УДК 004.932

Захарова Е.А.[1], Буланова Ю. А.[2]

МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ЗАБОЛЕВАЕМОСТИ РАКОМ МОЛОЧНОЙ ЖЕЛЕЗЫ

[1]НУЗ «Отделенческая больница на станции Муром ОАО «РЖД», г. Муром
[2]ГОУ ВПО Владимирский государственный университет, (филиал) Муромский институт, г. Муром

Резюме. *В статье рассматривается применение логлинейного анализа для разработки математических моделей молочных желез, выявляющих наиболее значимые показатели при развитии злокачественных и доброкачественных заболеваний. Материалы методы содержат математическое описание модели. Проверка на адекватность выполняется с использованием критерия χ^2 Пирсона. В качестве примера было проведено исследование модели для группы «репродуктивная функция + болезнь», которое показало, что основными факторами возникновения злокачественных заболеваний являются возраст первой беременности и возраст первых родов. Таким образом, используя данную модель, можно прогнозировать влияние тех или иных факторов на развитие опухолей молочной железы.*
Ключевые слова. *Молочная железа, заболевания, рак молочной железы, логлинейный анализ, математическая модель, параметры, адекватность, показатели.*

Zakharova E.A.[1], Bulanova Y. A.[2]

MATHEMATICAL MODEL OF BREAST CANCER

Summary. *The article discusses the use of log-linear analysis for the development of mathematical models of mammary glands, identifies the most significant indicators of the development of malignant and benign diseases. Materials methods contain the mathematical description of the model. Checking the adequacy performed using Pearson criterion χ^2. As an example, a study was conducted for the model of the "reproductive function + disease", which showed that the main factors of malignant diseases are age at first pregnancy and age at first birth . Thus, using this model can predict the influence of various factors on the development of breast tumors.*
Keywords. *Mammary gland diseases, breast cancer, log-linear analysis, mathematical model, parameters, value, performance.*

Введение. На основе своих знаний и опыта, врач определяет заболевания молочной железы [1] по небольшому числу наиболее значимых параметров [2]. Поэтому возникает необходимость в разработке инструментария, помогающего врачу при постановке диагноза с учетом сопутствующих факторов, например, данных о менструальной и репродуктивной функциях.

Целью настоящей работы является построение моделей заболеваний, таких как киста, фиброаденома и рак молочной железы, на основе математических методов и информационных технологий обработки данных [1, 3, 6], выявление наиболее значимых показателей для развития злокачественных опухолей, прогнозирование влияния факторов риска на заболеваемость раком молочной железы.

Построение математических моделей осуществлялось на основе логлинейного анализа данных, позволяющего установить силу и значимость связи между признаками с учетом их взаимодействия, а также определить степень влияния входных параметров на выходные результирующие признаки-отклики [4]. Расчет и построение логлинейных моделей проводились в Statistica 6.0.

Материалы и методы. Особенности логлинейного анализа состоят в следующем [5]. Рассмотрим двумерную таблицу сопряженности r×s. Представим теоретические частоты в ячейках такой таблицы в виде:

$$n_{ij}^{*} = e^{u_0 + u_i^a + u_j^b + u_{ij}^{ab}} . \qquad (1)$$

В логлинейной модели теоретические ожидаемые частоты n^*_{ij}преобразуются в их логарифмы, представляющие собой сумму из четырех параметров модели.

$$\ln n^*_{ij} = u_0 + u^a_i + u^b_j + u^{ab}_{ij}, \quad (2)$$

где n^*_{ij}– теоретическая частота в ячейке,
u – неизвестные параметры, называемые:
u^a_i – эффект i-ой градации первого признака,
u^b_j – эффект j-ой градации второго признака,
u^{ab}_{ij} – эффект взаимодействия двух признаков,
u_0 – общийэффект,
$i \in [1, r]$, $j \in [1, s]$.

Параметры логлинейной модели иногда называют также вкладами, вносимыми различными эффектами в теоретическую частоту.

Для проверки адекватности модели используется критерий $\chi^2_{инф}$Пирсона [4]:

$$\chi^2_{инф} = 2 \times \sum_i \sum_j n_{ij} * \ln \frac{n_{ij}}{n^*_{ij}}, \quad (3)$$

где n_{ij} – теоретические частоты, полученные по наблюдавшимся частотам,
n^*_{ij} – теоретические частоты, полученные по модели.

Поскольку логлинейная модель накладывает свои ограничения на количество исследуемых параметров, сгруппируем все полученные анкетные данные: «возраст + болезнь», «менструальная функция + болезнь», «репродуктивная функция + болезнь», «сопутствующие заболевания + болезнь», «травмы + болезнь», «индекс массы тела (ИМТ) + болезнь», «образование + место жительства + болезнь». Следовательно, логлинейная модель представляет собой систему моделей для каждой из группы параметров.

Результаты исследований и их обсуждение. В исследование включались женщины возрастом от 25 до 70 лет, проживающие в округе Муром. В анамнезе женщин отсутствовали хронические гинекологические заболевания, такие как миома матки, эндометриоз, НМЦ.

На основании данных анкетного опроса с помощью логлинейного анализа были проанализированы такие факторы риска, как возраст женщины, количество родов, сопутствующие гинекологические заболевания, количество абортов, наличие онкологических заболеваний молочных желез в семье, питание, стресс, место жительства, гормональная терапия, наследственность, зависимость от вредных привычек, были исследованы их взаимосвязи.

В данной статье внимание обращено на определение зависимости риска развития рака МЖ от возраста первой беременности и первых родов.

Рассмотрим расчет модели для группы «репродуктивная функция + болезнь», расчеты для других групп аналогичны.

Репродуктивная функция объединяет в себя 2 подгруппы: возраст женщины при первой беременности, возраст женщины при первых родах. По данной группе будет построено 2 модели, учитывающие взаимосвязь каждой подгруппы. На рис. 1 представлено формирование лучшей выборки по всем данным для логлинейного моделирования, результаты которого представлены на рис. 2.

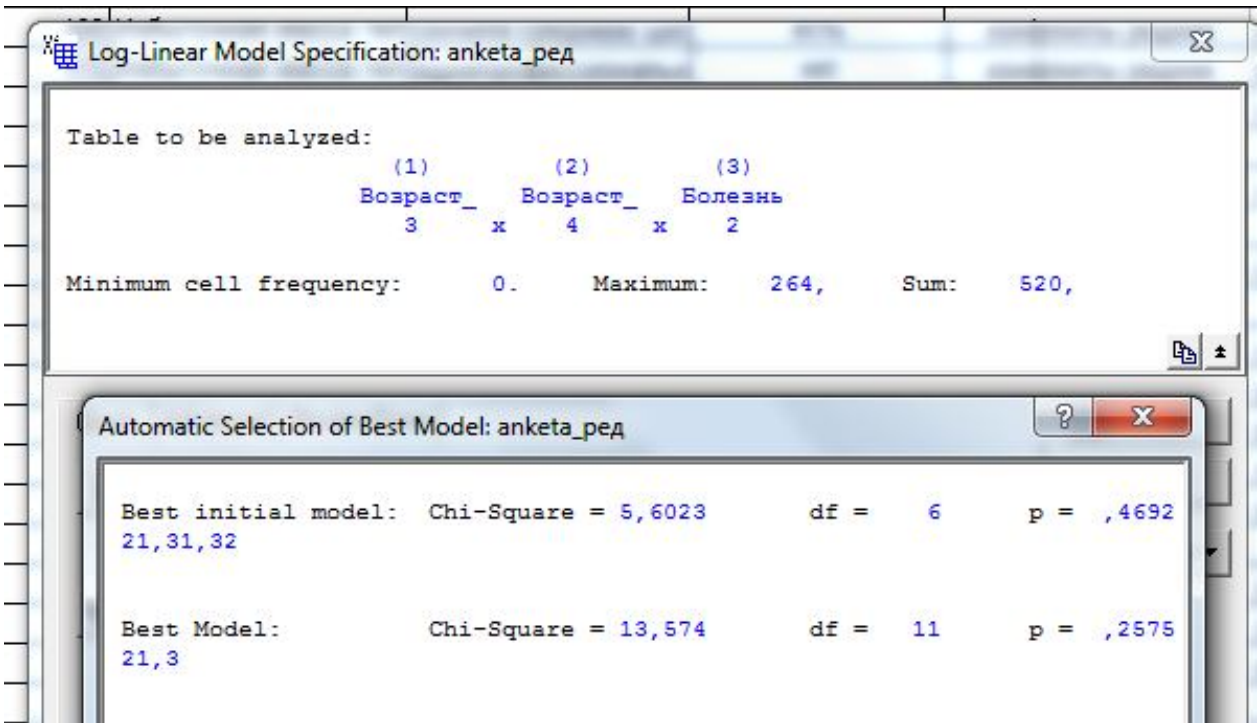

Рис. 1. Формирование лучшей выборки.

```
Table to be analyzed:
                    (1)         (2)         (3)
                Возраст_    Возраст_    Болезнь
                   3     x     4     x     2

Minimum cell frequency:      0.     Maximum:    264,     Sum:    520,

Model to be tested: 21,3

Delta: ,5000 ;   Maximum iterations:  50 ;  Conv. criterion: ,0100
Convergence reached after   2 iterations
                                          df       p
Maximum Likelihood Chi-square: 13,574      11    ,25749
          Pearson Chi-square: 12,316      11    ,34034
```

Рис. 2. Результаты.

Возраст женщины при первой беременности подразделяется на 5 групп: Группа 1 – до 18 лет, Группа 2 – 18-25 лет, Группа 3 – 26-30 лет, Группа 4 – 31-35 лет, Группа 5 - старше 36 лет.

Возраст женщины при первых родах можно также сгруппировать: Группа 1 – не было родов, Группа 2 – моложе 18 лет, Группа 3 – 18-25 лет, Группа 4 – старше 25 лет.

Болезнь	Obs. Freq. (+delta): Болезнь by Возраст_первой_берем w/in vars: (anketa_ред) Возраст_первых_родов:Группа1			
	Возраст_первой_берем Группа 1	Возраст_первой_берем Группа 2	Возраст_первой_берем Группа 3	Total
здорова	7,50000	4,500000	0,500000	12,50000
больна	10,50000	4,500000	0,500000	15,50000
Total	18,00000	9,000000	1,000000	28,00000

а)

Болезнь	Obs. Freq. (+delta): Болезнь by Возраст_первой_берем w/in vars: (anketa_ред) Возраст_первых_родов:Группа2			
	Возраст_первой_берем Группа 1	Возраст_первой_берем Группа 2	Возраст_первой_берем Группа 3	Total
здорова	3,500000	0,500000	0,500000	4,500000
больна	0,500000	0,500000	0,500000	1,500000
Total	4,000000	1,000000	1,000000	6,000000

б)

Болезнь	Obs. Freq. (+delta): Болезнь by Возраст_первой_берем w/in vars: (anketa_ред) Возраст_первых_родов:Группа3			
	Возраст_первой_берем Группа 1	Возраст_первой_берем Группа 2	Возраст_первой_берем Группа 3	**Total**
здорова	1,500000	157,5000	0,500000	159,5000
больна	2,500000	264,5000	0,500000	267,5000
Total	4,000000	422,0000	1,000000	427,0000

в)

Болезнь	Obs. Freq. (+delta): Болезнь by Возраст_первой_берем w/in vars: (anketa_ред) Возраст_первых_родов:Группа4			
	Возраст_первой_берем Группа 1	Возраст_первой_берем Группа 2	Возраст_первой_берем Группа 3	**Total**
здорова	1,500000	1,50000	22,50000	25,50000
больна	0,500000	14,50000	30,50000	45,50000
Total	2,000000	16,00000	53,00000	71,00000

г)

Рис. 3. Таблицы частот наблюдений для различных сочетаний уровней признаков:
а) сочетание диагноза и возраста женщины при первой беременности для группы 1,
б) сочетание диагноза и возраста женщины при первой беременности для группы 2,
в) сочетание диагноза и возраста женщины при первой беременности для группы 3,
г) сочетание диагноза и возраста женщины при первой беременности для группы 4.

Fitted Freq.: Болезнь by Возраст_первой_берем w/in vars: (anketa_ред)
Возраст_первых_родов:Группа1

Болезнь	Возраст_первой_берем Группа 1	Возраст_первой_берем Группа 2	Возраст_первой_берем Группа 3	**Total**
Здоров	6,83459	3,417293	0,379699	10,63158
Болен	11,16541	5,582707	0,620301	17,36842
Total	18,00000	9,000000	1,000000	28,00000

а)

Fitted Freq.: Болезнь by Возраст_первой_берем w/in vars: (anketa_ред)
Возраст_первых_родов:Группа2

Болезнь	Возраст_первой_берем Группа 1	Возраст_первой_берем Группа 2	Возраст_первой_берем Группа 3	**Total**
Здоров	1,518797	0,379699	0,379699	2,278196
Болен	2,481203	0,620301	0,620301	3,721805
Total	4,000000	1,000000	1,000000	6,000000

б)

Fitted Freq.: Болезнь by Возраст_первой_берем w/in vars: (anketa_ред)
Возраст_первых_родов:Группа3

Болезнь	**Возраст_первой_берем Группа 1**	Возраст_первой_берем Группа 2	Возраст_первой_берем Группа 3	Total
Здоров	1,518797	160,2331	0,379699	162,1316
Болен	2,481203	261,7669	0,620301	264,8684
Total	4,000000	422,0000	1,000000	427,0000

в)

Fitted Freq.: Болезнь by Возраст_первой_берем w/in vars: (anketa_ред)
Возраст_первых_родов:Группа4

Болезнь	Возраст_первой_берем Группа 1	Возраст_первой_берем Группа 2	Возраст_первой_берем Группа 3	**Total**
Здоров	0,759398	6,07519	20,12406	26,95865
Болен	1,240602	9,92481	32,87594	44,04135
Total	2,000000	16,00000	53,00000	71,00000

г)

Рис. 4. Расчет четырехпольных таблиц ожидаемых частот для всех сочетаний уровней признаков на основе полученной адекватной модели:
а) для группы 1, б) для группы 2, в) для группы 3, г) для группы 4.

Таблицы на рис. 4 показывают прогноз ожидаемых частот наблюдений для различных уровней факторов. Сравнение теоретических частот с наблюдавшимися (рис. 3) показало незначительные различия, что прямым образом указывает на адекватность модели.

Болезнь	Stdrd. Resid.: Болезнь by Возраст_первой_берем w/in vars: (anketa_ред) Возраст_первых_родов:Группа1			
	Возраст_первой_берем Группа 1	Возраст_первой_берем Группа 2	Возраст_первой_берем Группа 3	**Total**
Здоров	0,254528	0,585692	0,195231	1,035451
Болен	-0,199138	-0,458235	-0,152745	-0,810118
Total	0,055390	0,127457	0,042486	0,225333

а)

Болезнь	Stdrd. Resid.: Болезнь by Возраст_первой_берем w/in vars: (anketa_ред) Возраст_первых_родов:Группа2			
	Возраст_первой_берем Группа 1	Возраст_первой_берем Группа 2	Возраст_первой_берем Группа 3	**Total**
Здоров	1,60760	0,195231	0,195231	1,99807
Болен	-1,25776	-0,152745	-0,152745	-1,56325
Total	0,34984	0,042486	0,042486	0,43482

б)

Болезнь	Stdrd. Resid.: Болезнь by Возраст_первой_берем w/in vars: (anketa_ред) Возраст_первых_родов:Группа3			
	Возраст_первой_берем Группа 1	Возраст_первой_берем Группа 2	Возраст_первой_берем Группа 3	**Total**
Здоров	-0,015252	-0,215912	0,195231	-0,035933
Болен	0,011933	0,168926	-0,152745	0,028114
Total	-0,003319	-0,046985	0,042486	-0,007819

в)

Болезнь	Stdrd. Resid.: Болезнь by Возраст_первой_берем w/in vars: (anketa_ред) Возраст_первых_родов:Группа4			
	Возраст_первой_берем Группа 1	Возраст_первой_берем Группа 2	Возраст_первой_берем Группа 3	**Total**
Здоров	0,849865	-1,85622	0,529636	-0,476718
Болен	-0,664918	1,45227	-0,414378	0,372975
Total	0,184946	-0,40395	0,115259	-0,103742

г)

Рис. 5. Расчет четырехпольных таблиц разностей наблюдавшихся и ожидаемых частот и стандартизованных разностей для всех сочетаний уровней признаков:
а) для группы 1, б) для группы 2, в) для группы 3, г) для группы 4.

Анализируя стандартизованные разности (рис. 5) наблюдавшихся и ожидаемых частот (рис. 6) на всех сочетаниях уровней факторов, можно сказать, что закон их распределения близок к нормальному, так как стандартизованные разности не выходят за пределы интервала (-2; 2).

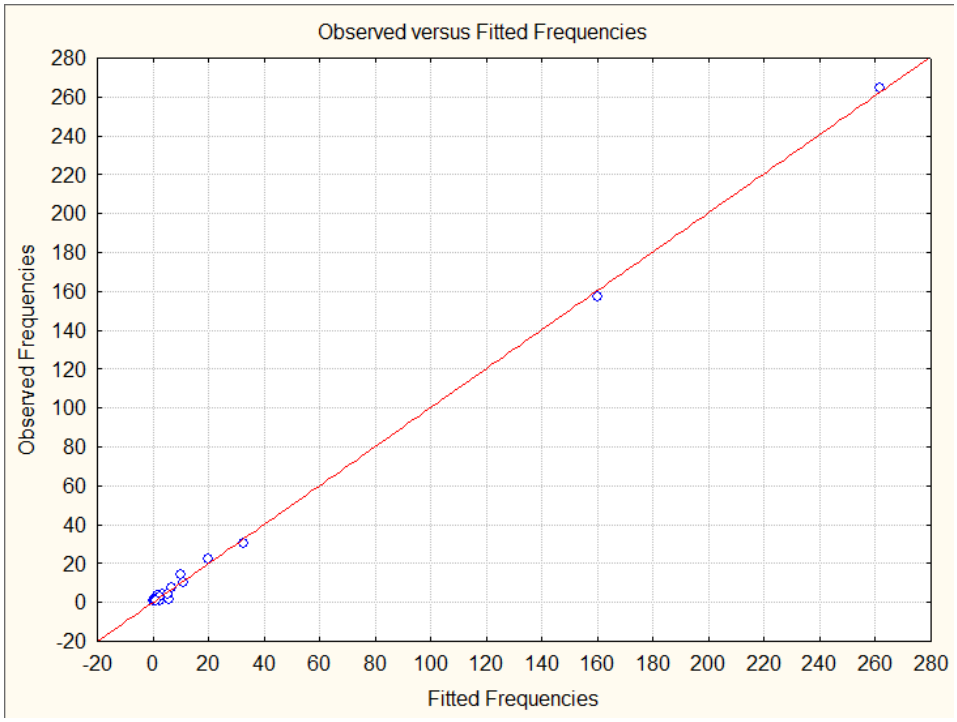

Рис. 6. Распределение отклонений наблюдавшихся частот от теоретических (по модели).

На рис. 7 представлен расчет наблюдений для групп «Возраст женщины при первой беременности» и «Возраст женщины при первых родах» попарно, который показал, что наибольшее количество наблюдений соответствует группе женщин в возрасте от 18 до 25 лет.

	Marg. Tabl. (freq+delta):Возраст_первой_берем by Возраст_первых_родов (anketa_ред)			
Возраст_первых_родов	Возраст_первой_берем Группа 1	Возраст_первой_берем Группа 2	Возраст_первой_берем Группа 3	**Total**
Группа 1	18,00000	9,0000	1,00000	28,0000
Группа 2	4,00000	1,0000	1,00000	6,0000
Группа 3	4,00000	422,0000	1,00000	427,0000
Группа 4	2,00000	16,0000	53,00000	71,0000
Total	28,00000	448,0000	56,00000	532,0000

Рис. 7. Расчет частот наблюдений для всех признаков попарно.

	Results of Fitting all K-Factor Interactions (anketa_ред) These are simultaneous tests that all K-Factor Interactions are simultaneously Zero.				
K-Factor	Degrs.of Freedom	Max.Lik. Chi-squ.	Probab. p	Pearson Chi-squ	**Probab. p**
1	6	1411,581	0,000000	3146,988	0,000000
2	11	373,191	0,000000	675,132	0,000000
3	6	5,610	0,468260	5,940	0,429890

Рис. 8. Значимость эффектов К-порядка.

Опираясь на таблицу из рис. 8 можно сказать, что значимыми являются эффекты первого и второго порядка (возраст первой беременности и возраст первых родов), вероятность появления эффекта третьего порядка более 0,05, что не обеспечивает требуемую достоверность в 95%, поэтому данный эффект не рассматриваем.

Степень влияния эффектов факторов и их взаимодействия на ожидаемые частоты наблюдений определяются по данным из таблицы коэффициентов парциальной и

маргинальной ассоциации с последующей оценкой значимости по методу χ^2 для полной насыщенной модели (рис. 9) [2].

Степень влияния каждого из эффектов определяют в соответствии с отношением критерия χ^2 данного эффекта к сумме χ^2 всех эффектов:

$$K_m = \frac{100 \times \chi^2}{\sum \chi_m^2}, \quad (4)$$

	Tests of Marginal and Partial Association (anketa_ред)				
Effect	Degrs.of Freedom	Prt.Ass. Chi-sqr.	Prt.Ass. p	Mrg.Ass. Chi-sqr.	**Mrg.Ass. p**
1	2	597,9124	0,000000	597,9124	0,000000
2	3	782,5673	0,000000	782,5673	0,000000
3	1	31,1013	0,000000	31,1013	0,000000
12	6	366,4903	0,000000	365,2275	0,000000
13	2	3,8471	0,146089	2,5844	0,274673
23	3	5,3793	0,146036	4,1165	0,249161

Рис. 9. Оценка значимости эффектов факторов и их взаимодействий в полной насыщенной модели.

По таблице из рис. 9 рассчитаем сумму χ^2 всех эффектов, она равна: $\sum \chi_m^2 = 1787{,}3$.

Таблица 1

Степень влияния эффектов факторов на частоты наблюдений

Эффекты факторов	**K_m, %**
1	33,45
2	43,78
3	1,74
12	20,5
13	0,22
23	0,31

В табл. 1 отражен расчет влияния каждого из факторов по отдельности на заболеваемость раком молочной железы (РМЖ), а также их сочетание. Оказалось, что самыми основными факторами является возраст первой беременности и возраст первых родов, а также их взаимное влияние.

Выводы.

Анализируя все выше сказанное, можно сделать следующие выводы:

1) построена модель заболевания раком молочной железы с применением логлинейного анализа;
2) установлено, что заболеваемость раком молочной железы существенно зависит от возраста первой беременности (33,45%) и возраста первых родов (43,78%);
3) выделение определенных групп пациентов по отношению к болезни и построение подобных логлинейных моделей позволит достоверно выявить и доказать влияние факторов риска на развитие данного заболевания. Метод может быть использован в кабинетах статистики для выявления наиболее значимых факторов риска развития рака молочной железы для данной местности, кроме того, использование данного метода возможно для выявления новых факторов риска развития рака молочной железы.

Литература.

1. Буланова Ю.А. Использование информационных технологий для локализации области рака молочной железы на маммограммах с преобладанием железистого компонента // Прикаспийский журнал: управление и высокие технологии. – 2013. – №3(23). – С. 100-111.

2. Захарова Е.АА. Обзор медицинской системы КМИС и формирование статистики заболеваемости молочных желез / Е.А. Захарова, Ю.А. Буланова // Алгоритмы, методы и системы обработки данных. – 2012. – №19. – С. 54-61.
3. Садыков С.С. Использование информационных технологий для выявления области кисты молочной железы на маммограммах / С.С. Садыков, Е.А. Захарова, Ю.А. Буланова // Вестник рентгенологии и радиологии. – 2013. – № 3. – С. 15-20.
4. Трошин Л.И. Статистический анализ нечисловой информации / Л.И. Трошин, В.А. Балаш, О.С. Балаш. – Московский государственный университет экономики, статистики и информатики. – М., 2003. – 67 с.
5. Юнкеров В.И. Математико-статистическая обработка медицинских исследований / В.И. Юнкеров, С.Г. Григорьев. – СПб.: ВМедА, 2002. - 266 с.
6. Sadykov S.S. Algorithm of localization of breast cancer in the background of mastopathy / S.S Sadykov, Y.A. Bulanova // 11th International Conference on Pattern Recognition and Image Analysis: New Information Technologies (PRIA-11-2013). – 2013. – № 2. – P. 717-721.