УДК 616:519.2
ББК К 519

Талатов Ёкубжон Талатович.

Методы статистического анализа и исследование зависимостей медицинских данных.

В статье приводятся результаты исследований различных методов статистического анализа с точки зрения их использования при анализе и обработке медицинских данных. Приводятся преимущества и недостатки каждого метода.

Маколада тиббиет маълумотларини тахлил килиш ва ишлов бериш учун турли статистик тахлил килиш усуллари устида олиб борилган изланишларнинг натижалари келтирилган. Хар бир усулни устунликлари ва камчиликлари берилган.

In the article results of researches of various methods of the statistical analysis from the point of view of their use at the analysis and processing of the medical data are resulted. The advantages and disadvantages of each method are given.

Статистический анализ медицинских данных основывается на исследование зависимостей между переменными. С этой целью применяются корреляционный анализ (для установления факта наличия или отсутствия зависимости между переменными, выраженной в виде числового значения), а также регрессионный анализ (для нахождения количественной зависимости между переменными, выраженной в виде уравнения и/или графика) [1,2].

Корреляционный анализ - взаимосвязь между двумя или более переменными (в последнем случае корреляция называется множественной или совокупной). Цель корреляционного анализа - установление наличия или отсутствия этой взаимосвязи. В случае, когда имеются две переменных,

значения которых измерены в шкале отношений, используется коэффициент линейной корреляции Пирсона **r**, который принимает значения от -1 до +1 (нулевое его значение свидетельствует об отсутствии корреляции).

Термин «линейный» свидетельствует о том, что исследуется наличие линейной связи между переменными.

Для данных, измеренных в порядковой шкале, следует использовать коэффициент ранговой корреляции Спирмена, так как он является непараметрическим и улавливает тенденцию - изменения переменных в одном направлении, который обозначается r_s и определяется сравнением *рангов* - номеров значений сравниваемых переменных в их упорядочении. Коэффициент корреляции Спирмена является менее чувствительным, чем коэффициент корреляции Пирсона.

Важно отметить, что близкое к плюс единице или к минус единице значение коэффициента корреляции говорит о силе взаимосвязи переменных прямой или обратной, но ничего не говорит о причинно-следственных отношениях между ними.

В отличие от корреляционного анализа, регрессионный анализ — не только говорит о наличии зависимости между независимой переменной и одной или несколькими зависимыми переменными, но и позволяет определить эту зависимость количественно. Независимые переменные называют регрессорами или предикторами, а зависимые переменные — критериальными. Опять же, терминология зависимых и независимых переменных отражает лишь математическую зависимость переменных, а не причинно-следственные отношения.

Существует несколько видов линейного и нелинейного регрессионного анализа, позволяющие обнаружить математическую зависимость между несколькими переменными, однако все эти методы являются параметрическими, что делает невозможным их применение для обработки качественных данных. Непараметрическим аналогом множественной регрессии является логистическая регрессия с двумя градациями зависимого

признака (бинарная логистическая регрессия) и более (мультиноминальная логистическая регрессия) [3,4].

С помощью метода бинарной логистической регрессии можно исследовать зависимость дихотомических (бинарных, имеющих только 2 категориальных значения) переменных от независимых переменных, имеющих любой вид шкалы. Как правило, в случае с дихотомическими переменными речь идёт о некотором событии, которое может произойти или не произойти; бинарная логистическая регрессия в таком случае рассчитывает вероятность наступления события в зависимости от значений независимых переменных с выводом коэффициентов регрессии для каждой такой переменной и её статистической значимости.

Вероятность наступления бинарного события рассчитывается по формуле:

$$F(z) = P(Y = 1|X) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}}$$

где $z = b_1 * X_1 + b_2 * X_2 + ... + b_n * X_n + a$,

X — значения независимых переменных, b_j — коэффициенты, расчёт которых является задачей бинарной логистической регрессии, a — константа полученного регрессионного уравнения.

Если рассчитанная вероятность имеет значение меньшее 0,5, то можно предположить, что событие не наступит; в противном случае предполагается наступление события.

Коэффициенты в полученной регрессии не следует интерпретировать как эффект от изменения X. Для правильной трактовки следует найти производную логистической функции по параметру X и вычислить предельный эффект (marginal effect) при конкретном значении переменной X (обычно вычисляется в среднем значении).

Экспоненты коэффициентов логистической регрессии с учётом 95% доверительного интервала используются как отношения шансов в качестве оценки вероятности наступления изучаемого бинарного события по

представляемой переменной в совокупности всех представленных статистически значимых переменных.

Мультиноминальная логистическая регрессия позволяет исследовать процессы, при которой зависимая переменная имеет больше двух категорий. В то время как, при бинарной логистической регрессии независимая переменная может иметь непрерывную шкалу, то мультиномиальная логистическая регрессия пригодна только для категориальных независимых переменных, причём имеет значение, относятся ли они к шкале наименований или к порядковой шкале.

Для построения мультиномиальной логистической регрессии формируется n сдублированных логитов для n+1 возможных значений независимой переменной, причём одна категория используется как эталонная, ее коэффициенты принимаются равными 0:

$$g1 = ln\frac{p1}{pn} = b_{10}+b_{11}+\ldots+b_{1(n-2)}$$

$$g2 = ln\frac{p2}{pn} = b_{20}+b_{21}+\ldots+b_{2(n-2)}$$

$$gn = 0$$

Нахождение коэффициентов b_{10}, b_{1b} b_{20} и b_{21} (называемых параметрическими оценками) является основной задачей мультиномиальной логистической регрессии. Первая цифра индекса указывает на номер логита, а вторая на порядковый номер коэффициента в данном логите, причём цифра 0 на второй позиции индекса означает константу, за которой далее следует ровно столько коэффициентов, сколько независимых переменных (факторов) взято в рассмотрение. Коэффициентам последней (эталонной) категории присваивается значение 0.

Получив значения для недублирующихся логитов, можно рассчитать значения дублирующихся логитов, используя правила вычисления логарифма.

$$ln\frac{p1}{p2} = ln\frac{p1}{pn} - ln\frac{p2}{pn}$$

Следует отметить, что прямое определение вероятности для каждой категории, значительно информативней, чем соотношение этих вероятностей между собой. Для каждой i-ой категории зависимых переменных эта вероятность может быть вычислена по следующей формуле:

$$p\ (i-te\ Kategorie) = \frac{\exp(g_i)}{\sum_{k=1}^{n} \exp(g_k)}$$

В случае наличия лишь одной независимой переменной проведение расчёта с применением столь громоздкого метода является достаточно бессмысленным — все соотношения могут быть выяснены проще, при помощи таблиц сопряженности.

В заключение можно отметить, что в зависимости от постановки задач анализа и обработки медицинских данных можно использовать ту или иную метод статистического анализа на основе проведенного исследования.

Список литературы

1. Бурдяк, А.Я. Применение метода «анализ наступления события (event history amlysis)» с помощью пакета SPSS / А.Я. Бурдяк // Spero. - 2007. - №6. - С.189-202.

2. Новиков, Д.А. Статистические методы в медико-биологическом эксперименте (типовые случаи) / Д.А. Новиков, В.В. Новочадов. - Волгоград: ВолГМУ, 2005. - 84 с.

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

4. Garcia-Perez, M.A. On the confidence interval for the binomial parameter / M.A. Garcia-Perez // Quality and quantity. - 2005. - N 39. - P. 467-481.