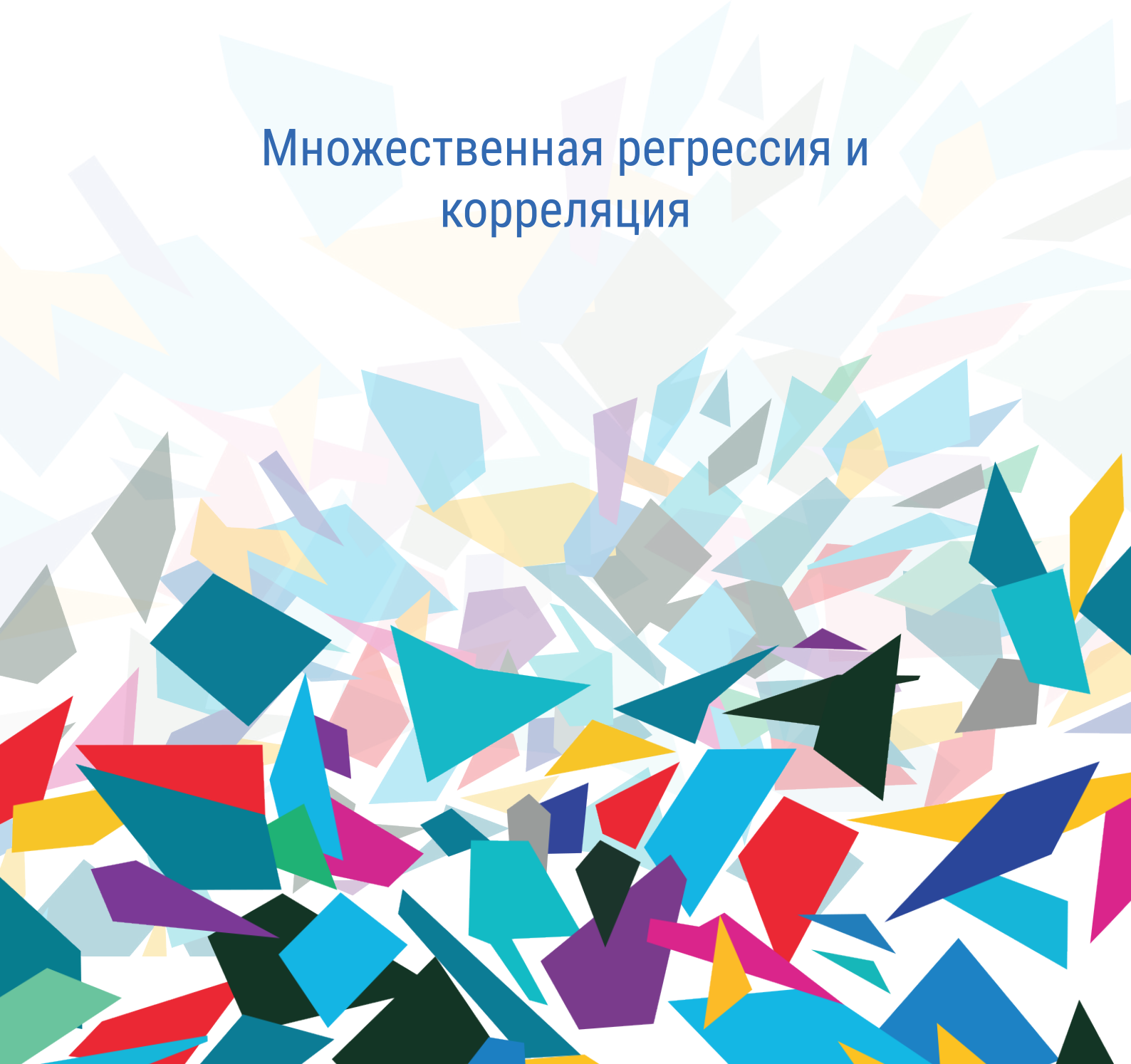
Лекция 19

ҚАЗАҚСТАННЫҢ АШЫҚ УНИВЕРСИТЕТІ

ОСНОВЫ СТАТИСТИКИ

Множественная регрессия и корреляция



Изучив данный материал у вас будет возможность строить модели множественной регрессии и давать оценку параметров модели, исключать мультиколлинеарные факторы и вводить фиктивные переменные.

Ранее описывалась модель линейной регрессии для двух переменных. В действительности социолог довольно редко сталкивается со столь простыми моделями данных. Влияние одного фактора обычно может объяснить лишь часть разброса наблюдаемых значений независимой переменной. Метод частной корреляции позволяет нам проконтролировать эффекты воздействия любых других контрольных переменных, которые мы в состоянии измерить.

(Стоит снова подчеркнуть здесь, что статистические методы изучения причинных взаимосвязей, в отличие от экспериментальных, позволяют нам контролировать лишь те источники вариации, которые мы способны концептуализировать и измерить). Однако еще более интересной задачей является контроль одновременного воздействия нескольких независимых на одну зависимую переменную, а также сравнение эффекта воздействия разных независимых переменных и предсказание “отклика” независимой переменной.

Именно эти задачи решают методы анализа, о которых пойдет речь в данной лекции. Изложение пока будет неполным, для общего понимания проблемы Уравнение множественной регрессии – это определенная модель порождения данных. Важные допущения, принимаемые в этой модели, касаются уже известного нам требования линейности, а также аддитивности суммарного эффекта независимых переменных. Последнее означает, что воздействия разных независимых переменных просто суммируются, а не, скажем, перемножаются (мультипликативный эффект, в отличие от аддитивного, имеет место тогда, когда величина воздействия одной независимой переменной на зависимую, в свою очередь, находится под влиянием другой независимой переменной, т. е. независимые переменные взаимодействуют друг с другом).

Множественная регрессия во многом аналогична простой (бивариантной) регрессии. Отличие состоит в том, что регрессия осуществляется по двум и более независимым переменным одновременно, причем каждая из них входит в регрессионное уравнение с коэффициентом, позволяющим предсказать значения зависимой переменной с минимальным количеством ошибок (критерием здесь снова является метод наименьших квадратов). Частные коэффициенты в уравнении множественной регрессии показывают, какой будет величина воздействия соответствующей независимой переменной на зависимую при контроле влияния других независимых переменных. Если воспользоваться простейшей системой обозначений, то уравнение множественной регрессии для трех независимых переменных можно записать как: где Y- это предсказываемое значение зависимой переменной, X1 ... X3 — независимые переменные, a b1, ... b3 — частные коэффициенты регрессии для каждой из зависимых переменных. Коэффициенты b могут быть интерпретированы как показатели влияния каждой из независимых переменных на зависимую при контроле всех других независимых переменных в уравнении. В отличие от коэффициентов частной корреляции коэффициенты регрессии обладают размерностью. Они показывают, на сколько единиц изменится зависимая переменная при увеличении независимой на одну единицу (при контроле всех остальных переменных модели). Пусть, например, мы построили уравнение множественной регрессии, описывающее зависимость дохода от интеллекта (X1) и стажа работы (X2). Если величина b1 оказалась равной 100, это означает, что каждый дополнительный балл по шкале интеллекта увеличивает доход на 100 денежных единиц. Значение b2 = 950 говорит нам, что год стажа прибавляет 950 денежных единиц. Однако “сырые” оценки интеллекта и стажа измерены в разных единицах. Для определения сравнительной значимости независимых переменных, входящих в уравнение множественной регрессии, мы должны подвергнуть все переменные стандартизации (т. е. перевести их в Z-оценки, см. выше). Стандартизованные коэффициенты множественной регрессии, которые удобнее всего обозначать как ß (либо греч. “бета”), меняются в пределах от - 1,0 до +1,0. Они сохраняют свою величину при изменении масштаба шкалы: переход от измерения возраста в годах к измерению в днях не изменит соответствующий ß.

Стандартизованные коэффициенты позволяют оценить “вклад” каждой из переменных-предикторов в предсказание значений независимой переменной. Если в примере с влиянием интеллекта и стажа работы на доход окажется, что ß1 = 0,25, а ß2 = 0,30, то можно заключить, что сравнительная значимость “веса” интеллекта и стажа в предсказании дохода различаются незначительно. Если же для одной переменной ß1= 0,80, тогда как ß1= 0,40, мы можем сказать, что эффект воздействия второй переменной в два раза меньше эффекта первой.

Чтобы определить ожидаемые значения зависимой переменной для отдельных индивидов, достаточно подставить в уравнение множественной регрессии соответствующие значения переменных-предикторов

и вычисленных коэффициентов b. Пусть, например, мы хотим рассчитать прогнозное значение величины дохода для человека, чей коэффициент интеллекта равен 110, а стаж работы — 20 годам. Если b1, как в вышеприведенном примере, составляет 100, b2 = 950, а слагаемое a = 50000, то мы получим:

ожидаемый доход = 50000 +100 x 110 + 950 x 20 = 80000 д.е.

Множественную регрессию можно использовать и для предсказания средних групповых значений, например, среднего дохода мужчин-врачей. Единственное различие в данном случае заключается в использовании средних значений независимых переменных для подстановки в уравнение множественной регрессии. В качестве независимой переменной множественной регрессии могут использоваться и фиктивные переменные, которым приписывают значения 0 и 1 (например, пол). Для того чтобы включить в уравнение номинальную переменную с более чем двумя категориями, нужно создать соответствующее число новых, “фиктивных” переменных, каждая из которых будет кодироваться как 0 или 1 в зависимости от наличия или отсутствия категории-признака. Скажем, состоящую из трех категорий переменную “цвет глаз” можно представить с помощью трех переменных: X1 — “голубые глаза”, X2 — “карие глаза”, X3 — “зеленые глаза”. (Человек с голубыми глазами получит 1 по X1 и 0 по двум другим переменным.)

Метод множественной регрессии очень популярен среди социологов. Вот, например, как выглядели результаты его применения в исследовании Л. Бэрона и М. Строса, изучавших факторы, влияющие на статистику изнасилований. Использованная в планировании этого исследования матрица данных включала в себя в качестве объектов (“случаев”) различные штаты США. Признаками, по которым описывались штаты, служили около десятка независимых и собственно контрольных переменных, предположительно воздействующих на зависимую переменную, – количество зарегистрированных полицией изнасилований на 100000 населения в год для данного штата (по данным ежегодных статистических отчетов ФБР). Предполагалось, что существующие различия между штатами в уровне изнасилований можно будет объяснить различиями в уровнях независимых переменных. Нужно отметить, что разброс “случаев” по зависимой переменной был весьма велик – от 71,9 на Аляске до 8,2 в Северной Дакоте (1979). Из десятка переменных, включенных в уравнение множественной регрессии, девять оказались статистически значимы.

При интерпретации результатов множественной регрессии стандартизованные коэффициенты, как уже говорилось, используют в качестве показателей значимости, “вклада” соответствующих переменных. Эта трактовка верна лишь в определенных пределах. При нарушении некоторых условий сравнение абсолютных величин стандартизованных коэффициентов может вести к неверным выводам. Дело в том, что коэффициенты регрессии подвержены влиянию случайных ошибок измерения. Использование ненадежных индикаторов “сдвигает” регрессионные коэффициенты к нулю. Иными, словами, более надежные индикаторы дают более высокие оценки коэффициентов.

Пусть, например, для предсказания риска сердечно-сосудистых заболеваний использовались две независимые переменные индивидуального уровня – “ориентация на достижения” и “склонность подавлять агрессию”, – причем шкала для измерения первой обладала более высоким коэффициентом надежности. Если стандартизованный коэффициент регрессии для достижения мотивации окажется выше, чем для подавления агрессии, это может рассматриваться как следствие таких содержательных различий между переменными, которые важны с точки зрения теории психосоциальных факторов заболеваемости. Но нельзя исключить и альтернативное объяснение, связывающее более высокий регрессионный коэффициент первой переменной с побочными эффектами методов измерения: влияние ориентации на достижения не превосходит влияния, оказываемого на риск инфаркта склонностью подавлять агрессию, а наблюдаемые различия регрессионных коэффициентов связаны лишь с ненадежностью использованных индикаторов склонности к подавлению.

Другая проблема, требующая некоторой осторожности в интерпретации коэффициентов регрессии, возникает вследствие того, что модель множественной регрессии не обязывает нас ни к каким строгим предположениям о причинных связях между независимыми переменными. Регрессионное уравнение, образно говоря, не делает никаких различий между собственно независимыми, т. е. теоретически специфицированными, переменными и дополнительными – контрольными, опосредующими и т.п.— факторами, вводимыми в модель с целью уточнения. В тех случаях, когда теоретическая гипотеза, проверяемая в ходе исследования, допускает: 1) существование взаимосвязей между независимыми переменными, 2) наличие прямых и косвенных (опосредованных) влияний, а также 3) использование нескольких индикаторов для каждого латентного фактора, могут понадобиться более совершенные статистические методы.

Итак, перейдем к детализации, рассмотрим модель множественной регрессии

Формула 19.1

$$y = f(x_1, x_2, ..., x_m)$$

где y – зависимая переменная (результативный признак), x_i – независимые, или объясняющие, переменные (признаки-факторы).

В настоящее время множественная регрессия – один из наиболее распространенных статистических методов. Основная цель множественной регрессии – построить модель с большим числом факторов, определив при этом влияние каждого из них в отдельности, а также совокупное их воздействие на моделируемый показатель.

Спецификация модели. Отбор факторов при построении уравнения множественной регрессии.

Построение уравнения множественной регрессии начинается с решения вопроса о спецификации модели. Он включает в себя два круга вопросов: отбор факторов и выбор вида уравнения регрессии.

Включение в уравнение множественной регрессии того или иного набора факторов связано прежде всего с представлением исследователя о природе взаимосвязи моделируемого показателя с другими экономическими явлениями. Факторы, включаемые во множественную регрессию, должны отвечать следующим требованиям.

Они должны быть количественно измеримы. Если необходимо включить в модель качественный фактор, не имеющий количественного измерения, то ему нужно придать количественную определенность.

Факторы не должны быть интеркоррелированы и тем более находиться в точной функциональной связи.

Включение в модель факторов с высокой интеркорреляцией, может привести к нежелательным последствиям – система нормальных уравнений может оказаться плохо обусловленной и повлечь за собой неустойчивость и ненадежность оценок коэффициентов регрессии.

Если между факторами существует высокая корреляция, то нельзя определить их изолированное влияние на результативный показатель и параметры уравнения регрессии оказываются неинтерпретируемыми.

Включаемые во множественную регрессию факторы должны объяснить вариацию независимой переменной. Если строится модель с набором факторов, то для нее рассчитывается показатель детерминации R^2, который фиксирует долю объясненной вариации результативного признака за счет рассматриваемых в регрессии m факторов. Влияние других, не учтенных в модели факторов, оценивается как $1-R^2$ с соответствующей остаточной дисперсией S^2.

При дополнительном включении в регрессию m+1 фактора коэффициент детерминации должен возрастать, а остаточная дисперсия уменьшаться:

$$R^2_{m+1} \geq R^2_m \qquad S^2_{m+1} \leq S^2_m$$

Если же этого не происходит и данные показатели практически не отличаются друг от друга, то включаемый в анализ фактор x_{m+1} не улучшает модель и практически является лишним фактором.

Насыщение модели лишними факторами не только не снижает величину остаточной дисперсии и не увеличивает показатель детерминации, но и приводит к статистической незначимости параметров регрессии по критерию Стьюдента.

Таким образом, хотя теоретически регрессионная модель позволяет учесть любое число факторов, практически в этом нет необходимости. Отбор факторов производится на основе качественного теоретико-экономического анализа. Однако теоретический анализ часто не позволяет однозначно ответить на вопрос о количественной взаимосвязи рассматриваемых признаков и целесообразности включения фактора в модель. Поэтому отбор факторов обычно осуществляется в две стадии: на первой подбираются факторы исходя из сущности проблемы; на второй – на основе матрицы показателей корреляции определяют статистики для параметров регрессии.

Коэффициенты интеркорреляции (т.е. корреляции между объясняющими переменными) позволяют исключать из модели дублирующие факторы. Считается, что две переменные явно коллинеарны, т.е. находятся между собой в линейной зависимости, если $r_{x_i x_j} \geq 0,7$. Если факторы явно коллинеарны, то они дублируют друг друга и один из них рекомендуется исключить из регрессии. Предпочтение при этом отдается не фактору, более тесно связанному с результатом, а тому фактору, который при достаточно тесной связи с результатом имеет наименьшую тесноту связи с другими факторами. В этом требовании проявляется специфика множественной регрессии как метода исследования комплексного воздействия факторов в условиях их независимости друг от друга.

Пусть, например, при изучении зависимости $y = f(x_1, x_2, x_3)$ матрица парных коэффициентов корреляции оказалась следующей:

Таблица 19.1 – Матрица парных коэффициентов корреляции

	y	x_1	x_2	x_3
y	1	0,8	0,7	0,6
x_1	0,8	1	0,8	0,5
x_2	0,7	0,8	1	0,2
x_3	0,6	0,5	0,2	1

Очевидно, что факторы x_1 и x_2 дублируют друг друга. В анализ целесообразно включить фактор x_2, а не x_1, хотя корреляция x_2 с результатом y слабее, чем корреляция фактора x_1 с y ($r_{yx2}=0,7<r_{yx1}=0,8$), но зато значительно слабее межфакторная корреляция $r_{x2x3}=0,2<r_{x1x3}=0,5$. Поэтому в данном случае в уравнение множественной регрессии включаются факторы x_2, x_3.

По величине парных коэффициентов корреляции обнаруживается лишь явная коллинеарность факторов. Наибольшие трудности в использовании аппарата множественной регрессии возникают при наличии мультиколлинеарности факторов, когда более чем два фактора связаны между собой линейной зависимостью, т.е. имеет место совокупное воздействие факторов друг на друга. Наличие мультиколлинеарности факторов может означать, что некоторые факторы будут всегда действовать в унисон. В результате вариация в исходных данных перестает быть полностью независимой и нельзя оценить воздействие каждого фактора в отдельности.

Включение в модель мультиколлинеарных факторов нежелательно в силу следующих последствий:

Затрудняется интерпретация параметров множественной регрессии как характеристик действия факторов в «чистом» виде, ибо факторы коррелированы; параметры линейной регрессии теряют экономический смысл.

Оценки параметров ненадежны, обнаруживают большие стандартные ошибки и меняются с изменением объема наблюдений (не только по величине, но и по знаку), что делает модель непригодной для анализа и прогнозирования.

Теперь можно осуществить отбор факторов по принципу устранения мультиколлинеарных факторов.

Существует ряд подходов преодоления сильной межфакторной корреляции. Самый простой путь устранения мультиколлинеарности состоит в исключении из модели одного или нескольких факторов. Другой подход связан с преобразованием факторов, при котором уменьшается корреляция между ними.

Отбор факторов, включаемых в регрессию, является одним из важнейших этапов практического использования методов регрессии. Подходы к отбору факторов на основе показателей корреляции могут быть разные. Они приводят построение уравнения множественной регрессии соответственно к разным методикам. В зависимости от того, какая методика построения уравнения регрессии принята, меняется алгоритм ее решения на ЭВМ.

Наиболее широкое применение получили следующие методы построения уравнения множественной регрессии:

Метод исключения – отсев факторов из полного его набора.

Метод включения – дополнительное введение фактора.

Шаговый регрессионный анализ – исключение ранее введенного фактора.

При отборе факторов также рекомендуется пользоваться следующим правилом: число включаемых факторов обычно в 6–7 раз меньше объема совокупности, по которой строится регрессия. Если это соотношение нарушено, то число степеней свободы остаточной дисперсии очень мало. Это приводит к тому, что параметры уравнения регрессии оказываются статистически незначимыми, а F-критерий меньше табличного значения.

Фиктивные переменные.

При исследовании влияния качественных признаков в модель можно вводить фиктивные переменные, принимающие, как правило, два значения: единица, если данный признак присутствует в наблюдении, и ноль при его отсутствии.

Если включаемый в рассмотрение качественный признак имеет не два, а несколько значений, то используют несколько фиктивных переменных, число которых должно быть на единицу меньше числа значений признака.

При назначении фиктивных переменных исследуемая совокупность по числу значений качественного признака разбивается на группы. Одну из групп выбирают как эталонную (группа 0) и определяют фиктивные переменные для остальных.

Например, если качественный признак имеет три значения, то две фиктивные переменные определяются следующим образом:

группа 0: z1=z21 = 0,
группа 1: z1= 1, z2 =0,
группа2: z1= 0, z2 =1

Введение в регрессию фиктивных переменных существенно улучшает качество ее оценивания.

Метод наименьших квадратов (МНК). Свойства оценок на основе МНК

Возможны разные виды уравнений множественной регрессии: линейные и нелинейные.

Ввиду четкой интерпретации параметров наиболее широко используется линейная функция. В линейной множественной регрессии $y=a+b_1x_1+b_2x_2+...+b_mx_m$ параметры при называются коэффициентами «чистой» регрессии. Они характеризуют среднее изменение результата с изменением соответствующего фактора на единицу при неизмененном значении других факторов, закрепленных на среднем уровне.

Рассмотрим линейную модель множественной регрессии

$$y=a+b_1x_1+b_2x_2+...+b_mx_m+\varepsilon$$

Классический подход к оцениванию параметров линейной модели множественной регрессии основан на методе наименьших квадратов (МНК). МНК позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака y от расчетных y минимальна:

Формула 19.2

$$\sum_i \left(y_i - \hat{y}_{x_i}\right)^2 \rightarrow \min$$

Как известно из курса математического анализа, для того чтобы найти экстремум функции нескольких переменных, надо вычислить частные производные первого порядка по каждому из параметров и приравнять их к нулю.

Итак. Имеем функцию m+1 аргумента:

$$S(a, b_1, b_2, ..., b_m) = \sum \left(y - a - b_1x_1 - b_2x_2 - ... - b_mx_m\right)^2$$

Находим частные производные первого порядка:

$$\begin{cases} \dfrac{\partial S}{\partial a} = -2\sum \left(y - a - b_1x_1 - b_2x_2 - ... - b_mx_m\right) = 0; \\ \dfrac{\partial S}{\partial b_1} = -2\sum x_1\left(y - a - b_1x_1 - b_2x_2 - ... - b_mx_m\right) = 0; \\ ..... \\ \dfrac{\partial S}{\partial b_m} = -2\sum x_m\left(y - a - b_1x_1 - b_2x_2 - ... - b_mx_m\right) = 0. \end{cases}$$

После элементарных преобразований приходим к системе линейных нормальных уравнений для нахождения параметров линейного уравнения множественной регрессии (19.3):

Формула 19.3

$$\begin{cases} na + b_1\sum x_1 + b_2\sum x_2 + \ldots + b_m\sum x_m = \sum y, \\ a\sum x_1 + b_1\sum x_1^2 + b_2\sum x_1x_2 + \ldots + b_m\sum x_1x_m = \sum yx_1, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a\sum x_m + b_1\sum x_1x_m + b_2\sum x_2x_m + \ldots + b_m\sum x_m^2 = \sum yx_m. \end{cases}$$

Для двухфакторной модели данная система будет иметь вид:

$$\begin{cases} na + b_1\sum x_1 + b_2\sum x_2 = \sum y, \\ a\sum x_1 + b_1\sum x_1^2 + b_2\sum x_1x_2 = \sum yx_1, \\ a\sum x_2 + b_1\sum x_1x_2 + b_2\sum x_2^2 = \sum yx_2. \end{cases}$$

Метод наименьших квадратов применим и к уравнению множественной регрессии в стандартизированном масштабе:

Формула 19.4

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \ldots + \beta_m t_{x_m} + \varepsilon,$$

где t_y, t_{x1}, ..., t_{xm} – стандартизированные переменные: $t_y = \frac{y - \overline{y}}{\sigma_y}$, $t_y = \frac{y - \overline{y}}{\sigma_y}$, для которых среднее значение равно нулю: $\overline{t}_y = \overline{t}_{x_i} = 0$, а среднее квадратическое отклонение равно единице: $\sigma_{t_y} = \sigma_{t_{x_i}} = 1$; – стандартизированные коэффициенты регрессии.

Стандартизованные коэффициенты регрессии показывают, на сколько единиц изменится в среднем результат, если соответствующий фактор x_i изменится на одну единицу при неизменном среднем уровне других факторов. В силу того, что все переменные заданы как центрированные и нормированные, стандартизованные коэффициенты регрессии β_i можно сравнивать между собой. Сравнивая их друг с другом, можно ранжировать факторы по силе их воздействия на результат. В этом основное достоинство стандартизованных коэффициентов регрессии в отличие от коэффициентов «чистой» регрессии, которые несравнимы между собой.

Применяя МНК к уравнению множественной регрессии в стандартизированном масштабе, получим систему нормальных уравнений вида

Формула 19.5

$$\begin{cases} r_{yx_1} = \beta_1 + \beta_2 r_{x_1x_2} + \beta_3 r_{x_1x_3} + \ldots + \beta_m r_{x_1x_m}, \\ r_{yx_2} = \beta_1 r_{x_1x_2} + \beta_2 + \beta_3 r_{x_1x_3} + \ldots + \beta_m r_{x_1x_m}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ r_{yx_m} = \beta_1 r_{x_1x_m} + \beta_2 r_{x_2x_m} + \beta_3 r_{x_3x_m} + \ldots + \beta_m, \end{cases}$$

где r_{yxi} и r_{xixj} – коэффициенты парной и межфакторной корреляции.

Коэффициенты «чистой» регрессии b_i связаны со стандартизованными коэффициентами регрессии β_i следующим образом:

Формула 19.6

$$b_i = \beta_i \frac{\sigma_y}{\sigma_{x_i}}$$

Поэтому можно переходить от уравнения регрессии в стандартизованном масштабе (2.4) к уравнению регрессии в натуральном масштабе переменных (19.1), при этом параметр a определяется как

$$a = \overline{y} - b_1\overline{x}_1 - b_2\overline{x}_2 - \ldots - b_m\overline{x}_m$$

Рассмотренный смысл стандартизованных коэффициентов регрессии позволяет их использовать при отсеве факторов – из модели исключаются факторы с наименьшим значением β_i.

На основе линейного уравнения множественной регрессии

Формула 19.7

$$y = a + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m + \varepsilon$$

могут быть найдены частные уравнения регрессии:

Формула 19.8

$$\begin{cases} y_{x_1 \cdot x_2, x_3, \ldots, x_m} = \varphi(x_1), \\ y_{x_2 \cdot x_1, x_3, \ldots, x_m} = \varphi(x_2), \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ y_{x_m \cdot x_1, x_2, \ldots, x_{m-1}} = \varphi(x_m), \end{cases}$$

т.е. уравнения регрессии, которые связывают результативный признак с соответствующим фактором при закреплении остальных факторов на среднем уровне. В развернутом виде систему можно переписать в виде:

$$\begin{cases} y_{x_1 \cdot x_2, x_3, \ldots, x_m} = a + b_1 x_1 + b_2 \bar{x}_2 + b_3 \bar{x}_3 + \ldots + b_m \bar{x}_m + \varepsilon, \\ y_{x_2 \cdot x_1, x_3, \ldots, x_m} = a + b_1 \bar{x}_1 + b_2 x_2 + b_3 \bar{x}_3 + \ldots + b_m \bar{x}_m + \varepsilon, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_{x_m \cdot x_1, x_2, \ldots, x_{m-1}} = a + b_1 \bar{x}_1 + b_2 \bar{x}_2 + b_3 \bar{x}_3 + \ldots + b_m x_m + \varepsilon. \end{cases}$$

При подстановке в эти уравнения средних значений соответствующих факторов они принимают вид парных уравнений линейной регрессии, т.е. имеем

Формула 19.9

$$\begin{cases} y_{x_1 \cdot x_2, x_3, \ldots, x_m} = A_1 + b_1 x_1, \\ y_{x_2 \cdot x_1, x_3, \ldots, x_m} = A_2 + b_2 x_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_{x_m \cdot x_1, x_2, \ldots, x_{m-1}} = A_m + b_m x_m, \end{cases}$$

Где

$$\begin{cases} A_1 = a + b_2 \bar{x}_2 + b_3 \bar{x}_3 + \ldots + b_m \bar{x}_m, \\ A_2 = a + b_1 \bar{x}_1 + b_3 \bar{x}_3 + \ldots + b_m \bar{x}_m, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ A_m = a + b_1 \bar{x}_1 + b_2 \bar{x}_2 + b_3 \bar{x}_3 + \ldots + b_{m-1} x_{m-1}. \end{cases}$$

В отличие от парной регрессии частные уравнения регрессии характеризуют изолированное влияние фактора на результат, ибо другие факторы закреплены на неизменном уровне. Эффекты влияния других факторов присоединены в них к свободному члену уравнения множественной регрессии. Это позволяет на основе частных уравнений регрессии определять частные коэффициенты эластичности:

Формула 19.10

$$Э_{y_{x_i}} = b_i \wedge \frac{x_i}{y_{x_i \cdot x_1, x_2, \ldots x_{i-1}, x_{i+1}, \ldots, x_m}}$$

где b_i – коэффициент регрессии для фактора x_i в уравнении множественной регрессии, $y_{x_i \cdot x_1, x_2, \ldots x_{i-1}, x_{i+1}, \ldots, x_m}$ — частное уравнение регрессии.

Наряду с частными коэффициентами эластичности могут быть найдены средние по совокупности показатели эластичности:

Формула 19.11

$$\bar{Э}_i = b_i \cdot \frac{\bar{x}_i}{\bar{y}_{x_i}}$$

Которые показывают на сколько процентов в среднем изменится результат, при изменении соответствующего фактора на 1%. Средние показатели эластичности можно сравнивать друг с другом и соответственно ранжировать факторы по силе их воздействия на результат.

Рассмотрим пример (для сокращения объема вычислений ограничимся только десятью наблюдениями). Пусть имеются следующие данные (условные) о сменной добыче нефти на одного рабочего y (т), мощности буровой установки x_1 (м) и уровне автоматизации работ x_2 (%), характеризующие процесс добычи нефти на 10 месторождениях.

Таблица 19.2– Исходные данные

№	1	2	3	4	5	6	7	8	9	10
x_1	8	11	12	9	8	8	9	9	8	12
x_2	5	8	8	5	7	8	6	4	5	7
y	5	10	10	7	5	6	6	5	6	8

Предполагая, что между переменными y, x_1, x_2 существует линейная корреляционная зависимость, найдем уравнение регрессии y по x_1 и x_2.

Для удобства дальнейших вычислений составляем таблицу ($\varepsilon=y-\hat{y}_x$):

Таблица 19.3 – Расчетная таблица

№	x_1	x_2	y	x_1^2	x_2^2	y^2	x_1*x_2	x_1*y	x_2*y	y_x	ε^2
1	2	3	4	5	6	7	8	9	10	11	12
1	8	5	5	64	25	25	40	40	25	5,13	0,016
2	11	8	10	121	64	100	88	110	80	8,79	1,464
3	12	8	10	144	64	100	96	120	80	9,64	0,127
4	9	5	7	81	25	49	45	63	35	5,98	1,038
5	8	7	5	64	49	25	56	40	35	5,86	0,741
6	8	8	6	64	64	36	64	48	48	6,23	0,052
7	9	6	6	81	36	36	54	54	36	6,35	0,121
8	9	4	5	81	16	25	36	45	20	5,61	0,377
9	8	5	6	64	25	36	40	48	30	5,13	0,762
10	12	7	8	144	49	64	84	96	56	9,28	1,631
Сумма	94	63	68	908	417	496	603	664	445	68	6,329
Среднее значение	9,4	6,3	6,8	90,8	41,7	49,6	60,3	66,4	44,5	–	–
σ^2	2,44	2,01	3,36	–	–	–	–	–	–	–	–
σ	1,56	1,42	1,83	–	–	–	–	–	–	–	–

Для нахождения параметров уравнения регрессии в данном случае необходимо решить следующую систему нормальных уравнений:

$$\begin{cases} 10a + 94b_1 + 63b_2 = 68, \\ 94a + 908b_1 + 603b_2 = 664, \\ 63a + 603b_1 + 417b_2 = 445. \end{cases}$$

Откуда получаем, что a=-3,54, b_1=0,854, b_2=0,367. Т.е. получили следующее уравнение множественной регрессии:

$$y_x=-3,54+0,854*x_1+0,367*x_2$$

Оно показывает, что при увеличении только мощности пласта x_1 (при неизменном x_2) на 1 м добыча угля на одного рабочего у увеличится в среднем на 0,854 т, а при увеличении только уровня механизации работ x_2 (при неизменном x_1) на 1% – в среднем на 0,367 т.

Найдем уравнение множественной регрессии в стандартизованном масштабе:

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \varepsilon,$$

при этом стандартизованные коэффициенты регрессии будут

$$\beta_1 = b_1 \frac{\sigma_{x_1}}{\sigma_y} = 0,854 \cdot \frac{1,56}{1,83} = 0,728$$

$$\beta_2 = b_2 \frac{\sigma_{x_2}}{\sigma_y} = 0,367 \cdot \frac{1,42}{1,83} = 0,285$$

Т.е. уравнение будет выглядеть следующим образом:

$$t_y = 0,728 \cdot t_{x_1} + 0,285 \cdot t_{x_2}$$

Так как стандартизованные коэффициенты регрессии можно сравнивать между собой, то можно сказать, что мощность пласта оказывает большее влияние на сменную добычу угля, чем уровень механизации работ.

Сравнивать влияние факторов на результат можно также при помощи средних коэффициентов эластичности (19.11):

$$\bar{Э}_i = b_i \cdot \frac{\bar{x}_i}{\bar{y}_{x_i}}$$

Вычисляем:

$$\bar{Э}_1 = 0,854 \cdot \frac{9,4}{6,8} = 1,18 \qquad \bar{Э}_2 = 0,367 \cdot \frac{6,3}{6,8} = 0,34$$

Т.е. увеличение только мощности пласта (от своего среднего значения) или только уровня механизации работ на 1% увеличивает в среднем сменную добычу угля на 1,18% или 0,34% соответственно. Таким образом, подтверждается большее влияние на результат у фактора x_1, чем фактора x_2.

Выводы:

Модель множественной регрессии, это модель, где имеется несколько независимых факторов и один зависимый.

Среди независимых факторов возможно наличие мультиколлинеарных факторов, от которых необходимо избавляться.

Для оценки качественных признаков фактора, возможно использование фиктивных переменных.

Существуют коэффициенты чистой регрессии и стандартизированные коэффициенты регрессии.