ӘДІСТЕМЕЛІК НҰСҚАУЛАР

УДК 613.1:610.6

ПОСТРОЕНИЕ МОДЕЛЕЙ, ОТРАЖАЮЩИХ ВЛИЯНИЕ ОКРУЖАЮЩЕЙ СРЕДЫ НА СОСТОЯНИЕ ЗДОРОВЬЯ НАСЕЛЕНИЯ, В ПРОГРАММЕ STATISTICA

О.В Гребенева., К.З. Сакиев, М.Б. Отарбаева, Н.М. Жанбасинова

РГКП «Национальный центр гигиены труда и профессиональных заболеваний» МЗСР РК, г. Караганда

Введение. Статистическая обработка данных, получаемых исследователем при проведении различного вида экспериментов, или составление врачом отчетов в повседневной медицинской практике требует проверки степени достоверности получаемых результатов, правильности их обобщения и выявления закономерностей [1]. Растет роль математических методов и в экологических исследованиях. Преимущества математического подхода в современной науке определяется двумя моментами: 1. Возрастает необходимость в уточнении понятий. Математика может оперировать только с четкими, конкретными понятиями. Поэтому, если мы хотим использовать математические методы, то должны четко формулировать задачу; 2. Сильная продвинутость математических теорий (линейная алгебра, математический анализ, теория вероятностей, корреляционный и регрессионный анализ, дифференциальные уравнения и т.д.) предоставляет к нашим услугам очень мощный и развитый математический аппарат [1]. Представления о современных информационных технологиях с их возможностями оценивания и прогнозирования различных зависимостей и построения математических моделей необходимо для решения задач в области медицинской экологии, демографических процессов, состояния здоровья. Представлена краткая информация о тех видах многофакторного анализа данных, которые позволяют анализировать одновременно два и более признака. К самым популярным в области медицинской экологии методам многофакторного анализа данных можно отнести различные типы регрессионного анализа.

Статистический анализ медицинских данных - это не только (и не столько!) расчет каких-то характеристик (чисел) по имеющейся базам данных, сколько достижение понимания того, какие характеристики изучаемой системы связаны между собой, как они связаны и почему? [2].

Для исследования зависимостей одних признаков от других при анализе медицинских и биологических данных чаще всего используют различные виды математических моделей. Построение регрессионных моделей в биомедицинских

исследованиях позволяет оценить направленность, силу, вид связи, прогнозировать значения. Но статистическая модель не позволяет выявлять биологические закономерности, а может лишь имитировать "поведение" одного признака при известном "поведении" других признаков, являясь лишь инструментом, позволяющая избегать трудоемких и дорогостоящих натурных экспериментов [3].

Диапазон и масштаб моделируемых процессов крайне велик - от глобальной экологии до прогнозирования динамики отдельных компонентов, что требует использования различных подходов. Многие авторы выделяют статические и динамические модели [4,5]. Статические модели формализуют связь между показателями без учета переменной времени и строятся при допущении, что исследуемый процесс случаен и может быть изучен с помощью статистических методов анализа [6]. Динамические модели используются для оценки явлений в их развитии, что позволяет использовать их для прогноза состояний объектов, которые не наблюдали ранее [7]. Наиболее известны динамические модели накопления и распада поллютантов в агроэкосистемах: пестицидов, нефтепро-дуктов [8,9], радионуклидов [10-12].

Сам процесс моделирования, по И.Я. Лиепа [13], можно разделить на четыре этапа: качественный анализ, математическая реализация, верификация и изучение моделей. Первый этап моделирования - качественный анализ - является основой любого объектного моделирования. На его основе формируются задачи и выбирается вид модели. Второй этап моделирования - это математическая реализация логической структуры модели. Третий этап моделирования предусматривает верификацию модели: проверку соответствия модели оригиналу, т.е. насколько адекватно отражает особенности оригинала. Модель может быть признана высококачественной, если прогнозы оправдываются. Четвертый этап моделирования - это изучение модели, экспериментирование с моделью и предметная интерпретация модельной информации. Основная цель этапа - выявление новых закономерностей и исследование возможностей оптимизации структуры и управление поведением моделируемой системы.

При описании неопределенных процессов в природных системах (социально-гигинические условия, миграция и трансформация веществ в атмосфере, в почве, возникновение вспышек болезней, динамика рождаемости и смертности) рекомендуют использовать вероятностные подходы [14,15]. Необходимо отметить, что моделирование данных в зависимости от интенсивности или выраженности воздействующих факторов - одна из самых сложных задач статистического анализа в медицинских исследованиях.

Известно, что регрессионный анализ базируются на ряде довольно жестких предпосылок, из которых назовем 3 наиболее важных: 1) результаты наблюдений должны быть независимыми случайными величинами, и часто быть нормально распределенными; 2) выборочные оценки наблюдений должны быть однородны, т.е. не должны зависеть от величины результатов наблюдений; 3) ошибки в

определении независимых переменных должны быть пренебрежимо малы по сравнению с ошибкой в определении величины результатов наблюдений. Однако многие из них не всегда могут быть выполнены, и нигде нет указаний на то, к чему приводит нарушение этих требований при использовании стандартных статистических программ [16].

Линейное программирование – является наиболее простым и лучше всего изученным разделом математического программирования, что часто используется при решении задач взаимосвязи здоровья населения с загрязнением окружающей среды [17]. Более сложным для медиков и биологов является наиболее востребованные методы логистический анализ и анализ выживания. Анализ временных рядов - еще одна область применения статистических методов. Для прогноза периодических процессов по известному спектру частот используется Фурье-анализ [18]. Методы моделирования и прогнозирования временных рядов позволяют выявить тенденции изменения фактических значений параметра Y во времени и прогнозировать его будущие значения [19]. Однако хороших руководств по многомерному регрессионному анализу и моделированию в эпидемиологических и экологических исследованиях, доступных пониманию специалистам медицинского и биологического профиля, крайне мало. Еще меньше книг, в которых бы было уделено внимание решению задач многомерного статистического анализа в программе Statistica, встречающихся в медицинских и биологических исследованиях [20-22].

Всем эти методам анализа, выполняемых в модулях программы Statistica, на отдельных примерах будет уделено место в представленных методических рекомендациях. В освоении методов и понимании логики различных видов регрессионного анализа большое влияние оказали лекции старшего советника НИОЗ (г.Осло, Норвегия), профессора университета г.Тромсё (Норвегия) Гржибовского А.М., которому выражаем сердечную благодарность.

1 Линейный регрессионный анализ

Линейный регрессионный анализ представляет собой метод исследования статистической (регрессионной) зависимости между одной зависимой переменной (количественной) и двумя и более независимыми переменными (предикторам). Он позволяет по параметрам модели количественно оценить и прогнозировать влияние вредных факторов окружающей среды на показатели здоровья. В зависимости от природы зависимой переменной (Y), с ней связывается определенная модель распределения случайной величины, за счет чего регрессионный анализ, по типу математической зависимости, подразделяется на линейный и нелинейный, а в зависимости от числа независимых переменных в уравнении регрессии – на простой (один предиктор) и многофакторный. В многофакторном линейном анализе зависимая переменная (в нашем случае - показатель здоровья) должно быть количественной переменной с нормальным распределением.

В отличие от корреляционного анализа, который изучает направление и силу связи признаков, регрессионный анализ изучает *вид* зависимости признаков, т.е. параметры функции зависимости одного признака (зависимого, объясняемого, исхода, доли больных) от одного или нескольких других признаков (независимых, объясняющих).

В отличие от дисперсионного анализа, с помощью которого исследуется зависимость количественного признака от одного или нескольких качественных признаков, в линейном регрессионном анализе может исследоваться зависимость (количественного) признака от одного или нескольких количественных или качественных признаков. Он позволяет определить на сколько увеличится зависимая переменная от изменения зависимой переменной на одну единицу.

Целью линейного регрессионного анализа является поиск таких комбинаций независимых признаков, которые точнее, полнее (в статистическом смысле) оценивали и прогнозировали значение (вариабельность) зависимого признака от изменения независимых признаков [21].

Задачей регрессионного анализа является расчет значений одного объясняемого признака по значению ряда объясняющих признаков. В ходе выполнения анализа мы проверяем нулевую гипотезу об отсутствии связи между зависимой переменными и независимыми переменными. Нулевую гипотезу отклоняем и принимаем альтернативную гипотезу о существовании связи переменных при условии получения значений коэффициентов регрессии, соответствующих или превышающих заданный уровень значимости. Существуют как формальные (проверка гипотез), так и неформальные (изучение графика остатков) способы проверки моделей. Основные параметры модели регрессионного анализа представлены в таблице 1.

Таблица 1 - Форма предоставления данных по многофакторному регрессионному анализу из модулей программы Statistica 10.0

Показатель	Обозначение	Пример
Коэффициент детерминации процент вариабельности зависимой переменной, которая объясняется данной моделью (если признак один)	R^2 (RI)	RI=0,75
Скорректированный коэффициент детерминации после включения второго признака	R^2 (RI)	RI=0,82
Коэффициент Фишера показывает значимость отличий модели от среднего при достигнутом	F –	F=18,5
уровне значимости 0.05), о том, что модель предсказывает данные лучше, чем среднее арифметическое при уровне альфа-ошибки 5%		

Продолжение таблицы 1

Средняя ошибка аппроксимации разница между фактическим и прогнозируемым значением зависимой переменной	A (ε)	A=44,3	
Свободный член, константа, это уровень Y при x=0	b_0	b_0 =2,8 (p=0,01)	
Весовые коэффициенты или коэффициенты регрессии - число, которое показывает насколько увеличится Y при увеличении X на 1	b_1, b_2	b_1=3,51 b_2=- 1,5	
Beta — как стандартизованный вариант весового коэффициента регрессии позволяет сравнивать силу влияния изменения различных независимых переменных на зависимую между собой в модели и определять их доли влияния	Beta	$Beta_1$=12,32 $Beta_1$=2,25	
Формулы уравнения регрессии: 1 признак	$Y= b_0+ b_1X_1$	$Y=2,8+3,51*X_1$	
2 признака	$Y= b_0+ b_1X_1+ +b_2X_2$	$Y=2,8+3,51*X_1-1,5*X_2$	
Оценка модели - это вероятность принятия или непринятия нулевой гипотезы	p	p_1=0,045, p_2=0,035	
Число объектов исследования	N	N=150	
Результаты парного корреляционного анализа зависимых с независимыми и между собой (не более 0,9) для независимых признаков	корреляционная матрица	$r_{x1,y}$=0,79	*0,02*
		$r_{x2,y}$=0,25	*0,27*
		$r_{x1,x2}$=0,35	*0,051*
ПРИМЕР: Вывод регрессионного анализа - установлено, что уровень зависимого показателя (уровень гемоглобина крови у девушек-подростков) статистически значимо снижается при увеличении уровня независимого показателя (от концентрации пыли в атмосферном воздухе).			

Условия для выполнения линейного регрессионного анализа:

- число объектов наблюдения должно быть в несколько раз больше числа предикторов (объясняющих) признаков;
- наблюдения должны быть независимыми (от разных объектов);
- зависимая переменная должна быть количественная непрерывная;
- между зависимой переменной и каждой независимой переменной должна быть линейная зависимость;

- дисперсия каждой из независимых переменных должна быть более 0;

- между независимыми переменными не должно быть сильных связей (мультиколлинеарности);

- остатки должны быть независимыми (Добсона-Уотсона 1-3 допустимо, а 2–оптимально), иметь нормальное распределение и одинаковое рассеяние (плотность облака распределений) при любом предсказанном значении зависимой переменной (гомоскедаксичность). Если все эти условия не выполняются, то смысла в модели нет.

Построение уравнения регрессии сводится к оценке ее параметров. Оценку качества построенной модели даст коэффициент (индекс) детерминации, а также средняя ошибка аппроксимации. Средняя ошибка аппроксимации – среднее отклонение расчетных значений зависимой переменной по модели от фактических. Допустимый предел значений А – не более 8-10% [25]. Коэффициент детерминации показывает, какая часть дисперсии зависимого признака может быть объяснена дисперсией независимого признака. С его помощью оценивается модель, но нельзя говорить, что зависимость распространяется на 75% объема выборки.

Основное описание модели включает рассчитанную формулу с описанием основных характеристик статистической значимости модели:

$$Y= 2,8+3,51\ X_1- 1,5\ X_2 +\varepsilon \qquad (1)$$

$R^2= 75\%;\ F= 18,5;\ p<0,05.$

Весовые коэффициенты регрессионной модели следует проверить на статистическую значимость $b_0=2,8$ ($t=4,5$; $p<0,05$), $b_1=3,51$($t=8,3$; $p<0,04$) $b_2= - 0,16$ ($t=3,9$; $p<0,05$). Саму модель оценивают на значимость (при уровне значимости $p\leq0,05$) и информативность ($R^2>70\%$), проверяют ее работоспособность (*например,* при увеличении концентрации пыли на 1 мг/м3 снижение уровня гемоглобина в группе слесарей составило 0,5 ед., что не противоречит данным других исследователей); оценивают на точность (качество) и надежность (генерализация) прогноза (95% ДИ). На качество модели может оказать влияние мультиколлинеарность ($r>0,9$), определяется по величине критерия толерантности (до 1).

Для того чтобы проверить, соответствует ли модель имеющимся данным в целом и не была ли она сильно подвержена влиянию отдельных наблюдений, а также можно ли переносить результатов модели на другие выборки или всю популяцию (генерализация), проводится диагностика модели.

Диагностика модели проводиться по «выбросам» и по остаткам. «Выброс» – такое наблюдение, которое «не вписывается» в модель, но может влиять на коэффициенты регрессии. 95% стандартизованных остатков модели с уровнем значимости 5% лежат в границах -1.96 и 1.96, случаи более 3 СКО должны быть

удалены. *Случаи, оказывающие сильное влияние на модель, но которые* невозможно увидеть при оценке остатков, можно обнаружить по величине расстояния Кука для любого наблюдения. Если оно более 1, то такие случаи должны быть удалены.

Остатки должны иметь нормальное распределение с M=0 –график (прямая на нормальном вероятностном графике остатков); быть независимыми (показатель Добсона-Уотсона с уровнем 1-3) и иметь одинаковое рассеяние (плотность облака распределений) при любом предсказанном значении зависимой переменной (гомоскедаксичность) на графике 2-х мерной диаграммы рассеивания.

Возможности метода для нашего примера 1: Между напряженностью ЭМП и уровнем гемоглобина установлена сильная, прямая, значимая корреляционная связь (r_{x2y}=0,79, p<0,02), а между концентрацией пыли в атмосферном воздухе и уровнем гемоглобина установлена средняя, обратная, значимая корреляционная связь (r_{x2y}=-0,25; p<0,05). Толерантность 0,996 менее 1. Случаев, оказывающих сильное влияние на модель, нет*:* величина Кука для всех наблюдений менее 1. Остатки имеют распределение близкое к нормальному – график - прямая на нормальном вероятностном графике остатков); независимы (показатель Добсона-Уотсона 1,96) и имеют одинаковое рассеяние (плотность облака распределений) при любом предсказанном значении зависимой переменной (гомоскедаксичность) на графике 2-х мерной диаграммы рассеивания.

Теперь статистически установлено, что уровень гемоглобина Y (ед.) зависит от напряженности ЭМП (X_1, В/см^2) и запыленности (X_2, мг/м^3), что выражается моделью Y= 2,8+3,51 X_1- 1,5 X_2. Полученная модель высоко информативна (R^2=0,75), значима (p<0,05) и качественна (гомоскедаксична, мультиколлинеарность и выбросы отсутствуют). Снижение гемоглобина на 1 ед. может быть следствием увеличения уровня запыленности на 0,3 мг/м^3 или снижения напряженности ЭМП на 0,7 В/см^2. При увеличении концентрации пыли на 10 мг/м^3 уровень гемоглобина снизится на 15 ед., а при снижении ЭМП на 10 В/см^{2}уровень гемоглобина возрастет на 35,1 ед. Поскольку модель для параметра Y в высокой степени информативна, можно считать, что количество наблюдений в эксперименте вполне достаточно.

Даже если все условия выполняются, ещё нет гарантий, что модель может быть применена на популяционном уровне, т.е. генерализована. Чтобы это проверить проводят кросс-валидацию модели: разбиение массива данных на две половины случайным способом с последующим сравнением результатов регрессионного анализа, полученных в каждой половине. Чтобы предусмотреть точность прогноза, надо предусмотреть расчет Y от средних X и от отдельных значений X с последующим сравнением результатов регрессионного анализа.

Для полного предоставления результатов многофакторного линейного регрессионного анализа следует приводить последовательно в таблицах средние значения для всех независимых и зависимой переменных; в следующей таблице -

коэффициенты парной корреляции; и в последней таблице – результаты регрессионного анализа с оценкой прироста скорректированного на каждом шаге коэффициента детерминации.

2 Нелинейные регрессии

Нелинейные регрессии делятся на два класса: а) регрессии, нелинейные относительно включенных в анализ объясняющих переменных, и б) регрессии нелинейные по оцениваемым параметрам [25].

Регрессии, нелинейные по оцениваемым параметрам (или линейные относительно параметров модели):

-степенная $y = a + b_1 x + b_2 x^2 + b_3 x^3 + b_m x^m + \varepsilon$ (2);

-показательная $y = a^* \cdot e^{b x}$ (3);

-обратная $y = a + b^* 1/ x + \varepsilon$ (4);

-экспоненциальная $y = e^{a+bx}$ (5);

-линейно-логарифмическая $y = a + b \ln x + \varepsilon$ (6);

Регрессии, нелинейные относительно включенных в анализ объясняющих переменных:

- полиномы разных степеней $y=a+b_1x_1+b_2x_2+ b_3x_3+\varepsilon$ (7);

- логистическая,
- регрессия пропорциональных рисков по Коксу,

-равносторонняя гипербола $y= a+ b / x +\varepsilon$ (8).

- пробит-регрессия и другие.

Все нелинейные регрессии по оцениваемым параметрам могут быть преобразованы в линейные, для чего часто используется принцип замены:

- для степенной модели x^2 заменяют на X_1 x^3 на X_2 , а x^m на X_m и модель приобретает вид линейной модели, в которой возможно определение всех параметров и коэффициентов:

$y= a + b_1 x + b_2 x^2 + b_3 x^3 + b_m x^m + \varepsilon$; (9);

-для обратной модели величину 1/x заменяют на X* и работают с линейной моделью стандартного вида:

$$y=a+ b*X* + \varepsilon \quad (10);$$

- для линейно-логарифмической ln x заменяют на X* и работают с линейной моделью вида:

$$y= a+ b*X* + \varepsilon \quad (11);$$

- в экспоненциальной модели логарифмируют левую и правые части и модель приобретает вид: ln y= a+bx. После этого заменяют ln y на z, что позволяет определять все параметры и коэффициенты модели. Затем проводят обратные действия потенцирования и получают окончательный вид модели.

- в показательной модели использование натурального логарифмирования преобразует модель в следующий вид:

$$\ln y= \ln a + b x \quad (12).$$

Заменяют ln y на z, a ln a на f и получают линейную модель вида: z=f+ b*x. После расчета параметров модели проводят обратные действия потенцирования и замены для исходных зависимых и независимых факторов.

В программе STATISTICA v.10 предусмотрено моделирование с использованием различных степенных функций (x^2, x^3, ..x^n, x^{-2}, x^{-3}), обратной функции 1/x, логарифмической функции для натурального (ln x) и десятичного логарифма (log x) или экспоненциальными функциями с различным основанием (e^x или 10^x).

3 Логистическая регрессия

При изучении логистической регрессии мы исследуем взаимосвязь между бинарной (дихотомической) переменной отклика (зависимой переменной) и любыми независимыми переменными (количественный, номинальный, ранговый предиктор). В этом случае появляется возможность получить вероятность прогнозировать принадлежность к той или иной группе для каждого изучаемого случая в зависимости от известных переменных-предикторов. При построении зависимости мы можем спрогнозировать следующее: во сколько раз возрастет вероятность попадания в нужную группу («выжил») зависимая переменная (игрек) при изменении величины анализируемых независимых переменных (иксов). В большинстве исследований, в которых изучаемый признак является дихотомической величиной, логистический регрессионный анализ является одним из самых популярных множественных методов обработки данных. Математически это можно записать как уравнение вида:

$$Y_i = b_0+b_1X_1+b_2X_2+b_3X_3\ldots b_nX_n+\varepsilon_i \quad (13)$$

Здесь зависимая переменная Y не является непрерывной величиной, а принимает всего два возможных значения. Обычно единицей в этом случае представляют осуществление какого-либо события (успех), а нулем - отсутствие его реализации (неуспех). Среднее значение Y, обозначенное через P, есть доля случаев, в которых Y принимает значение 1. Математически это выражается как:

$$P(Y_i) = 1/(1+e^{-z}) \quad (14),$$

где $z=b_0+b_1X_1+b_2X_2+b_3X_3\ldots b_nX_n+\varepsilon_i$ .P(Y) – Вероятность возникновения события Y (вероятность принадлежности случая к определенной категории); e – основание натурального логарифма (~2.72)

В этом случае нам хотелось бы уметь оценивать величину P и определять факторы (независимые переменные Xi (непрерывные, ранговые), которые влияют на переменную Y. Каждый предиктор (X_1-X_n) имеет свой коэффициент (b_1-b_n).

В отличие от стандартной линейной регрессионной модели Y - зависимая непрерывная переменная представляет вероятность P, значения которой ограничены интервалом (0,1), а правая часть уравнения, напротив, может иметь значения, лежащие вне указанного выше интервала. Преобразование (логарифмирование модели) устранило эти противоречия. Поэтому уравнение логистической регрессии представляет собой уравнение линейной регрессии на логарифмичесой шкале (проблема линейной взаимосвязи решена). Ее параметры (коэффициенты) рассчитываются путем построения моделей, предсказывающих эмпирические данные исходя из имеющихся предикторов. Лучшая модель та, которая при включении всех рассчитанных параметров дает величины Y наиболее близкие к эмпирическим данным.

В линейной регрессии для определения насколько модель соответствует данным, рассчитывается коэффициент детерминации модели (SS_M / SS_T)*100% = R^2

В логистической модели используется (-2LL статистика), которая показывает количество информации, которое осталось после построения модели (большие числа указывают на модели, которые плохо подходят для имеющихся данных).

В линейной регрессии базовая модель – это модель, в которой используется среднее значение зависимой переменной, а в логистической регрессии – это наибольшая частота (заболеваний или излеченных лиц). Число степеней свободы (df) = $k_н$ – kб. Затем каждая последующая рассчитываемая модель будет сравниваться с этой моделью с помощью коэффициента χ^2 для (-2LL статистики):

$$\chi^2 = 2[LL(\text{новая}) - LL(\text{базовая})] \quad (15).$$

Оценка модели. В линейной регрессии оценка качества моделью производилась с помощью R^2, а в логистической ее аналогом – критерий отношения правдоподобия (в англоязычной литературе - Log-likelihood (-2LL), которая также указывает, насколько хорошо модель соответствует эмпирическим данным.

Для оценки предикторов в логистической регрессии используется критерий Вальда (Wald). При проверке нулевой гипотезы если $b \neq 0$, значит предиктор оказывает влияние на способность модели прогнозировать исход. Критерий Wald может увеличивать вероятность ошибки II типа при больших значениях коэффициентов регрессии. Полученые коэффициенты предикторов интерпретируют через понятие «шанс». Шансы на то, что событие произойдет, равны отношению вероятности того, что событие произойдет (P(Y) к вероятности того, что событие не произойдет (1-P(Y)):

$$\text{Шансы (Odds)} = P(Y)/(1-P(Y)) \qquad (16)$$

Можем сравнивать, как изменятся шансы, если предиктор вместо 1 примет значение 0. ОШ (отношение шансов) = (шансы в случае, когда предиктор =1) / (шансы в случае, когда предиктор =0). ОШ можно представить в виде e^b. При этом, если b больше 1, то вероятность события возрастает в разы, а если b меньше 1, то вероятность уменьшается в разы. Считаем так: во сколько раз уменьшается вероятность при увеличении значения переменной - предиктора на 1 единицу: (1/0,6=1,67) в 1,67 раз

Условия для выполнения логистического анализа:

- число объектов наблюдения должно быть в несколько раз больше числа предикторов (объясняющих) признаков;
- наблюдения должны быть независимыми (от разных объектов);
- зависимая переменная должна быть дихотомической (номинальной);
- выбросов не должно быть (не менее и не более 3,29δ) и расстояние Кука не должно быть более 1.

В тех случаев, когда исследования проводятся впервые по данной теме или когда исследователь хочет найти наилучшую модель для имеющихся данных, предпочтительными являются пошаговые методы ввода данных: последовательный ввод (forward) или последовательное исключение. Однако методы пошагового ввода несут риск ошибки II типа. Лучшим методом оценки является метод LR, чем Conditional или Wald

Интерпретация результатов множественного логистического регресссионного анализа:

- проверяем улучшается предсказательная способность модели при введении в нее независимых по χ^2 ;
- сравниваем R^2 (-2LL) для базовой модели (наибольшая частота) и новой модели с предикторами;

- получаем классификационную таблицу с чувствительностью и специфичностью;

- находим в таблице значений коэффициентов константу и коэффициенты регрессии со стандартным отклонением, оценкой Вальда, степенями свободы, и статистической значимостью, а также даны Exp (B) с 95%ДИ.

Пример: Увеличатся ли шансы сдать экзамен положительно, если предварительный экзамен сдан. b_0=-1,344, b_1=2,76. Исходя из формулы логистической модели $P(Y)= 1/(1+e^{-z})$ рассчитываем $z= b_0+b_1X$. Если b_1=0 (предварительный экзамен не сдан), то $z= b_0$, а $P(Y)= 1/(1+e^{-b0})$=0,21, где 1-P(Y)=0,79. Шансы сдать экзамен, если не сдан предварительный = 0,21/0,79=0,26. Или шанс не сдать экзамен при предварительно несданном 3,8 к 1 (0,79/0,21=3,8). Если b_1=1 (предварительный экзамен сдан), то $z= b_0 +b_1X$, а P(Y)= 0,81. 1-P(Y)=0,19. Шансы сдать экзамен при предварительно сданном = 0,81/0,19=4,3. Изменением шансов при изменении предиктора на 1 будет величина от 4,3/0,26=16,5, то есть предварительно сданный экзамен увеличивает шанс сдать основной экзамен в 16,5 раз.

Представление результатов множественного логистического регрессионного анализа включает : B, SE(B), Exp(B), R^2, 95% ДИ, таблица с нескорректированными и скорректированными Exp(B) и 95% ДИ, классификационная таблица. Используя классификационную таблицу, проверяем главные характеристики модели для популяции: чувствительность и специфичность

Таблица 2 - Классификационная таблица результатов

Наблюдения	Предикторы		
	да	нет	проценты
Нет	30	5	85,7
Да	7	33	82,5
Всего	37	38	84,0

Из таблицы делают следующие выводы:

- модель предсказывает правильно исход: в 84%;
- чувствительность модели: 33/40=82,5%;
- специфичность модели 30/35=85,7%;
- прогностическая ценность положительного результата: 33/38=86,8%;
- прогностическая ценность отрицательного результата: 30/37=81,1%.

Для особо важных предикторов – важнее оценивать процент чувствительности, а для оценки вмешательства - важнее оценивать процент специфичности. Но лучше, чтобы чувствительность и специфичность были в сумме максимальными.

Последовательность выполнения этапов логистического анализа:

1) получить данные описательной статистики – представить в отдельной таблице;

2) если есть пропуски и их много, то их заменить их на средние, если их мало, то исключить;

3) по $\chi 2$ сравнить влияние всех категорированных переменных в наблюдениях;

4) построить логистические модели и по полученному значению b=2.76 определить величину $e^{2,76}$ =15,8. Это значит, что в 15,8 раз увеличиться шанс сдать экзамен. ДИ не должен включать 1, чтобы быть значимым.

5) оценить остатки.

M.H.Katz. предлагает: Когда переменных много (больше 20), то сначала оценить их влияние по $\chi 2$, выбрать только те, которые связаны с исходом при p <0,15. Если значения в бивариантном анализе хороши, но есть сочетанный эффект, т.е. воздействие в присутствии других - его пересчитать и сравнить. Отбор предикторов всегда сложный и обсчет в программе тоже сложный.

Заключение. Использование метода логистической регрессии возможно в компьютерных статистических пакетах. В них для получения коэффициентов логистической регрессии применяется метод максимального правдоподобия. Возможно, решения, как простой, так и множественной логистической регрессии (число независимых переменных два и более).

4 Анализ временных рядов

Одним из методов статистического анализа при обработке данных о среде и здоровье населения является анализ временных серий (рядов). Временной серией (рядом) называется последовательность наблюдений, упорядоченная по времени. В отличие от анализа случайных выборок, анализ временных рядов основывается на предположении, что последовательные значения в файле данных наблюдаются через равные промежутки времени (тогда как в других методах нам не важна и часто не интересна привязка наблюдений ко времени). В этом смысле после-довательные наблюдения будут зависимы друг от друга. Графически временной ряд принято представлять в форме точечной диаграммы, при этом при построении каждой точки на диаграмме время (t) - независимая переменная откладывается по оси X, а наблюдаемая величина - зависимая переменная - по оси Y.

Цель анализа временных рядов: определение природы ряда и предсказание будущих значений временного ряда по настоящим и прошлым значениям. Для этого нужно идентифицировать и описать модель. Как только модель определена, можно оценивать независимые переменные и предсказать его будущие значения

Анализ зависимости здоровья населения от вредных факторов окружающей среды методом временных серий сводится к установлению регрессионной зависимости между показателем здоровья (переменной Y(t)) и величинами, от которых зависит рассматриваемый показатель здоровья - независимыми переменными, описывающие состояние окружающей среды или условия жизни (Xi(t)),

либо установлению связи между двумя рядами данных, отслеженных за аналогичные промежутки времени.

Временной ряд имеет свои особенности дополнительные статистики: основная величина – это время, за которое должны быть получены анализируемые зависимые и независимые переменные. Временной интервал может быть любым: день, неделя, месяц, сезон, год.

Большинство регулярных составляющих временных рядов принадлежит к двум классам: они являются либо трендом, либо сезонной составляющей. Тренд представляет собой общую систематическую линейную или нелинейную компоненту, которая может изменяться во времени. Сезонная составляющая - это периодически повторяющаяся компонента. Тип модели временного ряда, в которой амплитуда сезонных изменений увеличивается вместе с трендом, называется моделью с мультипликативной сезонностью.

Если временные ряды содержат значительную ошибку, то первым шагом выделения тренда является сглаживание.

Анализ временных рядов предполагает, что данные содержат систематическую составляющую (обычно включающую несколько компонент) и случайный шум (ошибку), который затрудняет обнаружение регулярных компонент. Поэтому при обработке данных эпидемиологических исследований много внимания уделяется сглаживанию, как определенному виду фильтрации данных. Цель сглаживания - сильнее выявить тренд, то есть обеспечить более ясный обзор действительного поведения статистической переменной Y. Чаще всего используется сглаживание *методом скользящей средней,* скользящей взвешенной средней, скользящей медианы и другие.

Все методы сглаживания отфильтровывают шум и преобразуют данные в относительно гладкую кривую, пригодную для анализа. Многие монотонные временные ряды можно хорошо приблизить линейной функцией. Если же имеется явная монотонная нелинейная компонента, то данные вначале следует преобразовать, чтобы устранить нелинейность. Обычно для этого используют логарифмическое, экспоненциальное или (менее часто) полиномиальное преобразование данных.

Периодическая и сезонная зависимость может быть формально определена как корреляционная зависимость порядка k между каждым i-м элементом ряда и (i-k)-м элементом (Kendall, 1976). Ее можно измерить с помощью автокорреляции (т.е. корреляции между самими членами ряда), где k обычно называют лагом (сдвиг, запаздывание). Если ошибка измерения не слишком большая, то сезонность можно определить визуально, рассматривая поведение членов ряда через каждые k временных единиц.

Сезонные составляющие временного ряда могут быть найдены с помощью коррелограммы, которая показывает численно и графически автокорреляционную функцию (АКФ), или коэффициенты автокорреляции (и их стандартные ошибки)

для последовательности лагов из определенного диапазона ряда. Ее надежность определяется диапазоном в размере двух стандартных ошибок на каждом лаге, а сила - величиной автокорреляции. Поскольку автокорреляции последовательных лагов формально зависимы между собой, то после удаления автокорреляций первого порядка (после взятия разности с лагом 1) ряд станет более стационарным, что необходимо для применения АРПСС.

Процедуры фильтрации проводятся перед построением регрессионной модели. При этом можно проводить как фильтрацию ряда зависимых, так и рядов независимых переменных.

Зависимая переменная может быть количественной или дискретной (число случаев), исходя из чего, можно использовать различные модели: регрессия Пуассона или биноминальные. Регрессия Пуассона проводится с инфляцией нулей и позволяет получить порог изменения (p).

В зависимости от природы переменной Y, задающей результат заболевания, с ней связывается определенная модель распределения случайной величины. В эпидемиологических исследованиях нашли применение следующие виды распределений: Гаусовское (нормальное) распределение (используется при анализе больших выборок количественных наблюдений), пуассоновское, экспоненциальное, лог-нормальное и ряд других. До того, как начать оценивание, необходимо решить, какой тип модели будет подбираться к данным, и какое количество параметров присутствует в модели, иными словами, нужно идентифицировать модель АРПСС. Основными инструментами идентификации порядка модели являются графики, автокорреляционная функция (АКФ), частная автокорреляционная функция (ЧАКФ).

Метод АРПСС используется для идентификации модели (выбор типа модели и числа параметров по АКФ и ЧАКФ) и ее оценки (квазиньютоновский алгоритм максимизации правдоподобия (вероятности) наблюдения значений ряда по значениям параметров). Для всех оценок параметров вычисляются так называемые асимптотические стандартные ошибки.

Качество модели определяют по значению t статистики, она должна давать точный прогноз, быть экономной и иметь независимые остатки, Хорошей проверкой модели являются: (a) график остатков и изучение их трендов, (b) проверка АКФ остатков.

Модель АРПСС является подходящей только для рядов, которые являются стационарными (среднее, дисперсия и автокорреляция примерно постоянны во времени); для нестационарных рядов следует брать разности. Рекомендуется иметь, как минимум, 50 наблюдений в файле исходных данных.

В эпидемиологических исследованиях представляет интерес анализ распределенных лагов как специальный метод оценки запаздывающей зависимости между рядами. В эконометрике часто возникают такого рода зависимости с запаздыванием. Например, доход от инвестиций в новое оборудование отчетливо

проявится не сразу, а только через определенное время. Более высокий доход изменяет выбор жилья людьми; однако эта зависимость, очевидно, тоже проявляется с запаздыванием. Во всех этих случаях, имеется независимая или объясняющая переменная, которая воздействует на зависимые переменные с некоторым запаздыванием (лагом). Метод распределенных лагов позволяет исследовать такого рода зависимость.

Обобщенная регрессионная модель для анализа временных рядов в медицинской экологии может быть представлена следующим образом. Например, число случаев заболевания, вызванной набором факторов окружающей среды, имеет среднюю, равную E(Y). Эти переменные измеряются несколько раз в течение определенного отрезка времени. Тогда l{E(Y)} = r (X), где: l - некоторая связующая функция, преобразующая шкалу, в которой измеряется средняя E(Y), это может быть линейное уравнение:

$$Yt = i*Xt-i \qquad (17)$$

В этом уравнении значение зависимой переменной в момент времени t является линейной функцией переменной X, измеренной в моменты t, t-1, t-2 и т.д. Таким образом, зависимая переменная представляет собой линейные функции X и X, сдвинутых на 1, 2, и т.д. временные периоды. Бета коэффициенты (i) могут рассматриваться как параметры наклона в этом уравнении. Если коэффициент переменной с определенным запаздыванием (лагом) значим, то можно заключить, что переменная Y предсказывается (или объясняется) с запаздыванием. Для устранения такой проблемы множественной регрессии, как мультикол-линеарность, Алмон (1965) предложил оценивать коэффициенты альфа, что уменьшает мультиколлинеарность (этот метод оценивания коэффициентов бета называется полиномиальной аппроксимацией).

В случае преобразования l, как связующей функции в логарифмическую шкалу для E(Y), то можем экспоненциальное уравнение типа:

$$Ln (n) = a + b X \qquad (18)$$

где ln (логарифм); r (X) - линейное уравнение, связывающее переменные регрессии X.

Согласно модели, где зависимая переменная находится под логарифмом, то можно сделать вывод о том, на сколько % увеличиться риск возникновения события при увеличении X на одну единицу.

Построив модель, можем определить, как измениться зависимая (заболваемость, смертность) при увеличении температуры на 1 оС или при увеличении запыленности на 1 мг/м3. По результату анализа мы получаем величину исходных данных и дополнительно лаг. Лаг показывает, на сколько увеличивается частота

случаев заболеваний при увеличении независимой переменной на единицу. Например, если величина лага 1,15, то значит, число случаев увеличивается на 15%, что и является результатом модели.

В модуле «Анализа временных рядов» встроен кросс-спектральный анализ, который позволяет анализировать одновременно два ряда данных. При этом анализе взаимосвязь между двумя рядам обнаруживаются по величине корреляции для периодичности, которые присутствуют в обоих анализируемых рядах. В таблицах представляются результаты из кросс-спектрального анализа (независимой (X) переменной 1 и зависимая (Y) переменной 2, с указанием метода сглаживания). Указывают общие для обоих рядов основные периодичности на частотах (на частоте 0625 и 1875).

Пример: Ежедневные данные по качеству атмосферного воздуха и смертности населения гг. Екатеринбурга и Нижнего Тагила за 1995-1997гг. были проанализированы методом временных серий. Цель анализа - установить зависимость между ежедневной смертностью от сердечно-сосудистых и респираторных заболеваний и загрязнением воздуха (фенол, тонкая фракция пыли, формальдегид, NO_2, CO). Ежедневные показатели смертности и концентрации загрязняющих веществ имели месячные и сезонные колебания. Для их устранения ряды были сглажены фильтром Шамвэя (Shumway, 1998),который позволял удалить долгие циклы и шум, а короткие изменения (длительностью в несколько дней) оставить практически без изменения. Согласно методу анализа временных рядов для каждого значения переменной вычислялась взвешенная скользящая средняя за 19 дней, которая далее вычиталась из самого же этого значения, а для каждого из 19 значений Yt=t+i (i = -9...,0,...9), участвующих в вычислении средней, Шамвэй вычислил специальные весовые коэффициенты, которые определяли вклад значения Yt за определенный день в среднее значение и, таким образом, наилучшим образом отсеивают долгие циклы. При проверке рядов на автокорреляцию оказалось также, что примененный фильтр эффективно ее устраняет. Профильтровали ряд зависимой переменной (показатель ежедневной смертности) и ряды независимых переменных (концентраций загрязняющих веществ).

В виду того, что по некоторым видам смертности показатели были небольшие, регрессионный анализ временных рядов проводился исследователями из предположения, что переменные модели имеют либо нормальное распределение, либо распределение Пуассона. Проверялись зависимости смертности от уровня загрязненности воздуха в этот же день, предыдущий день (единичный лаг) и за два дня (двойной лаг), при этом формировалась многорегрессионная модель. Обе модели дали аналогичные результаты. Непротиворечивость результатов и нормальность распределения ошибки контролировалась.

Анализ показал статистически значимую зависимость между содержанием в воздухе респирабельных фракций пылевых частиц и смертностью, как от респираторных, так и сердечно-сосудистых заболеваний. В частности, было

определено: 11-16% смертей с диагнозом острое респираторное заболевание ассоциировалось с повышенным содержанием твердых пылевых частиц в воздухе в предыдущий день в г.Екатеринбурге, около 5-9% смертей с тем же диагнозом было вызвано высоким содержанием твердых пылевых частиц в воздухе г.Нижний Тагил за текущий день, 2% смертей с диагнозом сердечно-сосудистое заболевание ассоциировалось с высоким содержанием твердых пылевых частиц в воздухе г. Нижний Тагил за предыдущий день и фенола за текущий день.[24].

5 Анализ выживаемости

В анализе выживаемости также как и анализе временных рядов используются параметр время, однако, если в анализе временных рядов происходит поиск закономерностей (автокорреляций, трендов, сезонов) изменения во времени анализируемого фактора, то в анализе выживаемости – поиск числа выбываний или включения на конкретных временных точках. Именно при развитии медицинских и биологических исследований сформировался комплекс методы, которые позже стали применяться в социальных, экономических и технических науках. Для решения конкретной задачи, сохранения для анализа максимальной собранной информации, часть которой могла быть потеряна до завершения эксперимента (в результате утраты связи с кем-то из пациентов), было введено понятие о наблюдениях, которые содержат неполную информацию, которые и называют цензурированными наблюдениями. Цензурированные наблюдения в социальных науках позволяет изучать интенсивность выбытия студентов из высшего учебного заведения (времен до выбытия, и в конце периода наблюдения некоторые студенты продолжают учебу, а данные об этих студентах являются цензурированными. Мы делаем выводы, не дожидаясь того момента, когда все выбранные студенты покинут учебное заведение.

Отличительными особенностями этого метода является то, что для всех наблюдений известно время начала наблюдения и время окончания наблюдения, а так же статус (умер или выбыл) наблюдаемого лица. Выбор наблюдаемых лиц произведён случайно. В модуле анализа выживаемости используется так называемая функция выживания, представляющая собой вероятность того, что объект проживет время больше t. Время здесь является непрерывной переменной, допускается его любая размерность: год, месяц, неделя, день. В тексте при описании эксперимента (наблюдения) необходимо указывать даты начала и окончания наблюдения, причина окончания (до летального исхода или плановое окончание).

При создании базы данных каждому пациенту присваиваем не только идентификационный номер, но и статус (дихотомическая переменная): 1- умер, заболел, утратил какое-либо качество (это событие, которое изучаем) или 0- за время наблюдений изучаемое событие не произошло (это цензурированное событие). Для каждого пациента надо указать причину цензурированного события:

- выбыл (запишем, когда и почему выбыл);
- умер, но от других причин;

- событие не произошло за время наблюдения.

5.1 Анализ Каплана-Майера

Анализ Каплана-Майера часто используют в эксперименте или в рандомизированных клинических исследованиях. Для решения эпидемиологических задач у населения, проживающего в различных условиях, при воздействии комплекса внешних факторов, использование отдельных приемов может быть полезно. Когорты для анализа набирают либо всех одновременно, либо с последовательным набором пациентов.

Процедура оценивание функции выживания Каплана-Мейера с подгонкой распределения выживаемости, построением таблиц времен жизни являются описательными методами. Таблицу времен жизни можно выживаемости можно рассматривать как "расширенную" таблицу частот, где область возможных времен наступления критических событий (смертей и др.) разбивается на некоторое число интервалов. Для каждого интервала вычисляется число и доля объектов, которые в начале рассматриваемого интервала были "живы", число и доля объектов, которые "умерли" в данном интервале, а также число и доля объектов, которые были изъяты или цензурированы в каждом интервале. По этим частотам вычисляют долю умерших и долю выживших, кумулятивную долю выживших (иначе называют функцией выживания), плотность вероятности, функцию интенсивности и медиану ожидаемого времени жизни.

Доля умерших – отношение числа объектов, умерших в соответствующем интервале, к числу объектов, изучаемых на этом интервале. Доля выживших – доля равна единице минус доля умерших. Кумулятивная доля выживших – кумулятивная доля выживших к началу соответствующего временного интервала. Поскольку вероятности выживания на разных интервалах считаются независимыми, эта доля равна произведению долей выживших объектов по всем предыдущим интервалам. Полученная доля как f(t) называется выживаемостью или функцией выживания (оценка функции выживания).

Расчет кумулятивной выживаемости производят по формуле:

$$F(t)=\Pi\,[1- n_i /N_i] \qquad (19)$$

где n1 –текущее событие, а N1- количество событий на начало наблюдения (n1=1, n2=1, n3=1 и т.д., а N1=7, N2=6, N=5; П- произведение.

Плотность вероятности – оценка вероятности отказа в соответствующем интервале, определяемая таким образом:

$$Fi = (Pi – P\,i +1) / hi \qquad (20)$$

где Fi - оценка вероятности отказа в i-ом интервале, Pi - кумулятивная доля выживших объектов (функция выживания) к началу i-го интервала, hi - ширина i-ого интервала.

Функция интенсивности определяется как вероятность того, что объект, выживший к началу соответствующего интервала, умрет в течение этого интервала. Оценка функции интенсивности вычисляется как число смертей, приходящихся на единицу времени соответствующего интервала, деленное на среднее число объектов, доживших до момента времени, находящегося в середине интервала. Медиана ожидаемого времени жизни - это точка на временной оси, в которой кумулятивная функция выживания равна 0.5. Принято считать 25- и 75-процентили) кумулятивной функции выживания. Медиана кумулятивной функции выживаемости совпадает с точкой выживания 50% выборочных наблюдений только в случае, когда за прошедшее к этому моменту время не было цензурированных наблюдений.

Числом изучаемых объектов считают то число объектов, которые были "живы" в начале рассматриваемого временного интервала, минус половина числа изъятых или цензурированных объектов. Чтобы получить надежные оценки трех основных функций (функции выживания, плотности вероятности и функции интенсивности) и их стандартных ошибок на каждом временном интервале, рекомендуется использовать не менее 30 наблюдений.

В результатах следует указать среднее и медианное время выживаемости с доверительными интервалами (среднее±1,96*стандартная ошибка). Медиану часто нельзя посчитать, если умерло меньше половины людей. Таблица выживаемости в результатах получается громоздкой. В данном случае нагляднее графики кумулятивной выживаемости.

График кумулятивной выживаемости (survival function) - это график со снижением частоты встречаемости лиц с определенными качествами (отмеченными как 1). Можно построить и обратный график, где кумулятивная выживаемость (событие) будет нарастать, отражая частоту накопления лиц с определенными качествами (с заболеваниями).

График функции умирания (Hazard function) отражает функцию интенсивности риска или скорости умирания. Функция интенсивности риска, h(t) – это вероятность того, что субъект, выживший к началу соответствующего интервала, умрет в течение этого интервала. Этот показатель используется для сравнения моделей, где фактор риска есть, и где его нет. Интенсивность риска может быть больше1, не следует путать его с вероятностью. Общий вид представлен в формуле:

$$h(t)=\lim_{\Delta t\to 0}\frac{R(t)-R(t+\Delta t)}{\Delta t\cdot R(t)}. \qquad (21)$$

В модуле «Анализ выживаемости» имеется пять критериев для сравнения цензурированных данных: логарифмический ранговый критерий, критерий Вилкоксона, обобщенный Геханом, критерий Вилкоксона, обобщенный Пето и F-

критерий Кокса. Большинство этих критериев приводят соответствующие z-значения (значения стандартного нормального распределения); эти z-значения могут быть использованы для статистической проверки любых различий между группами. В SPSS кроме логрангового критерия используются критерий Бреслоу и критерий Тарона-Варе.

Сравнение выживаемости в двух группах. Если все случаи (события) имеют одинаковый вес, независимо от времени наступления события (во времени), то применяют логранговый критерий или критерий Кокса-Ментела которые являются наиболее мощным (безотносительно к цензурирования), обычно начинают с него. Известно, что F - критерий Кокса более мощный, чем критерий Вилкоксона – Гехана в случаях, если выборочные объемы малы (то есть, объем группы меньше 50), если выборки извлекаются из экспоненциального распределения или распределения Вейбулла, а цензурированных наблюдений нет.

Пример. Для сравнения каждой пары изменений с помощью логрангового критерия в случае событий 1, 2, 3 мы получим: между 1 и 2 – z= 0,041(>0,017 – различий нет), между 2 и 3 - z= 0,074 (>0,017 - различий нет), а в паре между 1 и 3 z=0,015 (< 0,017 - есть различия). Для нескольких попарных сравнений необходима поправка Банферрони, которая повышает требования к статистической значимости 95% для анализируемой разницы с 0,05 до 0,017= 0,05/3.

Сравнение выживаемости в трех и более группах (многовыборочный критерий). В случае сравнения выживаемости в трех и более группах вначале используют критерий Вилкоксона, обобщенный Геханом, критерий Вилкоксона, обобщенный Пето или логранговый (омнибусный) критерий. При этом каждому времени жизни приписывается его вклад в соответствии с процедурой Ментела; далее на основе этих вкладов (по группам) вычисляется значение статистики-χ^2. В дальнейшем для попарного сравнения выживаемости критерию используют критерий Вилкоксона, обобщенному Геханом. В продолжении нашего примера: для оперативной оценки можно воспользоваться оценкой разницы для 3 групп по логранговому критерию: $\chi 2$ =0,32, p=0,57, что требует принять нулевую гипотезу H_0, то есть различий среди трех групп не выявлено. На графике можно получить 2, 3 и более кривых выживаемости, а также таблицы средних и медианных времени выживания.

Если сравниваются две или более группы, то важно проверить доли цензурированных наблюдений в каждой, поскольку эти различия могут привести к смещению в статистических выводах. В медицинских исследованиях степень цензурирования может зависеть, например, от различий в методе лечения: пациенты, которым стало много лучше или стало хуже, с большой вероятностью теряются из наблюдения.

Для оценки линейного тренда направления увеличения риска можно посчитать ранговые переменные как количественные, но использовать только в качестве результата направленность для описания их соотношения.

Среди недостатков анализа выживаемости методом Каплана-Майера следует указать на то, что группировочная переменная и независимая переменная (X) могут быть только качественной переменной; возможно проведение только бивариантных сравнений, что повышает риск ошибки 1 рода при проведении большого количества попарных сравненений и стратификации; при сравнении между группами не предусмотрено показывать результаты в виде отношения рисков; не предусмотрено проводить коррекцию на конфаундеры (смешивающие факторы).

5.2 Метод таблиц дожития

Метод таблиц дожития (актуарный) в анализе выживаемости проще понять и объяснить результаты, но у него ниже мощность и он менее чувствительный метод, чем метод Каплан-Мейера.

Таблицы дожития (life-table method) основываются на разбивке времени наблюдения на равные интервалы, которым не всегда можно найти аналоги (декада, неделя и т.п.) в клинических исследованиях, чаще применяют в крупных демографических исследованиях. Все расчеты проводятся для каждого временного интервала, а сравнение выживаемости проводят с помощью критерия Вилкоксона-Гехана. Программа учитывает количество лиц, вступивших в каждый интервал времени, число выбывших (цензурированных) и число изучаемых исходов, из чего рассчитывается вероятность изучаемого исхода, вероятность того, что исход не наступит, а также кумулятивная выживаемость и функция интенсивности риска.

В общем случае метод таблиц жизни (дожития) дает хорошее представление о распределении смертей объектов во времени, но для прогноза часто надо знать форму рассматриваемой функции выживания. Для описания продолжительности жизни важны такие семейства распределений, как экспоненциальное распределение, распределение Вейбулла экстремальных значений и распределение Гомперца.

Процедура оценивания параметров выбранного распределения использует алгоритм метода наименьших квадратов, а в случае преобразованных моделей - оценки взвешенных наименьших квадратов двух типов. Зная параметрическое семейство распределений, можно вычислить функцию правдоподобия по имеющимся данным и найти ее максимум. Такие оценки называются оценками максимального правдоподобия.

В модуле можно строить графики как эмпирических, так и теоретических функций распределения и интенсивности, что позволяет собой прекрасное средство проверки.

Преимущество метода Каплана-Мейера (по сравнению с методом таблиц жизни) состоит в том, что оценки не зависят от разбиения времени наблюдения на интервалы, т.е. от группировки. Метод множительных оценок Каплана-Мейера и метод таблиц времен жизни приводят, по существу, к одинаковым результатам

только в том случае, если временные интервалы содержат, максимум, по одному наблюдению.

Для полного предоставления результатов анализа нужны таблицы и графики выживаемости, результаты их сравнения по Каплану-Мейеру, таблицы дожития, где приводят коэффициент Вилкоксона-Гехана с оценкой значимости (например, коэффициент Вилкоксона равен 0,16; статистическая значимость p=0,699; вывод - различий выживаемости в сравниваемых группах нет).

Последовательность выполнения анализа выживаемости:

1) определить долю выживших по методу Каплана-Мейера;

2) вычислить кумулятивную долю выживших;

3) рассчитать ошибки;

4) определить плотности вероятности – таблицы дожития;

5) провести сравнительную оценку групп по Вилкоксону- Гехану.

6 Анализ пропорциональных рисков Кокса

Анализ выживаемости содержит анализ регрессионных моделей для оценивания зависимостей между многомерными непрерывными переменными со значениями типа времена жизни. Выяснение того, являются ли некоторые непрерывные переменные связанными с наблюдаемыми временами жизни, является важной задачей медицинских и биологических статистических исследований. Классическая техника множественной регрессии в таких исследованиях не может быть использована, поскольку времена жизни не являются нормально распределенными (экспоненциальное распределение или распределение Вейбулла), а наблюдении являются не завершенными, а цензурированные.

В многомерной линейной регрессии результирующей является непрерывная переменная с нормальным распределением, а предикторами – любые переменные (цензурирование невозможно); в многомерной логистической регрессии результирующей является бинарная (дихотомическая) переменная, а предикторами – также любые переменные (цензурирование невозможно), в регрессия Кокса результирующей является непрерывная переменная со значениями типа времена жизни, а предикторами также могут быть любые переменные с введением цензурирования.

В регрессии Кокса независимые переменные могут быть любого типа, учитывается абсолютное время наблюдения, что, позволяет оценивать влияние факторов риска на результирующую, проводить коррекцию модели на конфаундеры (смешивающие факторы), снижает риск ошибки 1 рода при проведении большого количества попарных сравнений, представлять результаты сравнений между группами в виде отношения интенсивности рисков (Hazard ratio, HR).

Модель пропорциональных интенсивностей Кокса - это наиболее общая регрессионная модель, поскольку она не связана с какими-либо предположениями относительно распределения времени выживания. Эта модель предполагает, что функция интенсивности имеет некоторый уровень Y, являющийся функцией

независимых переменных (X). Никаких предположений о виде функции интенсивности не делается. Модель может быть записана в следующем виде:

$$h\{(t), (z_1, z_2, ..., z_m)\} = h_0(t)*\exp(b_1*z_1 + ... + bm*zm) \quad (22)$$

где h(t,) обозначает результирующую интенсивность, при заданных для соответствующего наблюдения значениях m ковариат (z_1, z_2, ..., z_m) и соответствующем времени жизни (t).

Множитель $h_0(t)$ называется базовой функцией интенсивности, она равна интенсивности в случае, когда все независимые переменные равны нулю. Можно линеаризовать эту модель, поделив обе части соотношения на $h_0(t)$ и взяв натуральный логарифм от обеих частей:

$$\log [h\{(t), (z...)\} / h_0(t)] = b_1*z_1 + ... + b_m*z_m \quad (23)$$

Получив "простую" линейную модель, которая легко поддается изучению. Модельное уравнение пропорциональных интенсивностей Кокса подразумевает два предположения: 1) зависимость между функцией интенсивности и логлинейной функцией ковариат является мультипликативной, что означает, что для двух заданных наблюдений отношения их функций интенсивности не зависит от времени, и 2) соотношения между функцией интенсивности и независимыми переменными логлинейно.

Модель пропорциональных интенсивностей Кокса с зависящими от времени ковариатами, которые могут быть категориальными (групповыми) переменными (например, пациент 1 прооперирован, а пациент 2 - нет. Согласно предположению пропорциональности риски для пациентов не меняются на протяжении наблюдения, но ясно же, что сразу после операции риск прооперированного пациента выше, с течением времени убывает и становится меньше риска, не оперированного пациента). В этом случае группирующая является ковариатой, зависящей от времени (пропорциональность рисков нарушена).

Проверка на нарушение пропорциональности, доказательство, что ковариант зависит от времени можно выполнить по параметру b_2 в модели:

$$h(t,z) = h_0(t)*\exp\{b_1*z + b_2*[z*\log(t)-5.4]\} \quad (24)$$

Если параметр b_2 статистически значим (например, если он, по крайней мере, в два раза больше своей стандартной ошибки), то можно сделать вывод, что ковариата z действительно зависит от времени, и поэтому предположение пропорциональности не выполняется. Обратите внимание, что функция интенсивности в момент t есть функция: (1) базовой функции интенсивности $h_0(t)$, (2) ковариаты z и (3) z-кратного логарифма времени.

Экспоненциальная регрессия предполагает, что распределение продолжительности жизни S(z) является экспоненциальным и связано со значениями некоторого множества независимых переменных (z_i):

$$S(z) = \exp(a + b_1*z_1 + b_2*z_2 + ... + b_m*z_m) \quad (25)$$

Оценка модели (влияние независимых переменных на время жизни значимо) проводится по критерию $\chi 2$, где сравнивается функция логарифма правдоподобия для модели со всеми оцененными параметрами (L_1) и функция логарифма правдоподобия модели, в которой все ковариаты обращаются в 0 (L_0). Если значение $\chi 2$ статистически значимо, то нулевую гипотезу отвергаем. Предположения экспоненциальности проводят по графику остатков времен жизни (сравнение их со значениями стандартных экспоненциальных порядковых статистик альфа).

Нормальная и логнормальная регрессия, где времена жизни (или их логарифмы) имеют нормальное распределение, идентична обычной модели множественной регрессии:

$$t = a + b_1*z_1 + b_2*z_2 + ... + b_m*z_m \quad (26)$$

Здесь t означает время жизни. Если принимается модель логнормальной регрессии, то t заменяется ln t. Модель нормальной регрессии особенно полезна, поскольку часто данные могут быть преобразованы в нормальные за счет применения нормализующих аппроксимаций. Таким образом, в некотором смысле это наиболее общая параметрическая модель (в противоположность модели пропорциональных интенсивностей Кокса, которая является непараметрической), оценки которой могут быть получены для большого разнообразия исходных распределений времен жизни. Оценка модели проводится по критерию $\chi 2$, как и в предыдущем случае.

Стратифицированный анализ позволяет проверить гипотезу о том, что зависимость между выживаемостью и регрессорами одна и та же для разных групп данных. При стратифицированном анализе строят регрессионные модели отдельно для каждой группы и общую для данных из двух групп. Оценка статистической значимости различий между группами ($\chi 2$) проводится по разности логарифмов правдоподобия для исходной и вновь рассчитанной модели.

Модель пропорциональных интенсивностей Кокса может быть использован для наших исследований как способ анализа рисков через функцию интенсивности риска (hazard function).Функция интенсивности риска является произведением двух факторов: базового риска и линейной экспоненциированной функции всех предикторов. Базовый риск для каждого участника исследования свой, а в момент времени i риск может быть представлен как:

$$h_i(t) = \lambda_0(t) \cdot e^{\beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} + ... + \beta_k \cdot X_{ik}} \quad (27)$$

Иначе говоря, при использовании этого вида анализа можно изучать отношение интенсивности рисков, и по ним сравнивать группы, поэтому нужна величина Exp (B). 1- референтная, а второй и третий факторы сравниваются с ней. Основываясь на функции интенсивности рисков, можем рассчитать вероятность достижения события.

При этом оценки значения функции интенсивности риска и оценки коэффициентов для каждого предиктора (λ) анализировать нет необходимости, важно определить влияние фактора на исход $H(t)=e^{\beta 1}$, помня, что вместо свободного члена в модели выступает базовый риск, т.е. вероятность отнесения изучаемого человека к конкретной группе.

Расчет относительных рисков (HR) проводим по формуле:

$$H(t) = \frac{h_i(t)}{h_j(t)} = \frac{\lambda_0(t) \cdot e^{\beta_1 \cdot x_{i1} + \beta_2 \cdot x_{i2} + ... + \beta_k \cdot x_{ik}}}{\lambda_0(t) \cdot e^{\beta_1 \cdot x_{j1} + \beta_2 \cdot x_{j2} + ... + \beta_k \cdot x_{jk}}} \quad (28)$$

$$H(t) = \frac{e^{\beta_1 \cdot x_{i1}}}{e^{\beta_1 \cdot x_{j1}}} = e^{\beta_1 (x_{i1} - x_{j1})} \quad (29)$$

Если предиктор x_1 изменяется на 1, то можно использовать упрощенную формулу:

$$H(t) = e^{\beta_1} \quad (30)$$

Относительный риск рассчитывается из оцененных параметров для каждого фактора, а его значимость представляется с помощью доверительных интервалов и оценивается статистическим критерием Вальда. Отношение рисков не зависит от времени (пропорционально e). Для частных случаев исхода (> 10%) лучше использовать логистическую регрессию, но часто объяснить риски не можем, необходимо провести дополнительный расчет по модели Кокса и полученные риски проще объяснить.

Используя модели пропорциональных рисков Кокса, мы получаем вероятность, которую интерпретируем как относительный риск, а не как отношение шансов, что делаем при логистическом анализе.

Модель проверяется на адекватность статистических критериев ($\chi 2$) и соблюдение условия пропорциональности рисков.

Введение дополнительных переменных может быть осуществлено методами форсированного ввода, блочного, пошагового ввода, пошагового исключения и других алгоритмов.

Таким образом, условиями использования анализа пропорциональных рисков Кокса являются:

1) использование предикторов любого типа (пол, образование, стадия заболевания, тип опухоли, АД на момент начала исследования);

2) неизменность предикторов на протяжении времени наблюдения;

3) соблюдение условий пропорциональности рисков, что проверяют для любого изучаемого фактора (иногда это условие может быть частично компенсировано);

4) зависимость между функцией интенсивности и логлинейной функцией ковариат (факторов, изменяющихся во времени) является мультипликативной, т.е. для двух заданных наблюдений с различными значениями независимых переменных отношения их функций интенсивности не зависит от времени;

5) соотношения между функцией интенсивности и независимыми переменными характеризуется логлинейной зависимостью.

Пример: Анализ времени жизни при раке щитовидной железы. Известно время наблюдения (даты). Исход за время наблюдения был представлен смертями в 77 случаях, цензурировано - 451 , исключен 1. Требуется получить и оценить риски влияния на исход 4 факторов «Морфология», «Возраст», «Стадия» и «Пол».

Проведение анализа и заполнение таблиц с результатами при использовании метода Каплана-Майера происходит поэтапно при блоковом, а затем пошаговом вводе факторов. В Таблице 1 следует указать перечень всех наблюдений с различным исходом: событие случилось -1; событие цензурировано (не случилось) – 0; событие исключено (неполная строка), в %. В таблице 2 отмечается в абсолютных числах все отобранные категории, включая референтные, для каждого фактора, которые как условие сохраняется на всем промежутке наблюдения. В таблице 3 представляют логранговый коэффициент и его оценки по хи-квадрат на каждом шаге дополнения данных. В таблице 4 представляют все коэффициенты модели: B, df, p и Exp (B) с ДИ. Доверительный интервал коэффициентов не может включать 1. В таблице 5 представляют все коэффициенты корреляции между факторами и исходом.

Анализ примера: X_1- морфология с вариантами 1 и 2, для которых надо рассчитать Exp (B). Для варианта 1 фактора «морфология» Exp (B) = 4,24, а для второго варианта Exp (B)= 10,1. Вывод можно сформулировать так: при фолликулярном раке в 4,24 быстрее наступает исход или при фолликулярном раке скорость умирания в 4,24 раза выше, чем при референтном рак. Добавили еще 2 фактор и скорректировали оценки.

Для оценки независимого влияния факторов обращают внимание на степень изменения Exp (B) после их включения. Если Exp (B) после включения

второго фактора изменился менее, чем на 15%, то факторы независимы, а если увеличение Exp (B) превышает 15%, делаем вывод о зависимости факторов. Если в модель включить такие факторы, которые будут снижать различия Exp (B) в предыдущих факторах, то можем говорить, что этот новый фактор будет главным и мешать предыдущему. При включении X_2 «стадия» Exp (B) в первой (с 4,24 до 3,2) и второй группах (с 10,13 до 3,58) X_1 «мофология» сократились.

В нашем примере фактор X_1 (стадия) стал для фактора X_2 (морфология) конфаундером (мешающим проявиться). Конфаундер имеет связь с исходом и имеет связь с морфологией, но не лежит на одной патогенетической прямой «морфология-исход». Показатели «морфология» и «стадия» не являются отражением одной стороны патологического процесса (действаия). Если при включении X_3-пол Exp (B) уменьшится, то значит у мужчин (согласно кодировке: 0-женщины, референтная и 1 - мужчины) менее благоприятная морфология и мужчины умирают чаще. Если для мужчин и женщин зависимости разные, то сначала анализируют пол как фактор, и считают общий эффект, а потом разделяют на 2 части по полу, снова считают эффект . А если взаимодействия исхода и пола нет, то нет необходимости проверять для мужчин и женщин.

При анализе включения возраста референтным будем считать возраст всей группы. Выявили, что для всей группы влияние значимо, а для 1 и 2 групп не значимо, а тренд определить нельзя. Считаем что возраст резидуальный конфаундер. Резидуальный конфаундер – это такой мешающий фактор, который можно только зафиксировать, но нельзя объяснить.

Результатом анализа будут графики кумулятивной выживаемости с коррекцией на возраст, стадию, морфологию и пол.

Особенностями регрессии Кокса является возможность оценивать независимые влияния каждого из факторов, включенных в модель. При этом сохраняется опасность наличия резидуальных конфаундеров и опасность наличия взаимодействия между переменными (interaction).

Важно понимать, что различные методы статистического анализа имеют свои возможности и ограничения, что определяет их использование в медицинских и экологических исследованиях.

Заключение.

В заключение можно отметить, что в математической статистике принято разграничивать любые переменные на четыре типа шкал: номинальную (наименований), ранговую (порядковую), интервалов и отношений (абсолютную). Как самостоятельный тип можно выделить бинарные данные, которые хотя и относятся к шкале наименований, но к ним можно применять целый ряд самостоятельных методов обработки. При совместном рассмотрении данных, измеренных в разных шкалах, с ними можно выполнять различные преобразования, переводящие все данные в одну шкалу. Переход от более грубой, «качественной» шкалы к шкале более высокого – «количественного» характера («оцифровка») не

всегда корректен и достаточно сложен. Обратный переход можно выполнять всегда, но часто это приводит к значительной потере информации. Для перехода от одной шкалы к другой необходимо выйти за границы понятий (классификации, оценки измерения), принятых в исходной шкале, и, используя некое дополнительное знание, по-другому оценить, измерить, квалифицировать тот же самый объект [23].

Таким образом, при выполнении любого из представленных видов анализа необходимо выполнить ряд последовательных действий: 1) определить цель и задачи; 2) определить тип переменной; 3) соблюсти все условия для выполнения выбранного типа анализа; 4) оценить качество построенной модели; 5) описать модель; 6) провести диагностику модели. Весь перечень последовательных действий с использованием модулей программы Statistica 10.0 мы попытались представить в виде алгоритма - отдельных этапов дороги, представленных в таблице Приложения.

Литература

1. Алесинская Т.В. Основы логистики. Общие вопросы логистического управления. Учебное пособие. Таганрог: Изд-во ТРТУ, 2005.- 121 с.

2. Вараксин А. Н. Почему ощущается нехватка специалистов в области статистического анализа медицинских данных?// Международный журнал медицинской практики.-М.:«Медиа Сфера», 2007.-№1.- С.76.

3. Хомяков Д.М., Искандарян Р.А. Информационные технологии и математическое моделирование в задачах природопользования // http://fadr.msu.ru/rin/ecol/model.htm.

4. Багоцкий С.В., Базыкин А.Д., Монастырская Н.П. Математические модели в экологии. Библиографический указатель отечественных работ. - М.: ВИНИТИ, 1981. -226 с.

5. Джефферс Дж. Введение в системный анализ: применение в экологии. М.: Мир, 1981. - 256 с.

6. Лапко А.В., Крохов С.В., Ченцов С.И., Фельдман Л.А. Обучающиеся системы обработки информации и принятия решений. - Новосибирск: Наука, 1996. - 284 с.

7. Хомяков Д.М., Хомяков П.М. Основы системного анализа. М.: Изд-во мех.-мат. ф-та. МГУ, 1996. - 107 с.

8. Быков А.А., Мурзин Н.В. Проблемы анализа безопасности человека, общества и природы. - СПб.: Наука, 1997. - 247 с.

9. Киреева Н.А., Водопьянов В.В. Математическое моделирование микробиологических процессов в нефтезагрязненных почвах // Почвоведение. - 1996. - №10. -С. 1222-1226.

10. Пых Ю.Г., Малкина-Пых И.Г. ПОЛМОД (версия 1.0) - Модель миграции загрязняющих веществ в элементарной экосистеме (На примере радионуклида Sr90). -М., 1992. -63 с.

11. Фесенко С.В., Санжарова Н.И., Алексахин Р.М., Спиридонов С.И. Изменение биологической доступности 137-Cs после аварии на ЧАЭС // Почвоведение.- 1995.- № 4. - С. 508-513.

12. Фесенко С.В., Яцало Б.И., Спиридонов С.И. Применение математических моделей в радиоэкологии // Вестник РАСХН.- 1996.-№4. - С. 29-31.

13. Лиепа И.Я. Математические методы в биологических исследованиях. Факторный и компонентный анализы. - Рига, - 1980, - 104 с.

14. Робертс Ф.С. Дискретные математические модели с приложениями к социальным, биологическим и экологическим задачам / Пер. с англ. А.М. Раппопорта, С.И. Травкина. Под ред. А.И. Теймана. - М.: Наука, 1986. - 496 с.

15. Джефферс Дж. Введение в системный анализ: применение в экологии. М.: Мир, 1981. - 256 с.

16. Максимов В.Н., Милованова Г.Ф., Булгаков Н.Г., Левич А.П. Индикация состояния экосистем методами детерминационного анализа// Теоретические проблемы экологии и эволюции. Тольятти, 2000. -С.113-120.

17. О некоторых принципах построения и анализа регрессионных моделей в задачах медико-экологического мониторинга / Т. А. Маслакова, А. Н. Вараксин, В.Н. Чуканов // Экологические системы и приборы. - 2004. - №9. - С. 27-31.

18. Дмитриев А.А. Алгоритм прогноза по известному спектру частот / Вопросы агроэкологического прогнозирования // Науч.-техн. бюл. /РАСХН. Сиб. отд-е. СибНИИЗХим. - Новосибирск, 1991. Вып. 5. -С. 33-37.

19. Алесинская Т.В.Основы логистики. Общие вопросы логистического управления Учебное пособие. Таганрог: Изд-во ТРТУ, 2005.- 121 с.

20. Боровиков В. STATISTICA: Искусство анализа данных на компьютере. Для профессионалов - СПб.: Питер, - 2001. - 656 с.

21. Реброва О.Ю. Статистический анализ медицинских данных. Применение пакета прикладных программ STATISTICA.. М.: Медиа Сфера, 2002. - 312 с.

22. Юнкеров В.И., Григорьев С.Г. Математико-статистическая обработка данных медицинских исследований.- СПб.: ВМедА, 2002.- 266 с.

23. Антомонов М. Ю. Математические аспекты анализа данных в медико-экологических исследованиях.- Киев, институт гигиены и медицинской экологии им. А.Н. Марзеева АМНУ http://www.health.gov.ua/publ/conf.nsf/0/c2e9d1e8ae-70d844c2256dc6003f7ea3?OpenDocument

24. Лебедева Н.В. Современные методы оценки влияния вредных факторов окружающей среды на здоровье населения/ Н.В Лебедева, В.Д. Фурман, В.А. Кислицин, Г.М. Земляная - М.: Центр подготовки и реализации международных проектов технического содействия (ЦПРП) - < txt/rus/articlour/art2.htm>

25. Куликов Е. Исследование операций математической модели.- Днепропетровск, 2007.- 17 с.

Приложение

Таблица – Алгоритм использования модулей программы Statistica 10.0 для проведения многомерного регрессионного анализа

	Назначение	Дорога в программе
1	Для количественных данных слева и справа. Оценка статистической значимости каждого коэффициента модели и свободного члена используется для обоснования (P < 0,05) и его включения в модель	STATISTICA:—> Меню: "Анализ" —> Модуль **"Множественная регрессия"** —>" Окно: "Переменные" —>ОК—> выбрать слева зависимую переменную, справа - одну или несколько независимых переменных—>ОК—>Выбрать опцию "продолжить как есть" —> ОК —> дополнительно—> выбрать «описательные статистики» и «корреляционная матрица» —> щелчок—> скопировать в электронные формат таблицы значений (M, SD, N) и (парные коэффициенты корреляции) —> ОК—> в процедуре выбрать формат «пошаговый с исключением» —> ОК в окне Результаты анализа признаков —> опция "итоговая таблица регрессии" —> ОК —> Величины коэффициентов и характеристики модели в таблице скопировать в электронные таблицы результатов—> дополнительно —> выбрать опцию «избыточность» —> значение толерантности записать. Выбрать опцию «итоги по шагам» —> ОК —>Результаты анализа признаков с изменением R^2 и F —> ОК—> скопировать в электронные таблицы результатов. Выбрать опцию "остатки/предсказанные/наблюдаемые значения" —> ОК—> в перечне подстрочных модулей выбрать опцию "анализ остатков" —> "остатки" —>ОК—> в окне анализа остатков "График остатков"—>ОК —>скопировать нормальный вероятностный график остатков —> затем в опциях дополнительно—> указать статистика Дорбана-Уотсона —> ОК—> скопировать значение—> опция «диаграммы рассеивания»—>щелчок—> выбрать опцию «предсказанные и квадраты остатков» —> скопировать график —> опция «выбросы» —> отметить кнопку «расстояние Кука»—> просмотреть список для поиска наблюдений со значением более 1—> ОК—> включить кнопку «стандартный остаток»—>активировать вкладку

Продолжение таблицы

		«построчный график выбросов» —>отметить наблюдения с СКО более 3 —>пересчитать модель без этих наблюдений и вновь сравнить ее качество.
2	Для качественных (слева) и количественных данных (справа)	STATISTICA:—> Меню: "Анализ" —> Модуль **"Углубленные методы анализа"** —>" Окно: "общие регрессионные модели —> опция "простая регрессия" или "множественная регрессия" или "общие линейные модели"—>" в окне состояний выбрать слева зависимую переменную, справа – предикторы (категорированные или непрерывные) —>ОК—>Выбрать опцию "продолжить как есть" —> ОК —> Результаты анализа признаков в окне состояний —>ОК —> "Все эффекты"—> Величины коэффициентов и характеристики модели в таблице —>скопировать в электронные таблицы результатов..
3	Используемые преобразования: X^2, X^3, X^4, X^5, X^{-2}, ln X, log X , e^X, 10^X, 1/X	STATISTICA:—>Меню:"Анализ"—>Модуль"**Углубленные методы анализа"**—>"Окно:"множественная нелинейная регрессия—>В окне состояний выбрать переменные—>ОК—>Выбрать опцию "продолжить как есть"—>ОК—>В окне состояний регрессия с нелинейными компонентами выбрать вид преобразования (степенная, эспоненциальная, обратная и т.п.) —> ОК—>В окне состояний в нижней строке выбрать слева зависимую переменную, справа - одну или несколько независимых переменных с учетом преобразования—> ОК —>Выбрать опцию "продолжить как есть" —> ОК—>Результаты анализа признаков в окне —> опция "итоговая таблица регрессии" —> ОК —> Величины коэффициентов и характеристики модели в таблице —>скопировать в электронные таблицы результатов.
4	Логистическая регрессия для дихотомических (слева) и качественных или количественных данных (справа)	STATISTICA: —> Меню: "Анализ" —> Модуль "**Углубленные методы анализа"** —>" Окно: "Обобщенные линейные и нелинейные модели" при выборе - "дополнительно" —> в окне состояний выбрать модель: "полиноминальная" и вид функции связи "логистическая"—>ОК—> В окне состояний выбрать слева зависимую переменную (дихотомическую), справа в окнах - одну или несколько независимых переменных (категорированных или непрерывных)—> ОК

Продолжение таблицы

		—>Выбрать опцию "продолжить как есть" —> ОК —> в окне результатов на вкладке «итог» активировать опцию «все эффекты», «критерий отношения правдоподобия» и «оценивание» —> скопировать в электронную версию таблицы значения коэффициента Вальда (оценка статистической значимости модели) и коэффициента лог-правдоподобия (2 LL), как аналога R^2 с оценкой по χ^2 и p; а также параметры модели b_0 и b_1 , b_2 —> активируя опцию «остатки1»—> активировать опцию «классификация и отношение шансов»—> скопировать таблицу (главная классификационная таблица) —> получаем и копируем график—>«наблюдаемые и предсказанные значения»; «нормальный график остатков» и гистограмму «остатки Пирсона» —>активируя опцию «остатки 2»—>график «предсказанные значения и расстояние Кука» —>проводим оценивание.
5	Временные ряды	STATISTICA:—>Меню: "Анализ"—>Модуль Углубленные методы анализа—> окно «Временные ряды и прогнозирование»—>ОК—> в окне «Анализ временных рядов» активизировать опцию «Анализ распределения лагов—> в окне состояний выбрать переменные (до 20) —> ОК—> выбрать "продолжить как есть" —>ОК—>в окне состояний выбрать независимые переменные —>ОК—>активизировать кнопку «Полиноминальные лаги Алмона» и опцию «ОК (начать анализ)» —>скопировать в результаты содержимое таблицы—>выбрать опцию «Прогноз»—>выбрать и скопировать гистограмму, нормальный вероятностный график и график без тренда. Вернуться в окно «Анализ временных рядов» и активизировать опцию «Фурье (спектральный) анализ»—>активизировать опцию "Двумерный анализ Фурье»—>в окне результатов активировать «периодограмма и графики плотности»—>активировать кнопку «итог»—>скопировать графики результатов—>активировать опцию «дополнительно» —> активировать опцию «отобразить N наиболее значительных»—> скопировать из таблицы наибольшие частоты.

Продолжение таблицы

6	Анализ выживаемости	STATISTICA:—> Меню: "Анализ" —>"Углубленные методы анализа" —>" Модуль «Анализ выживаемости»—> ОК—> окно «Таблицы времени жизни» —> ОК—> Выбрать переменные в окнах (в левом - время жизни, в правом - индикатор цензурирования), в окне «код для полных» указать 1, а в окне «код для цензурированных» - 0; активизировать кнопку числа интервалов и указать их число (до) —> ОК—> в окне «Результатов времени жизни» выбрать кнопку «Таблицы времени жизни»—> скопировать в результаты содержимое таблицы; активизировать кнопку анализа в левом углу экрана и выбрать кнопку «Оценка параметров» —> скопировать в результаты содержимое таблицы; активизировать кнопку анализа в левом углу экрана и выбрать кнопку Графики функций —> скопировать в результаты график функции выживаемости. Выбрать окно: «Метод множительных оценок Каплана-Мейера» —> ОК—> Выбрать переменные в окнах (в левом - время жизни, в правом - индикатор цензурирования), в окне «код для полных» указать 1, а в окне «код для цензурированных» - 0—> ОК—> продолжить как есть—>в окне «Результатов времени жизни» активизировать функцию «дополнительно» и выбрать кнопку «Процентили функции выживаемости»—> скопировать в результаты содержимое таблицы; активизировать кнопку «графики» и выбрать «времени жизни и кумулятивной доли выживших»—> скопировать в результаты график. Выбрать окно: «Сравнение 2 группа» —> ОК—> Выбрать переменные в окнах (в левом - время жизни, в среднем- индикатор цензурирования, в правом - группировочная), в окне «код для полных» указать 1, а в окне «код для цензурированных» - 0, в окне группировочной указать коды для сравниваемых групп: 1 и 2—> ОК—> в окне «Результатов сравнения 2 групп» активизировать функцию «двувыборочные критерии —> выбрать «логранговый критерий»—>скопировать верх таблицы; выбрать «доля выживших по группам»—>скопировать таблицу; активизировать функцию

Продолжение таблицы

		«графики функций»—>выбрать «график функции выживаемости по группам» —>скопировать график. Выбрать окно: «Сравнение нескольких выборок» —> ОК—> Выбрать переменные в окнах (в левом - время жизни, в среднем- индикатор цензурирования, в правом- группировочная), в окне «код для полных» указать 1, а в окне «код для цензурированных» - 0, в окне группировочный указать коды для сравниваемых групп (1-4)—> ОК—> в окне «Результатов сравнения выживаемости в нескольких группах» активизировать функцию «дополнительно» —> выбрать «времена жизни и вклады по группам» —> скопировать таблицу «итоговые статистики»; выбрать «график функции выживания по группам» —>скопировать график; выбрать «процент времени жизни по группам» —> скопировать таблицу; активизировать функцию «описательные статистики» —> выбрать «описательные статистики» —> скопировать таблицу. Выбрать окно: «Регрессионные модели» —> ОК—> Выбрать переменные в окнах (в левом - время жизни, в среднем- независимые переменные, в правом - индикатор цензурирования), в окне «код для полных» указать 1, а в окне «код для цензурированных» - 0 —> ОК—> в окне «Результатов регрессии» активизировать функцию «дополнительно» —> выбрать «ковариации и корреляции оценок» —> скопировать таблицу; выбрать «средние и стандартные отклонения» —> скопировать таблицу; выбрать «график функций» —>скопировать график функций для средних. STATISTICA:—> Меню: "Анализ" —>"Углубленные методы анализа" —>" Модуль «Пропорциональные интенсивности Кокса»—> ОК—> Выбрать переменные в окнах (в левом - время жизни или факторы, преобразованные как ВЖ, в правом - индикатор цензурирования), во втором- ковариванты, в третьем - факторы, в окне «код для полных» указать 1, а в окне «код для цензурированных» - 0;—> ОК—> в окне «Результатов пропорциональных интенсивностей Кокса» активизировать опцию «быстро» и выбрать «качество подгонки»»—> скопировать в результаты содержимое

Продолжение таблицы

		таблицы; выбрать «оценка параметров»—> скопировать в результаты содержимое таблицы; выбрать «тесты Тип3»—> скопировать в результаты содержимое таблицы; выбрать «остатки»—> скопировать в результаты содержимое таблицы.