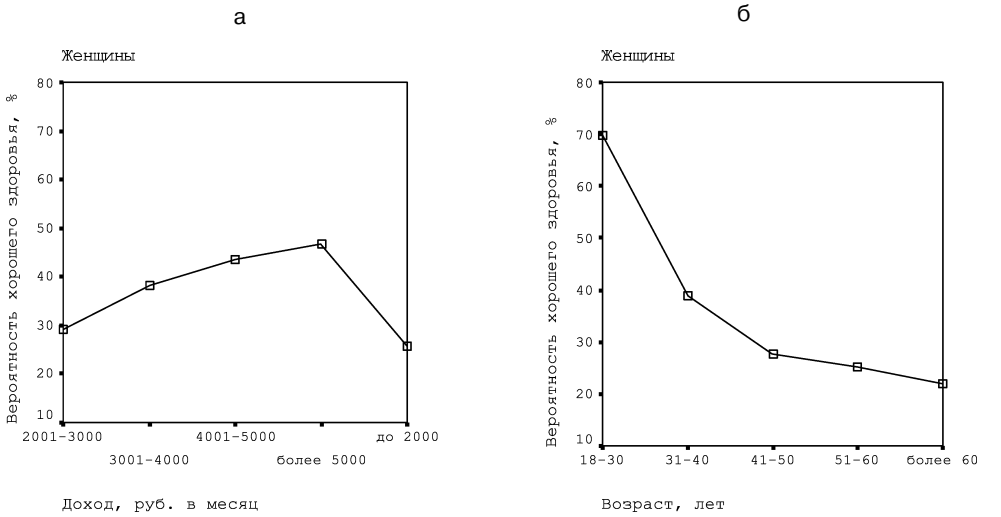


Рис. 9. Диаграмма вероятности хорошего здоровья женского населения Орловского региона для уровней статусных факторов: *а* – дохода; *б* – возраста. Расчет по модели (3)

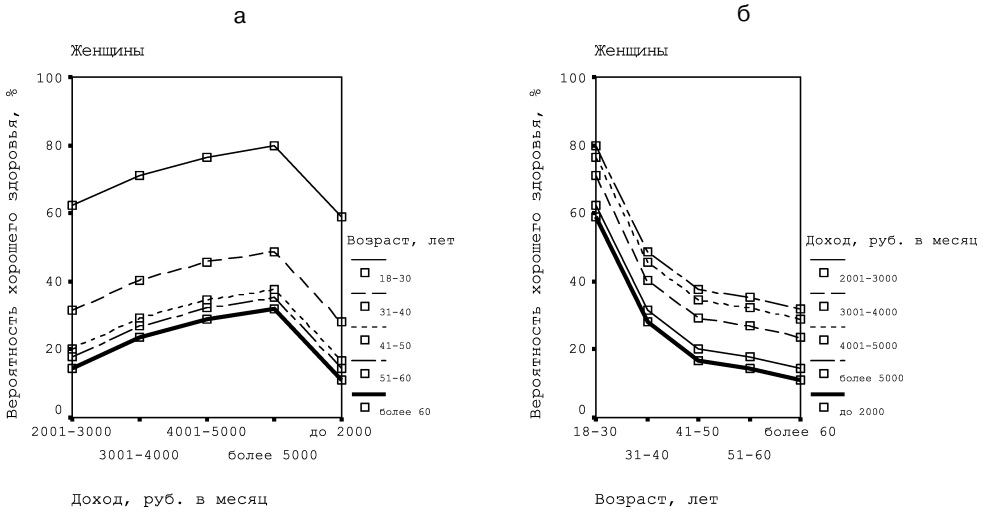


Рис. 10. Диаграмма вероятности хорошего здоровья женского населения Орловского региона для уровней каждого из статусных факторов при различных значениях другого: *а* – дохода; *б* – возраста. Расчет по модели (3)

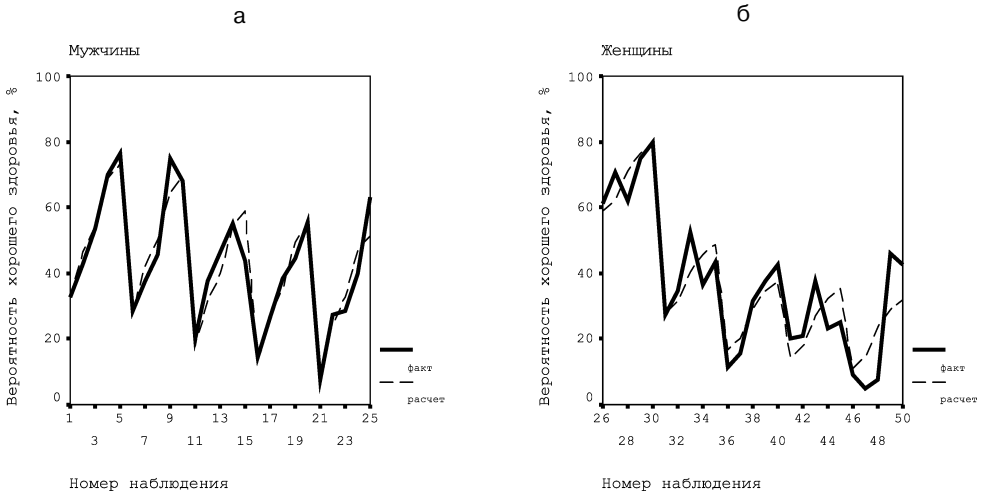


Рис. 11. Графическая иллюстрация погрешности моделей вероятности хорошего здоровья мужского (*а*) и женского (*б*) населения Орловского региона

О погрешности моделей (2) и (3) можно судить по графикам рис. 11. Видно, что модель (2) более точная, чем модель (3).

Итак, в результате применения технологии дисперсионного анализа по схеме «Общая линейная модель» получены модели, отражающие влияние двух факторов – возраста и уровня дохода – на здоровье мужского и женского населения Орловского региона. Выявлены существенные различия моделей здоровья мужского и женского населения: вероятность хорошего здоровья мужчин в большей мере определяется доходом, чем возрастом, в то время как для женщин более значимым оказывается возраст. Вероятность хорошего здоровья и мужчин, и женщин монотонно увеличивается с ростом доходов и уменьшается с возрастом.

Полученные результаты позволяют рекомендовать применение методов дисперсионного анализа в прикладных социологических исследованиях в случаях, когда ставится задача выявления связи результативного признака, измеренного по количественным шкалам, с градациями (уровнями) одного и более статусных или факторных признаков, измеренных по номинальным или порядковым шкалам. В качестве инструмента моделирования при этом рекомендуется использовать процедуру «General Linear Model» (общая линейная модель) пакета анализа данных общественных наук SPSS Base.

---

[1] Шуметов В.Г., Лясковская О.В. О применении количественных методов анализа в прикладных социологических исследованиях: проблемы и пути их решения // Среднерусский вестник общественных наук. 2013. № 3.

[2] Игнатова Т.В., Перевозкин Л.М. Социологический мониторинг в системе муниципального управления: вопросы теории и практики // Среднерусский вестник общественных наук. 2012. № 4 (2). С. 35–39.

[3] Уварова В.И., Шуметов В.Г., Лясковская О.В. Здоровье населения Центральной России: типология, модели, прогнозы: Монография. Орел: ОрелГТУ, 2008.

[4] Бююль А., Цёфель П. SPSS: Искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. СПб.: ООО «ДиаСофтЮП», 2002.

[5] Афифи А., Эйзен С. Статистический анализ: Подход с использованием ЭВМ. М.: Мир, 1982.

[6] Шуметов В.Г., Лясковская О.В. О применении количественных методов анализа в прикладных социологических исследованиях: проблемы и пути их решения // Среднерусский вестник общественных наук. 2013. № 3.

[7] Там же.