***Тема №10. Основы выбора метода исследования взаимосвязи двух групп медицинских данных.***

Одной из задач большинства медико-биологических исследований, является выявление взаимной связи одного или нескольких явлений.

Свет в окне может означать (с той или иной вероятностью), что хозяева находятся дома, кашель с мокротой может означать заболевание хроническим бронхитом. Если в серии повторяющихся наблюдений один из признаков (или его часть) появляется одновременно с другим чаще, чем можно объяснить случайным стечением обстоятельств, то это служит основанием говорить о взаимосвязи, сопряженности появления этих признаков.

Постановка задачи в такого рода исследованиях обычно выглядит следующим образом: определить наличие и силу статистической связи какого-либо признака от одного или нескольких других признаков. Знание взаимосвязи отдельных признаков дает возможность решать одну из *основных задач любого научного исследования*: возможность предвидеть, прогнозировать развитие ситуации при изменении тех или иных известных характеристик объекта исследования.

Термин *зависимость* при статистической обработке медико-биологических исследований должен использоваться весьма осторожно. С помощью статистических методов можно дать только формальную оценку взаимосвязи. Попытки механически перенести данные статистических расчетов в объективную реальность могут привести к ошибочным выводам.

*Например,* утверждение: «Чем громче утром кричат воробьи, тем выше встает солнце», несмотря на явную несуразность, с точки зрения формальной статистики, вполне правомерно. Таким образом, термин *«зависимость»* в статистическом анализе подразумевает только статистическую оценку взаимосвязи.

Любые явления в окружающем нас мире могут быть связаны прямой или обратной связью. Эта характеристика называется направленностью связи.

**По направленности связь может быть прямой или обратной.**

***Прямая (или положительная) связь*** характеризует зависимость, при которой увеличение или уменьшение значения одного признака ведет, соответственно, к увеличению или уменьшению – второго. *Например,* при увеличение температуры возрастает давление газа (при сохранении неизменным его объема). При уменьшении температуры – снижается и давление.

***Обратная (или отрицательная) связь*** характеризуется такой зависимостью, когда при увеличении одного признака второй уменьшается или, наоборот, при уменьшении одного, второй – увеличивается. Обратная зависимость или обратная связь является основой нормального регулирования почти всех процессов жизнедеятельности любого организма.

**По характеру связь может быть функциональной или корреляционной (статистической).**

***Функциональная зависимость*** – такой вид зависимости, когда каждому значению одного признака соответствует точное значение другого (зависимость может быть задана функцией). *Например:* взаимосвязь радиуса и длины окружности*.* Такую зависимость можно считать полной (исчерпывающей). Она полностью объясняет изменение одного признака изменением другого. Этот вид связи характерен для объектов, являющихся точкой приложения точных наук. В медико-биологических исследованиях сталкиваться с функциональной связью приходится крайне редко, поскольку объекты исследований имеют большую индивидуальную изменчивость. С другой стороны, характеристики биологических объектов зависят, как правило, от комплекса большого числа сложных взаимосвязей и не могут быть сведены к отношению двух или трех факторов.

***Корреляционная зависимость*** – существует в том случае, когда при изменении величины одного признака наблюдается тенденция соответствующего изменения значений другого признака.

*Например,* при изменении роста человека меняется и масса тела. Однако, эта зависимость не является полной, т.е. функциональной. У людей с одинаковым ростом может быть разная масса тела, поскольку на нее влияют и многие другие факторы (питание, здоровье и т.п.). При оценке статистических связей можно говорить только о *тенденции,* когда возрастание одного признака вызывает тенденцию возрастания или уменьшения другого признака.

Корреляционная связь описывается с помощью различных статистических характеристик. Выбор характеристики для определения взаимосвязи обусловлен видом исследуемых признаков, способами их группировки и предполагаемым характером связи. Подчас, для выявления реально существующих взаимосвязей достаточно правильно составить статистическую таблицу распределения или построить наглядный график этого распределения.

***Корреляционный анализ*** занимается измерением степени связи между двумя переменными *(х и у).* Вначале предполагаем, что как *х,* так и *у* — количественные величины, например, рост и вес.

Предположим, что есть пара величин *(х, у),* измеренных у каждого из пациентов в выборке. Мы можем отметить точку, соответствующую паре величин каждого пациента, на ***двухмерном графике рассеяния точек*** (рис 1,2,3). Обычно переменную х располагают на горизонтальной оси, а *у* — на вертикальной в той же диаграмме. Размещая точки для всех пациентов, получаем график рассеяния точек (**корреляционное поле**), который говорит о взаимосвязи между этими двумя переменными.

В результате могут возникнуть следующие ситуации:

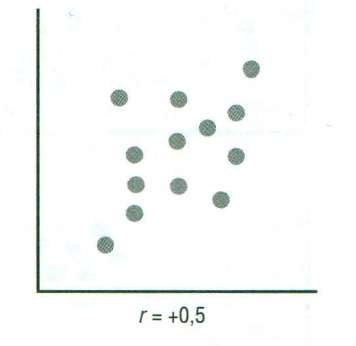


Рисунок 1. Положительная (прямая) корреляционная связь

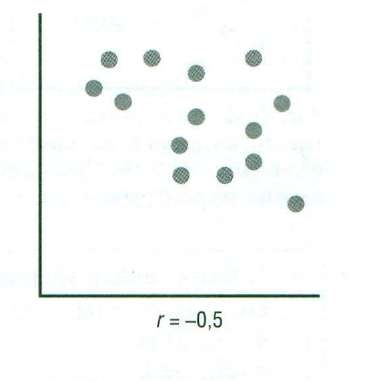
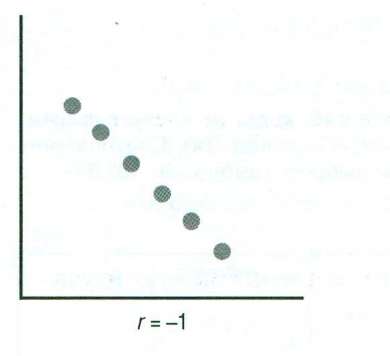


Рисунок 2. Отрицательная (обратная) корреляционная связь



Рисунок 3. Корреляционная связь отсутствует

Если на графике рассеяния точек построить прямую линию, наилучшим образом описывающую изображенные данные (расстояния от точек до прямой минимальны), то полученная прямая является **линией регрессии**. Расчет коэффициентов корреляции дает численную характеристику того, насколько близко находятся наблюдения к линии регрессии. Основными коэффициентами корреляции являются **коэффициент корреляции Пирсона и коэффициент корреляции Спирмэна**.

***Свойства коэффициентов корреляции:***

* Значения коэффициента корреляции изменяются в пределах от *-1* до *+1*.
* Знак коэффициента корреляции показывает направление связи, увеличивается (положительный *r*, прямая связь) или уменьшается (отрицательный *r*, обратная связь) одна переменная, по мере того как увеличивается другая.
* Величина коэффициента корреляции указывает, как близко расположены точки к прямой линии. В частности, если *r = +1* или *r* = *-1*, то имеется абсолютная (функциональная) корреляция по всем точкам, лежащим на линии (рис 1, рис. 2); если *r = 0*, то линейной корреляции нет (рис. 3). Чем ближе *r* к крайним точкам *(±1)*, тем больше степень линейной связи.
* Коэффициент корреляции безразмерен, т.е. не имеет единиц измерения.
* Величина коэффициента корреляции действительна только в диапазоне значений *х* и *у* в выборке. Невозможно заключить, что коэффициент будет иметь ту же величину при рассмотрении значений *х* или *у,* значительно больших, чем в выборке.
* Неважно, какой из признаков обозначить за *х*, а какой за *у; х* и *у* могут заменять друг друга, не влияя на величину r (rху~rух).
* Корреляция между *х и у* необязательно означает соотношение «причины и следствия».

Следует отметить, что в случае биологических факторов тот или иной характер связи сохраняется, как правило, только в определенном интервале изменений признаков. За пределами этого интервала связь может ослабнуть, стать прямо противоположной по направлению либо совсем исчезнуть.

*Например,* при увеличении возраста ребенка сила скелетной мускулатуры увеличивается. В зрелом возрасте такой связи уже нет. А в старших возрастных группах тенденция становится обратной.

***Сила корреляционной связи*** между признаками оценивается по величине коэффициента корреляции согласно *Таблице 1*:

*Таблица 1*

**Распределение значений коэффициента линейной корреляции**

|  |  |  |
| --- | --- | --- |
| Характеристики связи | Прямая | Обратная |
| Связи нет | 0 | 0 |
| Слабая | от 0 до 0,3 | от 0 до -0,3 |
| Средняя | от 0,3 до 0,7 | от - 0,3 до -0,7 |
| Сильная | от 0,7 до 1 | от - 0,7 до - 1 |
| Полная (функциональная) | + 1 | -1 |

***Случаи, в которых не следует рассчитывать коэффициент линейной корреляции:***

* получено нелинейное соотношение между признаками, например, квадратичное соотношение (рис. 4,а);
* данные включают более одного наблюдения по каждому пациенту;
* присутствуют аномальные значения (рис. 4,б);
* данные содержат подгруппы пациентов, для которых средние уровни наблюдений, по крайней мере, по одной из переменных, отличаются (рис. 4,в).

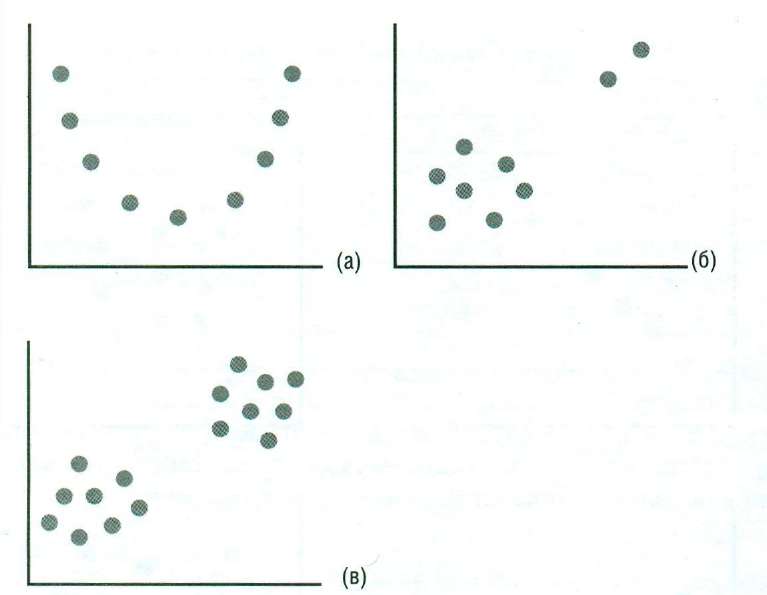


Рисунок 4. Диаграммы, показывающие, когда не следует рассчитывать коэффициент корреляции, (а) - соотношение нелинейно, (б) - при наличии выброса (выбросов), (в) - данные состоят из подгрупп.

**Коэффициент корреляции Пирсона**

***Коэффициент корреляции Пирсона* ()** определяет силу и направление связи только для количественных данных (*x, y* – значения исследуемых признаков, *n* –количество пар данных):

***Условия для расчета коэффициента корреляции Пирсона:***

* исследуемые признаки являются количественными;
* выборка состоит из независимых пар величин *х* и *у;*
* по крайней мере, одна из этих двух переменных нормально распределена.

***Достоверность коэффициента корреляции устанавливается по величине средней ошибки.*** Поскольку коэффициент корреляции в клинических исследованиях рассчитывается обычно для ограниченного числа наблюдений, нередко возникает вопрос о надежности полученного коэффициента. С этой целью определяют среднюю ошибку коэффициента корреляции. При достаточно большом числе наблюдений (больше 100) средняя ошибка коэффициента корреляции () вычисляется по формуле:

*n* – число наблюдений.

В том случае, если число наблюдений меньше 100 точнее определять среднюю ошибку коэффициента корреляции, по формуле:

С достаточной для медицинских исследований надежностью о наличии той или иной степени связи можно утверждать только тогда, когда величина коэффициента корреляции превышает или равняется величине трех своих ошибок (*r ≥3mr*). Обычно это отношение коэффициента корреляции (*r*) к его средней ошибке (*mr*) обозначают буквой *tr*:

Если ***tr≥3,*** то коэффициент корреляции является статистически значимым.

***Пример расчета коэффициента корреляции Пирсона***

Необходимо определить, существует ли связь между количеством часов, посвященных студентом подготовке к тестовому экзамену по статистике и итоговым количеством правильных ответов (и соответственно итоговой оценкой). В тестирование включает в себя 100 вопросов из банка тестовых заданий. В таблице приведены данные о шести случайно выбранных студентах.

*Таблица 2*

**Данные экзамена по статистике**

|  |  |
| --- | --- |
| Количество часов подготовки | Балл на экзамене (правильные ответы из 100 вопросов) |
| 3 | 86 |
| 5 | 95 |
| 4 | 92 |
| 4 | 83 |
| 2 | 78 |
| 3 | 82 |

*Решение:*

Очевидно, что количество часов напрямую отражается на финальной оценке. Переменная «Часы подготовки» (*х*) является независимой переменной, т.к. она приводит к наблюдаемой вариации переменной «Балл на экзамене» (*у*). Причинная связь между зависимыми и независимыми переменными существует только в одном направлении: Независимая переменная (х)→ Зависимая переменная (у). В обратном направлении эта связь не работает.

Коэффициент корреляции Пирсона (r) вычисляется при помощи следующего уравнения

Таблица, приведенная ниже, поможет разбить это уравнение на несколько несложных вычислений.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Часы изучения  *х* | Балл на экзамене  *у* | Расчеты | | |
| *ху* | *х2* | *у2* |
| 3 | 86 | 258 | 9 | 7396 |
| 5 | 95 | 368 | 25 | 8464 |
| 4 | 92 | 475 | 16 | 9025 |
| 4 | 83 | 332 | 16 | 6889 |
| 2 | 78 | 156 | 4 | 6084 |
| 3 | 82 | 246 | 9 | 6724 |
|  |  |  | =79 |  |

Используя эти значения и n=6 (общее количество студентов), получаем:

Теперь рассчитаем среднюю ошибку коэффициента корреляции

Установим, надежной, ли является установленная нами связь

Т.к. *tr≥3*, то коэффициент корреляции является статистически значимым.

*Таким образом,* между числом часов, посвященных изучению предмета, и экзаменационной оценкой существует статистически значимая сильная положительная (прямая) корреляция. Отсюда следует, что экзаменационные результаты можно предугадать на основе определенного количества часов, посвященных изучению предмета.

**Коэффициент корреляции Спирмэна**

***Ранговый коэффициент корреляции Спирмэна (rs) –*** непараметрический аналог корреляционного коэффициента Пирсона.

Применение этого коэффициента корреляции может быть рекомендовано в случаях:

* когда необходимо **быстро ориентировочно** определить связь между какими-то признаками;
* если необходимо оценить связь между **качественными (ранговыми) и количественными признаками** или **только между качественными признаками**;
* когда **распределение** значений учетных признаков (в том числе и количественных) **не соответствует нормальному** распределению или **распределение неизвестно**.

***Вычисление:***

1. Располагают величины *х* в возрастающем порядке, начиная с наименьшей величины, и придают им последовательные ранги (номера 1, 2, 3, .., n). Равные варианты получают среднее значение из суммы их порядковых номеров.

2. Подобным образом ранжируют *у*.

3. Рассчитывается rs — коэффициент корреляции между рангами *х* и *у* по формуле:

где – разности между рангами соответствующих пар *y* и *x;*

n – число сопоставляемых пар.

**Пример расчета коэффициента корреляции Спирмэна.**

Необходимо определить по *Таблице 2*, существует ли связь между количеством часов, посвященных студентом подготовке к тестовому экзамену по статистике, и итоговым количеством правильных ответов (и, соответственно, итоговой оценкой). Тестирование включает в себя 100 вопросов из банка тестовых заданий.

*Решение:*

Составляем вариационный ряд *x* и ранжируем:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***x*** | 2 | 3 | 3 | 4 | 4 | 5 |
| ***Rx*** | 1 | 2.5 | 2.5 | 4.5 | 4.5 | 6 |

Составляем вариационный ряд *y* и ранжируем:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***y*** | 78 | 82 | 83 | 86 | 92 | 95 |
| ***Ry*** | 1 | 2 | 3 | 4 | 5 | 6 |

Для удобства расчета заполняем следующую таблицу:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| ***x*** | 3 | 5 | 4 | 4 | 2 | 3 |
| ***y*** | 86 | 95 | 92 | 83 | 78 | 82 |
| ***Rx*** | 2.5 | 6 | 4.5 | 4.5 | 1 | 2.5 |
| ***Ry*** | 4 | 6 | 5 | 3 | 1 | 2 |
| ***Ry- Rx*** | 1.5 | 0 | 0.5 | -1.5 | 0 | -0.5 |
| **(*Ry- Rx*)2** | 2.25 | 0 | 0.25 | 2.25 | 0 | 0.25 |

Таким образом, получено, что исследуемая корреляционная связь является прямой и сильной.

В ходе корреляционного анализа или анализа корреляционной связи решается целая группа взаимосвязанных ***задач:***

1. Установление направления (прямая или обратная) и формы (линейная или нелинейная) корреляционной связи.
2. Оценка тесноты (силы, плотности) корреляционной связи.
3. Оценка репрезентативности статистических оценок взаимосвязей, полученных по выборочным данным (величина ошибки, доверительный интервал, уровень значимости).
4. Установление величины детерминации (доли взаимовлияния) коррелируемых факторов.

Таким образом, **статистические методы** изучения связи между переменными зависят от:

* характера переменных *(качественные, количественные)*
* характера распределения количественных переменных *(нормальное, ненормальное, неизвестное*)
* числа наблюдений *(большое, малое)*
* взаимоотношения между наблюдениями *(зависимые, независимые).*

**Статистические методы** изучения связи между переменными могут быть:

* **однофакторными,** *т.е. принимающими во внимание только взаимоотношения между двумя анализируемыми переменными*
* **многофакторными,** *т.е. учитывающими влияние на изучаемую связь между двумя переменными со стороны некоторых других переменных.*

***Понятие о регрессионном анализе***

**Регрессия** определяет **математическую зависимость** между зависимой переменной (отклик) и одной или более независимыми переменными (предикторами).

**Регрессионный анализ** с помощью коэффициента регрессии позволяет количественно прогнозировать изменения одной переменной при изменении другой.

Для описания связи могут использоваться различные математические функции, основными из которых являются:

* линейная
* экспоненциальная
* логистическая

***Простая линейная регрессия*** или множественная регрессия могут применяться для непрерывных признаков, например, давление, вес.

***Логистическая регрессия*** применима в тех случаях, когда зависимые признаки являются бинарными (например, умер/жив, выздоровел/не выздоровел).

**Линейная регрессия**

Математическое уравнение, которое оценивает линию простой линейной регрессии:

***Y = a + bx***

*х* – называется предиктором – независимой или объясняющей переменной.

Для данной величины *х, Y* — значение переменной *у* (называемой зависимой, выходной переменной, или переменной отклика), которое расположено на линии оценки. Это есть значение, которое мы ожидаем для *у* (в среднем), если мы знаем величину *х,* и называется она «предсказанное значение у» (рис. 5).

а – свободный член (пересечение) линии оценки; это значение *Y,* когда х=0.

b – угловой коэффициент или градиент оценённой линии; он представляет собой величину, на которую *Y* увеличивается в среднем, если мы увеличиваем х на одну единицу (рис. 5). Коэффициент *b* называют коэффициентом регрессии.

*Например*: при увеличении температуры тела человека на 1оС, частота пульса увеличивается в среднем на 10 ударов в минуту.



Рисунок 5. Линия линейной регрессии, показывающая коэффициент *а* и угловой коэффициент *b* (величину возрастания *Y* при увеличении *х* на одну единицу)

Математически решение уравнения линейной регрессии сводится к вычислению параметров *а* и *b* таким образом, чтобы точки исходных данных корреляционного поля ***как можно ближе лежали к прямой регрессии***.

Статистическое использование слова «регрессия» исходит из явления, известного как регрессия к среднему, приписываемого Френсису Гальтону (1889). Он показал, что, хотя высокие отцы имеют тенденцию иметь высоких сыновей, средний рост сыновей меньше, чем у их высоких отцов. Средний рост сыновей «регрессировал» или «двигался вспять» к среднему росту всех отцов в популяции. Таким образом, в среднем высокие отцы имеют более низких (но всё-таки высоких) сыновей, а низкие отцы имеют сыновей более высоких (но всё-таки довольно низких).

Мы наблюдаем регрессию к среднему при скрининге и клинических исследованиях, когда подгруппа пациентов может быть выбрана для лечения потому, что их уровни определённой переменной, скажем, холестерина, крайне высоки (или низки). Если это измерение через некоторое время повторяется, средняя величина второго считывания для подгруппы обычно меньше, чем при первом считывании, имея тенденцию (т.е. регрессируя) к среднему, подобранному по возрасту и полу в популяции, независимо от лечения, которое они могут получить. Пациенты, набранные в клиническое исследование на основе высокого уровня холестерина при их первом осмотре, таким образом, вероятно, покажут в среднем падение уровня холестерина при втором осмотре, даже если в этот период они не лечились.

Часто метод регрессионного анализа применяется для разработки нормативных шкал и стандартов физического развития.

Насколько хорошо линия регрессии согласуется с данными, можно судить, рассчитав коэффициент R (обычно выраженный в процентах и называемый коэффициентом детерминации), который равняется квадрату коэффициента корреляции (r2). Он представляет собой долю или процент дисперсии *у,* который можно объяснить связью с *х,* т.е. долю вариации признака-результата, сложившуюся под влиянием независимого признака. Может принимать значения в диапазоне от 0 до 1, или соответственно от 0 до 100%. Разность (100% - R) представляет собой процент дисперсии *у,* который нельзя объяснить этим взаимодействием.

**Пример**

Соотношение между ростом (измеренным в см) и систолическим артериальным давлением (САД, измеренным в мм рт. ст.) у детей. Мы провели анализ парной линейной регрессии зависимости САД от роста (рис. 6). Имеется существенное линейное соотношение между ростом и САД.



Рисунок 6. Двумерный график, показывающий соотношение между систолическим артериальным давлением и ростом. Изображена оценённая линия регрессии, систолическое артериальное давление.

Уравнение линии оценённой регрессии имеет следующий вид:

САД = 46,28 + 0,48 х рост.

В этом примере свободный член не представляет интереса (рост, равный нулю, явно вне диапазона величин, наблюдаемых в исследовании). Однако мы можем интерпретировать угловой коэффициент; предсказано, что у этих детей САД увеличивается в среднем на 0,48 мм рт.ст. при увеличении роста на один сантиметр

Мы можем применить уравнение регрессии для предсказания САД, которое мы ожидаем у ребёнка при данном росте. Например, ребёнок ростом 115 см имеет предсказанное САД, равное 46,28 + (0,48 х 115)=101,48 мм рт. ст., ребёнок ростом 130 имеет предсказанное САД, 46,28 + (0,48 х 130) = 108,68 мм рт. ст.

При расчете коэффициента корреляции, установлено, что он равен 0,55, что указывает на прямую корреляционную связь средней силы. В этом случае коэффициент детерминации *r2 = 0,552 = 0,3*. Таким образом, можно сказать, что доля влияния роста на уровень артериального давления у детей не превышает 30%, соответственно на долю других факторов приходится 70% влияния.

Линейная (простая) регрессия ограничивается рассмотрением связи между зависимой переменной и только одной независимой переменной. Если в связи присутствует более одной независимой переменной, тогда нам необходимо обратиться к множественной регрессии. Уравнение для такой регрессии выглядит так:

*y = a + bx1+b2x2 +.... + bnхn*

Можно интересоваться результатом влияния нескольких независимых переменных х1*, х2, .., хn* на переменную отклика *у.* Если мы полагаем, что эти х могут быть взаимозависимы, то не должны смотреть по отдельности на эффект изменения значения одного х на *у,* но должны одновременно принимать во внимание величины всех других х.

**Пример**

Поскольку между ростом и массой тела ребёнка существует сильная зависимость, можно поинтересоваться, изменяется ли также соотношение между ростом и систолическим артериальным давлением, если принять во внимание также и массу тела ребёнка и его пол. Множественная линейная регрессия позволяет изучить совместный эффект этих нескольких независимых переменных на *у.*

*Уравнение множественной регрессии в этом случае может иметь такой вид:*

*САД=79,44 –(0,03* х *рост)+ (1,18* х *вес) + (4,23* х *пол)\**

*\* - (для признака пол используют значения 0 – мальчик, 1 - девочка)*

*Согласно этому уравнению, девочка, рост которой 115 см и масса тела 37 кг, будет иметь прогнозируемое САД:*

*САД = 79,44 – (0,03* х *115) + (1,18* х *37) + (4,23* х *1) = 123,88 мм.рт.ст.*

Логистическая регрессия очень похожа на линейную; её применяют, когда есть интересующий нас бинарный исход (т.е. наличие/отсутствие симптома или субъекта, который имеет/не имеет заболевания) и ряд предикторов. Из уравнения логистической регрессии можно определить, какие предикторы влияют на исход, и, используя значения предикторов пациента, оценить вероятность того, что он/она будет иметь определённый исход. Например: возникнут или нет осложнения, будет лечение эффективным или не будет.

Начинают создания бинарной переменной, чтобы представить эти два исхода (например, «имеет болезнь»=1, «не имеет болезни»=0). Однако мы не можем применить эти два значения как зависимую переменную в анализе линейной регрессии, поскольку предположение нормальности нарушено, и мы не можем интерпретировать предсказанные величины, которые не равны нулю или единице. Фактически, вместо этого мы берём вероятность того, что субъект классифицируется в ближайшую категорию (т.е. «имеет болезнь») зависимой переменной, и чтобы преодолеть математические трудности, применяют логистическое, преобразование, в уравнении регрессии — натуральный логарифм отношения вероятности «болезни» (p) к вероятности «нет болезни» (1-p).

Интегративный процесс, называемый методом максимального правдоподобия, а не обычная регрессия (так как мы не можем применить процедуру линейной регрессии) создаёт из данных выборки оценку уравнения логистической регрессии

*logit (p) = a + bx1+b2x2 +.... + bnхn*

* **.***logit (р)*— оценка значения истинной вероятности того, что пациент с индивидуальным набором значений для *х1 ... хn* имеет заболевание;
* а — оценка константы (свободный член, пересечение);
* *b1, b2,*... ,*bn —* оценки коэффициентов логистической регрессии.