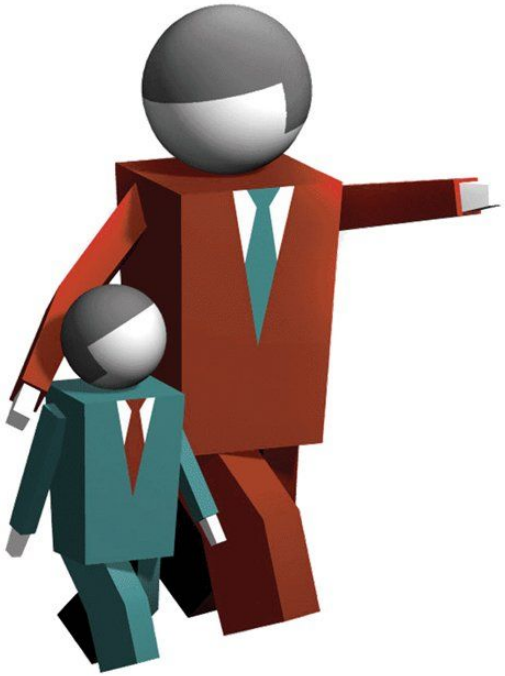


Часть 3. Множественная регрессия и корреляция

Голлай Александр

Множественная регрессия

Парная регрессия может дать хороший результат при моделировании, если влиянием других факторов, воздействующих на объект исследования, можно пренебречь. Если же этим влиянием пренебречь нельзя, то в этом случае следует попытаться выявить влияние других факторов, введя их в модель, т.е. построить уравнение множественной регрессии.

Множественная регрессия

Уравнение множественной регрессии:

$$y = \hat{f}(x_1, x_2, \ldots, x_m)$$

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

Множественная регрессия. Применение

Множественная регрессия широко используется в решении проблем спроса, доходности акций, при изучении функции издержек производства, в макроэкономических расчетах и целом ряде других вопросов эконометрики. В настоящее время множественная регрессия – один из наиболее распространенных методов в эконометрике. 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

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

1. отбор факторов,
2. выбор вида уравнения регрессии.

Построение уравнения множественной 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.

Построение уравнения множественной регрессии. Требования к факторам.

Факторы, включаемые во множественную регрессию, должны отвечать следующим требованиям.

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

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

Интеркоррелированность

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.

Построение уравнения множественной регрессии. Коэффициент детерминации.

Включаемые во множественную регрессию факторы должны объяснить вариацию независимой переменной. Если строится модель с набором m 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.

Построение уравнения множественной регрессии. Коэффициент детерминации.

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

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

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.

Построение уравнения множественной регрессии. Коэффициент детерминации.

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.

Построение уравнения множественной регрессии

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.

Построение уравнения множественной 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 $r_{x_i x_j} \geq 0{,}7$.

Построение уравнения множественной регрессии.

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.

Построение уравнения множественной регрессии. Подбор переменных.

Пусть, например, при изучении зависимости

$$y = \hat{f}(x_1, x_2, x_3)$$

матрица парных коэффициентов корреляции оказалась следующей:

	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

Построение уравнения множественной регрессии. Подбор переменных.

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 $\left(r_{yx_2} = 0{,}7 < r_{yx_1} = 0{,}8\right)$, но зато значительно слабее межфакторная корреляция $r_{x_2x_3} = 0{,}2 < r_{x_1x_3} = 0{,}5$. Поэтому в данном случае в уравнение множественной регрессии включаются факторы x_2, x_3.

Построение уравнения множественной регрессии. Подбор переменных.

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.

Построение уравнения множественной регрессии. Подбор переменных.

Включение в модель мультиколлинеарных факторов нежелательно в силу следующих последствий:

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

2. 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.

Построение уравнения множественной регрессии. Матрица парных коэффициентов

Для оценки мультиколлинеарности факторов может использоваться определитель матрицы парных коэффициентов корреляции между факторами.

Если бы факторы не коррелировали между собой, то матрица парных коэффициентов корреляции между факторами была бы единичной матрицей, поскольку все недиагональные элементы $r_{x_i x_j}$ $(i \neq j)$. были бы равны нулю.

Построение уравнения множественной регрессии. Матрица парных коэффициентов

Так, для уравнения, включающего три объясняющих переменных

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + b_3 x_3$$

матрица коэффициентов корреляции между факторами имела бы определитель, равный единице:

$$\mathrm{Det}\,\mathbf{R} = \begin{vmatrix} r_{x_1x_1} & r_{x_1x_2} & r_{x_1x_3} \\ r_{x_2x_1} & r_{x_2x_2} & r_{x_2x_3} \\ r_{x_3x_1} & r_{x_3x_2} & r_{x_3x_3} \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 1$$

Построение уравнения множественной регрессии. Матрица парных коэффициентов

Если же, наоборот, между факторами существует полная линейная зависимость и все коэффициенты корреляции равны единице, то определитель такой матрицы равен нулю:

$$\mathrm{Det}\,\mathbf{R} = \begin{vmatrix} r_{x_1x_1} & r_{x_1x_2} & r_{x_1x_3} \\ r_{x_2x_1} & r_{x_2x_2} & r_{x_2x_3} \\ r_{x_3x_1} & r_{x_3x_2} & r_{x_3x_3} \end{vmatrix} = \begin{vmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{vmatrix} = 0$$

Построение уравнения множественной регрессии. Матрица парных коэффициентов

Чем ближе к нулю определитель матрицы межфакторной корреляции, тем сильнее мультиколлинеарность факторов и ненадежнее результаты множественной регрессии. И, наоборот, чем ближе к единице определитель матрицы межфакторной корреляции, тем меньше мультиколлинеарность факторов.

Подходы к преодолению сильной межфакторной корреляции.

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.

Подходы к преодолению сильной межфакторной корреляции.

Рассматриваемое уравнение включает взаимодействие первого порядка (взаимодействие двух факторов). Возможно включение в модель и взаимодействий более высокого порядка, если будет доказана их статистическая значимость по F-критерию Фишера, но, как правило, взаимодействия третьего и более высоких порядков оказываются статистически незначимыми.

Подходы к преодолению сильной межфакторной корреляции.

Одним из путей учета внутренней корреляции факторов является переход к совмещенным уравнениям регрессии, т.е. к уравнениям, которые отражают не только влияние факторов, но и их взаимодействие. Так, если $y = f(x_1, x_2, x_3)$, то возможно построение следующего совмещенного уравнения:

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + b_{12}x_1x_2 + \\ +b_{13}x_1x_3 + b_{23}x_2x_3 + \varepsilon$$

Отбор факторов

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.

Отбор факторов

Наиболее широкое применение получили следующие методы построения уравнения множественной регрессии:

1. Метод исключения – отсев факторов из полного его набора.

2. Метод включения – дополнительное введение фактора.

3. Шаговый регрессионный анализ – исключение ранее введенного фактора.

Отбор факторов

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.

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

Виды уравнений множественной регрессии

Возможны разные виды уравнений множественной регрессии:

- линейные,
- нелинейные.

Ввиду четкой интерпретации параметров наиболее широко используется линейная функция.

Линейная множественная регрессия

Линейная множественная регрессия

В линейной множественной регрессии

$$\hat{y}_x = a + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m$$

параметры при x называются коэффициентами «чистой» регрессии. Они характеризуют среднее изменение результата с изменением соответствующего фактора на единицу при неизмененном значении других факторов, закрепленных на среднем уровне.

Линейная множественная регрессия

Рассмотрим линейную модель множественной регрессии

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

Классический подход к оцениванию параметров линейной модели множественной регрессии основан на методе наименьших квадратов (МНК).

Линейная множественная регрессия

МНК позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака y от расчетных $\hat{y}$ минимальна:

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

Линейная множественная регрессия

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.

Линейная множественная регрессия

Итак. Имеем функцию $m + 1$ аргумента:

$$S(a,\ b_1,\ b_2,\ \ldots,\ b_m) =$$

$$= \sum (y - a - b_1 x_1 - b_2 x_2 - \ldots - b_m x_m)^2$$

Линейная множественная регрессия

Находим частные производные первого порядка:

$$\begin{cases} \dfrac{\partial S}{\partial a} = -2\sum (y - a - b_1 x_1 - b_2 x_2 - \ldots - b_m x_m) = 0; \\ \dfrac{\partial S}{\partial b_1} = -2\sum x_1 (y - a - b_1 x_1 - b_2 x_2 - \ldots - b_m x_m) = 0; \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \dfrac{\partial S}{\partial b_m} = -2\sum x_m (y - a - b_1 x_1 - b_2 x_2 - \ldots - b_m x_m) = 0. \end{cases}$$

Линейная множественная регрессия

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

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

Линейная множественная регрессия

Для двухфакторной модели данная система будет иметь вид:

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

Линейная множественная регрессия

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

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

где $t_y, t_{x_1}, \ldots, t_{x_m}$ – стандартизированные переменные: $t_y = \frac{y-\bar{y}}{\sigma_y}$, $t_{x_i} = \frac{x_i - \bar{x}_i}{\sigma_{x_i}}$, для которых среднее значение равно нулю: $\bar{t}_y = \bar{t}_{x_i} = 0$, а среднее квадратическое отклонение равно единице: $\sigma_{t_y} = \sigma_{t_{x_i}} = 1$. β_i - – стандартизированные коэффициенты регрессии.

Линейная множественная регрессия

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.

Линейная множественная регрессия

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.

Линейная множественная регрессия

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

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

где r_{yx_i} и $r_{x_ix_j}$ – коэффициенты парной и межфакторной корреляции.

Линейная множественная регрессия

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

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

Линейная множественная регрессия

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

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

Линейная множественная регрессия

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

Линейная множественная регрессия

На основе линейного уравнения множественной регрессии

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

могут быть найдены частные уравнения регрессии:

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

Линейная множественная регрессия

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

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

Линейная множественная регрессия

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

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

Линейная множественная регрессия

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

где

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

Линейная множественная регрессия

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.

Линейная множественная регрессия

Это позволяет на основе частных уравнений регрессии определять частные коэффициенты эластичности:

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

где b_i – коэффициент регрессии для фактора x_i в уравнении множественной регрессии,

$\hat{y}_{x_i \cdot x_1, x_2, \ldots x_{i-1}, x_{i+1}, \ldots, x_m}$ – частное уравнение регрессии.

Линейная множественная регрессия

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

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

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.

Пример

Условия

Рассмотрим пример (для сокращения объема вычислений ограничимся только десятью наблюдениями). Пусть имеются следующие данные (условные) о сменной добыче угля на одного рабочего y (т) , мощности пласта x_1 (м) и уровне механизации работ x_2 (%), , характеризующие процесс добычи угля в 10 шахтах.

№	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

Решение

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

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

Решение

№	x_1	x_2	y	x_1^2	x_2^2	y^2	$x_1 \cdot x_2$	$x_1 \cdot y$	$x_2 \cdot y$	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	–	–	–	–	–	–	–	–

Решение

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

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

Решение

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

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

$$\hat{y}_x = -3{,}54 + 0{,}854 \cdot x_1 + 0{,}367 \cdot x_2$$

Решение

$$\hat{y}_x = -3{,}54 + 0{,}854 \cdot x_1 + 0{,}367 \cdot x_2$$

Оно показывает, что при увеличении только мощности пласта x_1 (при неизменном x_2) на 1 м добыча угля на одного рабочего y 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.

Решение

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

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

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

Решение

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:

$$\hat{t}_y = 0{,}728 \cdot t_{x_1} + 0{,}285 \cdot t_{x_2}$$

Решение

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.

Решение

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

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

Вычисляем:

$$\overline{Э}_1 = 0{,}854 \cdot \frac{9{,}4}{6{,}8} = 1{,}18$$

$$\overline{Э}_2 = 0{,}367 \cdot \frac{6{,}3}{6{,}8} = 0{,}34$$

Решение

$$\overline{Э}_1 = 0{,}854 \cdot \frac{9{,}4}{6{,}8} = 1{,}18$$

$$\overline{Э}_2 = 0{,}367 \cdot \frac{6{,}3}{6{,}8} = 0{,}34$$

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 y фактора x_1, , чем фактора x_2.

Проверка существенности факторов и показатели качества регрессии

Практическая значимость уравнения множественной регрессии оценивается с помощью показателя множественной корреляции и его квадрата – показателя детерминации.

Показатель множественной корреляции характеризует тесноту связи рассматриваемого набора факторов с исследуемым признаком или, иначе, оценивает тесноту совместного влияния факторов на результат.

Показатель множественной корреляции

Независимо от формы связи показатель множественной корреляции может быть найден как:

$$R_{yx_1x_2\ldots x_m} = \sqrt{1 - \frac{\sigma^2_{ост}}{\sigma^2_y}}$$

где σ^2_y – общая дисперсия результативного признака; $\sigma^2_{ост}$ – остаточная дисперсия.

Показатель множественной корреляции

Границы изменения индекса множественной корреляции от 0 до 1. Чем ближе его значение к 1, тем теснее связь результативного признака со всем набором исследуемых факторов. Величина индекса множественной корреляции должна быть больше или равна максимальному парному индексу корреляции:

$$R_{yx_1x_2\ldots x_m} \geq r_{yx_i(\max)} \quad (i = \overline{1,m})$$

Показатель множественной корреляции

При правильном включении факторов в регрессионную модель величина индекса множественной корреляции будет существенно отличаться от индекса корреляции парной зависимости. Если же дополнительно включенные в уравнение множественной регрессии факторы третьестепенны, то индекс множественной корреляции может практически совпадать с индексом парной корреляции (различия в третьем, четвертом знаках). Отсюда ясно, что сравнивая индексы множественной и парной корреляции, можно сделать вывод о целесообразности включения в уравнение регрессии того или иного фактора.

Показатель множественной корреляции

Расчет индекса множественной корреляции предполагает определение уравнения множественной регрессии и на его основе остаточной дисперсии:

$$\sigma^2_{ост} = \frac{1}{n}\sum\left(y - \hat{y}_{x_1 x_2 \ldots x_m}\right)^2$$

Можно пользоваться следующей формулой индекса множественной детерминации:

$$R^2_{yx_1x_2\ldots x_m} = 1 - \frac{\sum\left(y - \hat{y}_{x_1 x_2 \ldots x_m}\right)^2}{\sum(y - \bar{y})^2}$$

Показатель множественной корреляции

При линейной зависимости признаков формула индекса множественной корреляции может быть представлена следующим выражением:

$$R_{yx_1x_2\ldots x_m} = \sqrt{\sum \beta_i \cdot r_{yx_i}}$$

где β_i – стандартизованные коэффициенты регрессии; r_{yx_i} – парные коэффициенты корреляции результата с каждым фактором.

Показатель множественной корреляции

Формула индекса множественной корреляции для линейной регрессии получила название линейного коэффициента множественной корреляции, или, что то же самое, совокупного коэффициента корреляции.

Показатель множественной корреляции

Возможно также при линейной зависимости определение совокупного коэффициента корреляции через матрицу парных коэффициентов корреляции:

$$R_{yx_1x_2,\ldots,x_p} = \sqrt{1 - \frac{\Delta r}{\Delta r_{11}}}$$

Показатель множественной корреляции

$$R_{yx_1x_2,\ldots,x_p} = \sqrt{1 - \frac{\Delta r}{\Delta r_{11}}}$$

– определитель матрицы парных коэффициентов корреляции;

$$\Delta r = \begin{vmatrix} 1 & r_{yx_1} & r_{yx_2} & \ldots & r_{yx_p} \\ r_{yx_1} & 1 & r_{x_1x_2} & \ldots & r_{x_1x_p} \\ r_{yx_2} & r_{x_2x_1} & 1 & \ldots & r_{x_2x_p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{yx_p} & r_{x_px_1} & r_{x_px_2} & \ldots & 1 \end{vmatrix}$$

Показатель множественной корреляции

$$R_{yx_1x_2,\ldots,x_p} = \sqrt{1 - \frac{\Delta r}{\Delta r_{11}}}$$

– определитель матрицы межфакторной корреляции.

$$\Delta r_{11} = \begin{vmatrix} 1 & r_{x_1x_2} & \ldots & r_{x_1x_p} \\ r_{x_2x_1} & 1 & \ldots & r_{x_2x_p} \\ \ldots & \ldots & \ldots & \ldots \\ r_{x_px_1} & r_{x_px_2} & \ldots & 1 \end{vmatrix}$$

Показатель множественной корреляции

Как видим, величина множественного коэффициента корреляции зависит не только от корреляции результата с каждым из факторов, но и от межфакторной корреляции. Рассмотренная формула позволяет определять совокупный коэффициент корреляции, не обращаясь при этом к уравнению множественной регрессии, а используя лишь парные коэффициенты корреляции.

Показатель множественной корреляции

В рассмотренных показателях множественной корреляции (индекс и коэффициент) используется остаточная дисперсия, которая имеет систематическую ошибку в сторону преуменьшения, тем более значительную, чем больше параметров определяется в уравнении регрессии при заданном объеме наблюдений n. Если число параметров при x_i равно m и приближается к объему наблюдений, то остаточная дисперсия будет близка к нулю и коэффициент (индекс) корреляции приблизится к единице даже при слабой связи факторов с результатом. Для того чтобы не допустить возможного преувеличения тесноты связи, используется скорректированный индекс (коэффициент) множественной корреляции.

Скорректированный индекс множественной корреляции

Скорректированный индекс множественной корреляции содержит поправку на число степеней свободы, а именно остаточная сумма квадратов $\sum\left(y-\hat{y}_{x_1x_2\ldots x_m}\right)^2$ делится на число степеней свободы остаточной вариации $(n-m-1)$, а общая сумма квадратов отклонений $\sum(y-\bar{y})^2$ на число степеней свободы в целом по совокупности $(n-1)$.

Скорректированный индекс множественной детерминации

Формула скорректированного индекса множественной детерминации имеет вид:

$$\widehat{R}^2 = 1 - \frac{\sum (y - \hat{y})^2 / (n - m - 1)}{\sum (y - \bar{y}) / (n - 1)}$$

где m – число параметров при переменных x; n – число наблюдений.

Скорректированный индекс множественной детерминации

Поскольку

$$\frac{\sum \left(y - \hat{y}_{x_1 x_2 \ldots x_m}\right)^2}{\sum (y - \bar{y})^2} = 1 - R^2$$

то величину скорректированного индекса детерминации можно представить в виде:

$$\widehat{R}^2 = 1 - \left(1 - R^2\right) \cdot \frac{n-1}{n-m-1}$$

Чем больше величина m, , тем сильнее различия $\widehat{R}^2$ и R^2.

Скорректированный индекс множественной детерминации

Как было показано выше, ранжирование факторов, участвующих во множественной линейной регрессии, может быть проведено через стандартизованные коэффициенты регрессии (β - коэффициенты). Эта же цель может быть достигнута с помощью частных коэффициентов корреляции (для линейных связей). Кроме того, частные показатели корреляции широко используются при решении проблемы отбора факторов: целесообразность включения того или иного фактора в модель можно доказать величиной показателя частной корреляции.

Частные коэффициенты корреляции

Частные коэффициенты корреляции характеризуют тесноту связи между результатом и соответствующим фактором при элиминировании (устранении влияния) других факторов, включенных в уравнение регрессии.

Показатели частной корреляции представляют собой отношение сокращения остаточной дисперсии за счет дополнительного включения в анализ нового фактора к остаточной дисперсии, имевшей место до введения его в модель.

Частные коэффициенты корреляции

В общем виде при наличии m факторов для уравнения

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

коэффициент частной корреляции, измеряющий влияние на y фактора x_i, при неизменном уровне других факторов, можно определить по формуле:

$$r_{yx_i \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_m} = \sqrt{1 - \frac{1 - R^2_{yx_1 x_2 \ldots x_i \ldots x_m}}{1 - R^2_{yx_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_m}}}$$

Частные коэффициенты корреляции

При двух факторах предыдущая формула примет вид:

$$r_{yx_1 \cdot x_2} = \sqrt{1 - \frac{1 - R^2_{yx_1x_2}}{1 - r^2_{yx_2}}}$$

$$r_{yx_2 \cdot x_1} = \sqrt{1 - \frac{1 - R^2_{yx_1x_2}}{1 - r^2_{yx_1}}}$$

Частные коэффициенты корреляции

$$r_{yx_i \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_m} = \sqrt{1 - \frac{1 - R^2_{yx_1 x_2 \ldots x_i \ldots x_m}}{1 - R^2_{yx_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_m}}}$$

где $R^2_{yx_1 x_2 \ldots x_i \ldots x_m}$ – множественный коэффициент детерминации всех m факторов с результатом; $R^2_{yx_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_m}$ – тот же показатель детерминации, но без введения в модель фактора x_i.

Частные коэффициенты корреляции

Порядок частного коэффициента корреляции определяется количеством факторов, влияние которых исключается. Например, $r_{yx_1 \cdot x_2}$ – коэффициент частной корреляции первого порядка. Соответственно коэффициенты парной корреляции называются коэффициентами нулевого порядка.

Частные коэффициенты корреляции

Коэффициенты частной корреляции более высоких порядков можно определить через коэффициенты частной корреляции более низких порядков по рекуррентной формуле:

$$r_{yx_i \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_m} = \frac{r_{yx_i \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_{m-1}} - r_{yx_m \cdot x_1 x_2 \ldots x_{m-1}} \cdot r_{x_i x_m \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_{m-1}}}{\sqrt{\left(1 - r^2_{yx_m \cdot x_1 x_2 \ldots x_{m-1}}\right) \cdot \left(1 - r^2_{x_i x_m \cdot x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_{m-1}}\right)}}$$

Частные коэффициенты корреляции

При двух факторах данная формула примет вид:

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_2}^2\right) \cdot \left(1 - r_{x_1x_2}^2\right)}}$$

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_1}^2\right) \cdot \left(1 - r_{x_1x_2}^2\right)}}$$

Частные коэффициенты корреляции

Для уравнения регрессии с тремя факторами частные коэффициенты корреляции второго порядка определяются на основе частных коэффициентов корреляции первого порядка. Так, по уравнению

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \varepsilon$$

возможно исчисление трех частных коэффициентов корреляции второго порядка:

$$r_{yx_1 \cdot x_2 x_3}, r_{yx_2 \cdot x_1 x_3}\ r_{yx_3 \cdot x_1 x_2}$$

Частные коэффициенты корреляции

$$r_{yx_1 \cdot x_2 x_3}, r_{yx_2 \cdot x_1 x_3}\ r_{yx_3 \cdot x_1 x_2}$$

каждый из которых определяется по рекуррентной формуле. Например, при $i = 1$ имеем формулу для расчета $r_{yx_1 \cdot x_2 x_3}$

$$r_{yx_1 \cdot x_2 x_3} = \frac{r_{yx_1 \cdot x_2} - r_{yx_3 \cdot x_2} \cdot r_{x_1 x_3 \cdot x_2}}{\sqrt{\left(1 - r^2_{yx_3 \cdot x_2}\right)\left(1 - r^2_{x_1 x_3 \cdot x_2}\right)}}$$

Частные коэффициенты корреляции

Рассчитанные по рекуррентной формуле частные коэффициенты корреляции изменяются в пределах от –1 до +1, а по формулам через множественные коэффициенты детерминации – от 0 до 1. Сравнение их друг с другом позволяет ранжировать факторы по тесноте их связи с результатом. Частные коэффициенты корреляции дают меру тесноты связи каждого фактора с результатом в чистом виде.

Частные коэффициенты корреляции

Если из стандартизованного уравнения регрессии $t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \beta_3 t_{x_3} + \varepsilon$ следует, что $\beta_1 > \beta_2 > \beta_3$, т.е. по силе влияния на результат порядок факторов таков: x_1 , x_2, x_3, то этот же порядок факторов определяется и по соотношению частных коэффициентов корреляции,

$$r_{yx_1 \cdot x_2 x_3} > r_{yx_2 \cdot x_1 x_3} > r_{yx_3 \cdot x_1 x_2}$$

Частные коэффициенты корреляции

В эконометрике частные коэффициенты корреляции обычно не имеют самостоятельного значения. Их используют на стадии формирования модели. Так, строя многофакторную модель, на первом шаге определяется уравнение регрессии с полным набором факторов и рассчитывается матрица частных коэффициентов корреляции. На втором шаге отбирается фактор с наименьшей и несущественной по t -критерию Стьюдента величиной показателя частной корреляции.

Частные коэффициенты корреляции

Исключив его из модели, строится новое уравнение регрессии. Процедура продолжается до тех пор, пока не окажется, что все частные коэффициенты корреляции существенно отличаются от нуля. Если исключен несущественный фактор, то множественные коэффициенты детерминации на двух смежных шагах построения регрессионной модели почти не отличаются друг от друга, $R^2_{m+1} \approx R^2_m$, где m – число факторов.

Частные коэффициенты корреляции

Из приведенных выше формул частных коэффициентов корреляции видна связь этих показателей с совокупным коэффициентом корреляции. Зная частные коэффициенты корреляции (последовательно первого, второго и более высокого порядка), можно определить совокупный коэффициент корреляции по формуле:

$$R_{yx_1x_2\ldots x_m} = \sqrt{1-\left(1-r^2_{yx_1}\right)\cdot\left(1-r^2_{yx_2\cdot x_1}\right)\cdot\left(1-r^2_{yx_3\cdot x_1x_2}\right)\cdot\ldots\cdot\left(1-r^2_{yx_m\cdot x_1x_2\ldots x_{m-1}}\right)}$$

Частные коэффициенты корреляции

В частности, для двухфакторного уравнения формула принимает вид:

$$R_{yx_1x_2\ldots x_m} = \sqrt{1 - \left(1 - r^2_{yx_1}\right) \cdot \left(1 - r^2_{yx_2 \cdot x_1}\right)}$$

Частные коэффициенты корреляции

При полной зависимости результативного признака от исследуемых факторов коэффициент совокупного их влияния равен единице. Из единицы вычитается доля остаточной вариации результативного признака $\left(1 - r^2\right)$ обусловленная последовательно включенными в анализ факторами. В результате подкоренное выражение характеризует совокупное действие всех исследуемых факторов.

Оценка значимости уравнения множественной регрессии

Значимость уравнения множественной регрессии в целом, так же как и в парной регрессии, оценивается с помощью F -критерия Фишера:

$$F = \frac{S_{факт}}{S_{ост}} = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m}$$

где $S_{факт}$ – факторная сумма квадратов на одну степень свободы; $S_{ост}$ – остаточная сумма квадратов на одну степень свободы; R^2 – коэффициент (индекс) множественной детерминации; m – число параметров при переменных x (в линейной регрессии совпадает с числом включенных в модель факторов); n – число наблюдений.

Оценка значимости уравнения множественной регрессии

Оценивается значимость не только уравнения в целом, но и фактора, дополнительно включенного в регрессионную модель. Необходимость такой оценки связана с тем, что не каждый фактор, вошедший в модель, может существенно увеличивать долю объясненной вариации результативного признака. Кроме того, при наличии в модели нескольких факторов они могут вводиться в модель в разной последовательности. Ввиду корреляции между факторами значимость одного и того же фактора может быть разной в зависимости от последовательности его введения в модель. Мерой для оценки включения фактора в модель служит частный F -критерий, т.е. F_{x_i}

Оценка значимости уравнения множественной регрессии

Частный F –критерий построен на сравнении прироста факторной дисперсии, обусловленного влиянием дополнительно включенного фактора, с остаточной дисперсией на одну степень свободы по регрессионной модели в целом. В общем виде для фактора x_i частный F –критерий определится как

$$F_{x_i} = \frac{R^2_{yx_1 \dots x_i \dots x_m} - R^2_{yx_1 \dots x_{i-1} x_{i+1} \dots x_m}}{1 - R^2_{yx_1 \dots x_i \dots x_m}} \cdot \frac{n - m - 1}{1}$$

Оценка значимости уравнения множественной регрессии

$$F_{x_i} = \frac{R^2_{yx_1 \dots x_i \dots x_m} - R^2_{yx_1 \dots x_{i-1} x_{i+1} \dots x_m}}{1 - R^2_{yx_1 \dots x_i \dots x_m}} \cdot \frac{n - m - 1}{1}$$

где $R^2_{yx_1 \dots x_i \dots x_m}$ – коэффициент множественной детерминации для модели с полным набором факторов, $R^2_{yx_1 \dots x_{i-1} x_{i+1} \dots x_m}$ – тот же показатель, но без включения в модель фактора x_i, n – число наблюдений, m – число параметров в модели (без свободного члена).

Оценка значимости уравнения множественной регрессии

Фактическое значение частного F – критерий сравнивается с табличным при уровне значимости α и числе степеней свободы: 1 и $n-m-1$. Если фактическое значение F_{x_i} превышает $F_{табл}(\alpha, k_1, k_2)$, то дополнительное включение фактора x_i в модель статистически оправданно и коэффициент чистой регрессии b_i при факторе x_i статистически значим. Если же фактическое значение F_{x_i} меньше табличного, то дополнительное включение в модель фактора x_i не увеличивает существенно долю объясненной вариации признака y , следовательно, нецелесообразно его включение в модель; коэффициент регрессии при данном факторе в этом случае статистически незначим.

Оценка значимости уравнения множественной регрессии

Для двухфакторного уравнения частные F – критерий имеют вид:

$$F_{x_1} = \frac{R^2_{yx_1x_2} - r^2_{yx_2}}{1 - R^2_{yx_1x_2}} \cdot (n-3)$$

$$F_{x_2} = \frac{R^2_{yx_1x_2} - r^2_{yx_1}}{1 - R^2_{yx_1x_2}} \cdot (n-3)$$

Оценка значимости уравнения множественной регрессии

С помощью частного F –критерия можно проверить значимость всех коэффициентов регрессии в предположении, что каждый соответствующий фактор x_i вводился в уравнение множественной регрессии последним.

Частный F – критерий оценивает значимость коэффициентов чистой регрессии. Зная величину F_{x_i} можно определить и t -критерий для коэффициента регрессии при i -м факторе, t_{b_i}, а именно:

$$t_{b_i} = \sqrt{F_{x_i}}$$

Оценка значимости уравнения множественной регрессии

Оценка значимости коэффициентов чистой регрессии по t –критерию Стьюдента может быть проведена и без расчета частных F -критериев. В этом случае, как и в парной регрессии, для каждого фактора используется формула:

$$t_{b_i} = \frac{b_i}{m_{b_i}}$$

где b_i– коэффициент чистой регрессии при факторе x_i, m_{b_i} – средняя квадратическая (стандартная) ошибка коэффициента регрессии b_i.

Оценка значимости уравнения множественной регрессии

Для уравнения множественной регрессии

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m$$

средняя квадратическая ошибка коэффициента регрессии может быть определена по следующей формуле:

$$m_{b_i} = \frac{\sigma_y \sqrt{1 - R^2_{yx_1 \ldots x_m}}}{\sigma_{x_i} \sqrt{1 - R^2_{x_i x_1 \ldots x_m}}} \cdot \frac{1}{\sqrt{n - m - 1}}$$

Оценка значимости уравнения множественной регрессии

$$m_{b_i} = \frac{\sigma_y \sqrt{1 - R^2_{yx_1 \dots x_m}}}{\sigma_{x_i} \sqrt{1 - R^2_{x_i x_1 \dots x_m}}} \cdot \frac{1}{\sqrt{n - m - 1}}$$

где σ_y – среднее квадратическое отклонение для признака y, σ_{x_i} – среднее квадратическое отклонение для признака x_i, $R^2_{yx_1 \dots x_m}$ – коэффициент детерминации для уравнения множественной регрессии, $R^2_{x_i x_1 \dots x_m}$ – коэффициент детерминации для зависимости фактора x_i со всеми другими факторами уравнения множественной регрессии; $n - m - 1$ – число степеней свободы для остаточной суммы квадратов отклонений.

Оценка значимости уравнения множественной регрессии

Как видим, чтобы воспользоваться данной формулой, необходимы матрица межфакторной корреляции и расчет по ней соответствующих коэффициентов детерминации $R^2_{x_i x_1 \dots x_m}$. Так, для уравнения

$$\hat{y} = a + b_1 x_1 + b_2 x_2 + b_3 x_3$$

оценка значимости коэффициентов регрессии b_1, b_2, b_3 предполагает расчет трех межфакторных коэффициентов детерминации: $R^2_{x_1 \cdot x_2 x_3}$, $R^2_{x_2 \cdot x_1 x_3}$, $R^2_{x_3 \cdot x_1 x_2}$.

Оценка значимости уравнения множественной регрессии

Взаимосвязь показателей частного коэффициента корреляции, частного F-критерия и t - критерия Стьюдента для коэффициентов чистой регрессии может использоваться в процедуре отбора факторов. Отсев факторов при построении уравнения регрессии методом исключения практически можно осуществлять не только по частным коэффициентам корреляции, исключая на каждом шаге фактор с наименьшим незначимым значением частного коэффициента корреляции, но и по величинам t_{b_i} и F_{x_i}. Частный F-критерий широко используется и при построении модели методом включения переменных и шаговым регрессионным методом.

Пример

Условия

Оценим качество уравнения, полученного в предыдущем разделе.

Решение

Сначала найдём значения парных коэффициентов корреляции:

$$r_{yx_1} = \frac{\overline{y \cdot x_1} - \bar{y} \cdot \bar{x}_1}{\sigma_y \cdot \sigma_{x_1}} = \frac{66{,}4 - 6{,}8 \cdot 9{,}4}{1{,}83 \cdot 1{,}56} = 0{,}869$$

$$r_{yx_2} = \frac{\overline{y \cdot x_2} - \bar{y} \cdot \bar{x}_2}{\sigma_y \cdot \sigma_{x_2}} = \frac{44{,}5 - 6{,}8 \cdot 6{,}3}{1{,}83 \cdot 1{,}42} = 0{,}639$$

$$r_{x_1x_2} = \frac{\overline{x_1 \cdot x_2} - \bar{x}_1 \cdot \bar{x}_2}{\sigma_{x_1} \cdot \sigma_{x_2}} = \frac{60{,}3 - 9{,}4 \cdot 6{,}3}{1{,}56 \cdot 1{,}42} = 0{,}488$$

Решение

Значения парных коэффициентов корреляции указывают на достаточно тесную связь сменной добычи угля на одного рабочего y с мощностью пласта x_1 и на умеренную связь с уровнем механизации работ x_2. В то же время межфакторная связь $r_{x_1x_2}$ не очень сильная ($r_{x_1x_2} = 0{,}49 < 0{,}7$), что говорит о том, что оба фактора являются информативными, т.е. и x_1, и x_2 необходимо включить в модель.

Решение

Теперь рассчитаем совокупный коэффициент корреляции $R_{yx_1x_2}$. Для этого сначала найдем определитель матрицы парных коэффициентов корреляции:

$$\Delta r = \begin{vmatrix} 1 & 0,87 & 0,64 \\ 0,87 & 1 & 0,49 \\ 0,64 & 0,49 & 1 \end{vmatrix} = 0,139064$$

и определитель матрицы межфакторной корреляции:

$$\Delta r_{11} = \begin{vmatrix} 1 & 0,49 \\ 0,49 & 1 \end{vmatrix} = 0,7599$$

Решение

Тогда коэффициент множественной корреляции по формуле

$$R_{yx_1x_2} = \sqrt{1 - \frac{\Delta r}{\Delta r_{11}}} = \sqrt{1 - \frac{0{,}139064}{0{,}7599}} = 0{,}904$$

Т.е. можно сказать, что 81,7% (коэффициент детерминации $R^2_{yx_1x_2} = 0{,}817$) вариации результата объясняется вариацией представленных в уравнении признаков, что указывает на весьма тесную связь признаков с результатом.

Решение

Примерно тот же результат (различия связаны с ошибками округлений) для коэффициента множественной регрессии получим, если воспользуемся формулами

$$R_{yx_1x_2} = \sqrt{1 - \frac{\sigma^2_{ост}}{\sigma^2_y}} = \sqrt{1 - \frac{0{,}6329}{3{,}36}} = 0{,}901$$

$$R_{yx_1x_2} = \sqrt{\sum \beta_i \cdot r_{yx_i}} = \sqrt{0{,}728 \cdot 0{,}87 + 0{,}285 \cdot 0{,}64} = 0{,}903$$

Решение

Скорректированный коэффициент множественной детерминации

$$\widehat{R} = 1 - \left(1 - R^2\right) \cdot \frac{n-1}{n-m-1}$$

$$= 1 - (1 - 0{,}817) \cdot \frac{10-1}{10-2-1} = 0{,}765$$

указывает на умеренную связь между результатом и признаками. Это связано с малым количеством наблюдений.

Решение

Теперь найдем частные коэффициенты корреляции по формулам

$$r_{yx_1 \cdot x_2} = \sqrt{1 - \frac{1 - R^2_{yx_1x_2}}{1 - r^2_{yx_2}}} = \sqrt{1 - \frac{1 - 0{,}817}{1 - 0{,}408}} = 0{,}831$$

$$r_{yx_2 \cdot x_1} = \sqrt{1 - \frac{1 - R^2_{yx_1x_2}}{1 - r^2_{yx_1}}} = \sqrt{1 - \frac{1 - 0{,}817}{1 - 0{,}755}} = 0{,}503$$

Решение

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_2}^2\right) \cdot \left(1 - r_{x_1x_2}^2\right)}} = \frac{0{,}869 - 0{,}639 \cdot 0{,}488}{\sqrt{\left(1 - 0{,}489^2\right)\left(1 - 0{,}639^2\right)}}$$

$$= 0{,}830$$

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_1}^2\right) \cdot \left(1 - r_{x_1x_2}^2\right)}} = \frac{0{,}639 - 0{,}869 \cdot 0{,}488}{\sqrt{\left(1 - 0{,}488^2\right)\left(1 - 0{,}869^2\right)}}$$

$$= 0{,}498$$

Т.е. можно сделать вывод, что фактор x_1, оказывает более сильное влияние на результат, чем признак x_2.

Решение

Оценим надежность уравнения регрессии в целом и показателя связи с помощью F-критерия Фишера. Фактическое значение F-критерия

$$F_{факт} = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m} = \frac{0{,}817}{1 - 0{,}817} \cdot \frac{10 - 2 - 1}{2} = 15{,}63$$

Табличное значение F-критерия при пятипроцентном уровне значимости ($\alpha = 0{,}05$, $k_1 = 2$, $k_2 = 10 - 2 - 1 = 7$) $F_{табл} = 4{,}74$. Так как $F_{факт} = 15{,}63 > F_{табл} = 4{,}10$, то уравнение признается статистически значимым.

Решение

Оценим целесообразность включения фактора x_1 после фактора x_2 и x_2 после фактора x_1 с помощью частного F-критерия Фишера

$$F_{x_1} = \frac{R^2_{yx_1x_2} - r^2_{yx_2}}{1 - R^2_{yx_1x_2}} \cdot (n-3) = \frac{0{,}817 - 0{,}408}{1 - 0{,}817} \cdot 7 = 15{,}65$$

$$F_{x_2} = \frac{R^2_{yx_1x_2} - r^2_{yx_1}}{1 - R^2_{yx_1x_2}} \cdot (n-3) = \frac{0{,}817 - 0{,}755}{1 - 0{,}817} \cdot 7 = 2{,}37$$

Решение

Табличное значение частного F -критерия при пятипроцентном уровне значимости ($\alpha = 0{,}05$, $k_1 = 1,\ k_2 = 10 - 2 - 1 = 7)$: $F_{табл} = 5{,}59$. Так как $F_{x_1} = 15{,}65 > F_{табл} = 5{,}59$, а $F_{x_2} = 2{,}37 < F_{табл} = 5{,}59$, то включение фактора x_1 в модель статистически оправдано и коэффициент чистой регрессии b_1 статистически значим, а дополнительное включение фактора x_2, после того, как уже введен фактор x_1, нецелесообразно.

Уравнение регрессии, включающее только один значимый аргумент

$$\hat{y} = -2{,}754 + 1{,}016x_1$$

Линейные регрессионные модели с гетероскедастичными остатками

При оценке параметров уравнения регрессии применяется метод наименьших квадратов (МНК). При этом делаются определенные предпосылки относительно случайной составляющей ε. В модели

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

случайная составляющая представляет собой ненаблюдаемую величину.

После того как произведена оценка параметров модели, рассчитывая разности фактических и теоретических значений результативного признака y , , можно определить оценки случайной составляющей $y - \hat{y}_x$. Поскольку они не являются реальными случайными остатками, их можно считать некоторой выборочной реализацией неизвестного остатка заданного уравнения, т.е. ε_i.

При использовании критериев Фишера и Стьюдента делаются предположения относительно поведения остатков ε_i - – остатки представляют собой независимые случайные величины и их среднее значение равно 0; они имеют одинаковую (постоянную) дисперсию и подчиняются нормальному распределению.

После построения уравнения регрессии проводится проверка наличия у оценок ε_i (случайных остатков) тех свойств, которые предполагались. Связано это с тем, что оценки параметров регрессии должны отвечать определенным критериям. Они должны быть несмещенными, состоятельными и эффективными. Эти свойства оценок, полученных по МНК, имеют чрезвычайно важное практическое значение в использовании результатов регрессии и корреляции.

Несмещенность

Несмещенность оценки означает, что математическое ожидание остатков равно нулю. Если оценки обладают свойством несмещенности, то их можно сравнивать по разным исследованиям.

Эффективность

Оценки считаются *эффективными*, если они характеризуются наименьшей дисперсией. В практических исследованиях это означает возможность перехода от точечного оценивания к интервальному.

Состоятельность

Состоятельность оценок характеризует увеличение их точности с увеличением объёма выборки. Большой практический интерес представляют те результаты регрессии, для которых доверительный интервал ожидаемого значения параметра регрессии b_i имеет предел значений вероятности, равный единице. Иными словами, вероятность получения оценки на заданном расстоянии от истинного значения параметра близка к единице.

Исследования остатков

Исследования остатков ε_i предполагают проверку наличия следующих пяти предпосылок МНК:

1) случайный характер остатков;

2) нулевая средняя величина остатков, не зависящая от x_i,

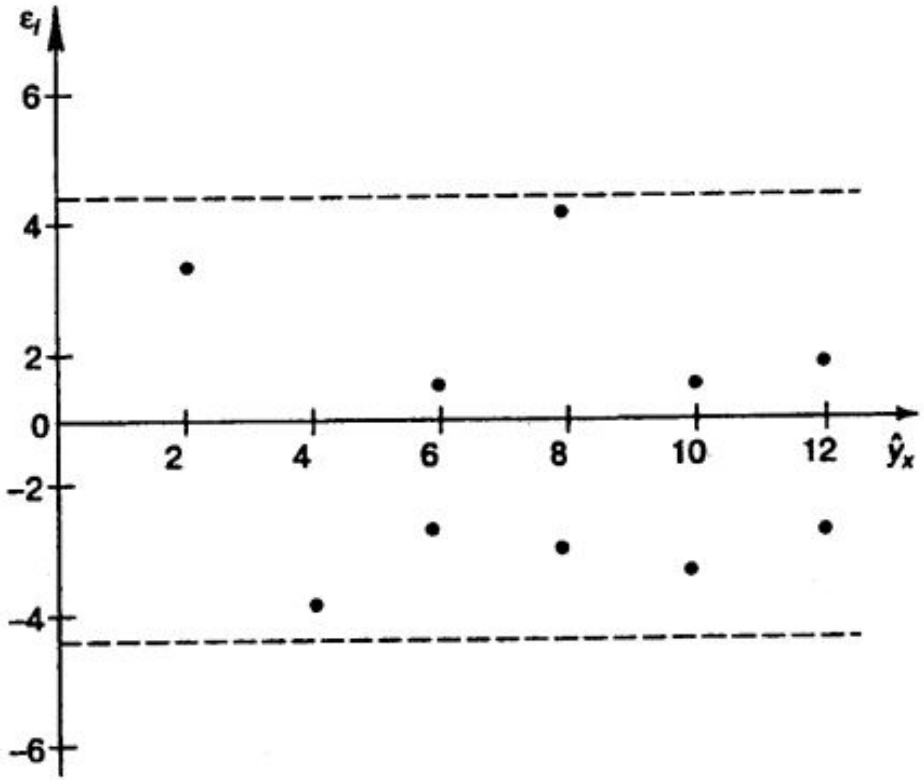
3) гомоскедастичность – дисперсия каждого отклонения ε_i, одинакова для всех значений x;

4) отсутствие автокорреляции остатков – значения остатков ε_i распределены независимо друг от друга;

5) остатки подчиняются нормальному распределению.

Исследования остатков

Прежде всего, проверяется случайный характер остатков ε_i . С этой целью стоится график зависимости остатков ε_i от теоретических значений результативного признака.



Исследования остатков

Если на графике получена горизонтальная полоса, то остатки ε_i представляют собой случайные величины и МНК оправдан, теоретические значения $\hat{y}_x$ хорошо аппроксимируют фактические значения y.

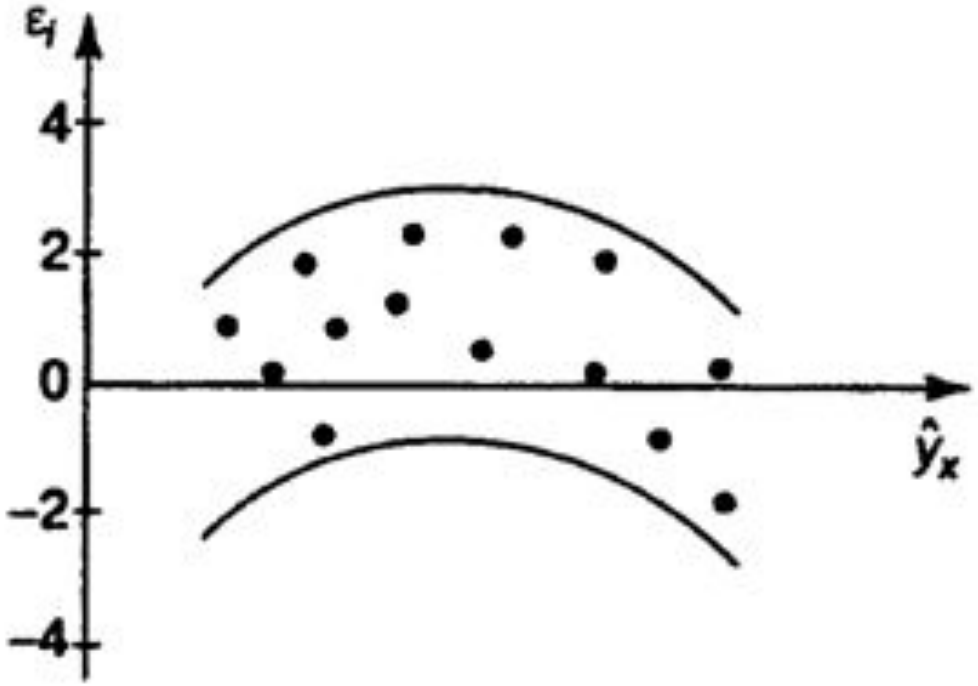
Исследования остатков

Если на графике получена горизонтальная полоса, то остатки ε_i представляют собой случайные величины и МНК оправдан, теоретические значения $\hat{y}_x$ хорошо аппроксимируют фактические значения y.

Исследования остатков

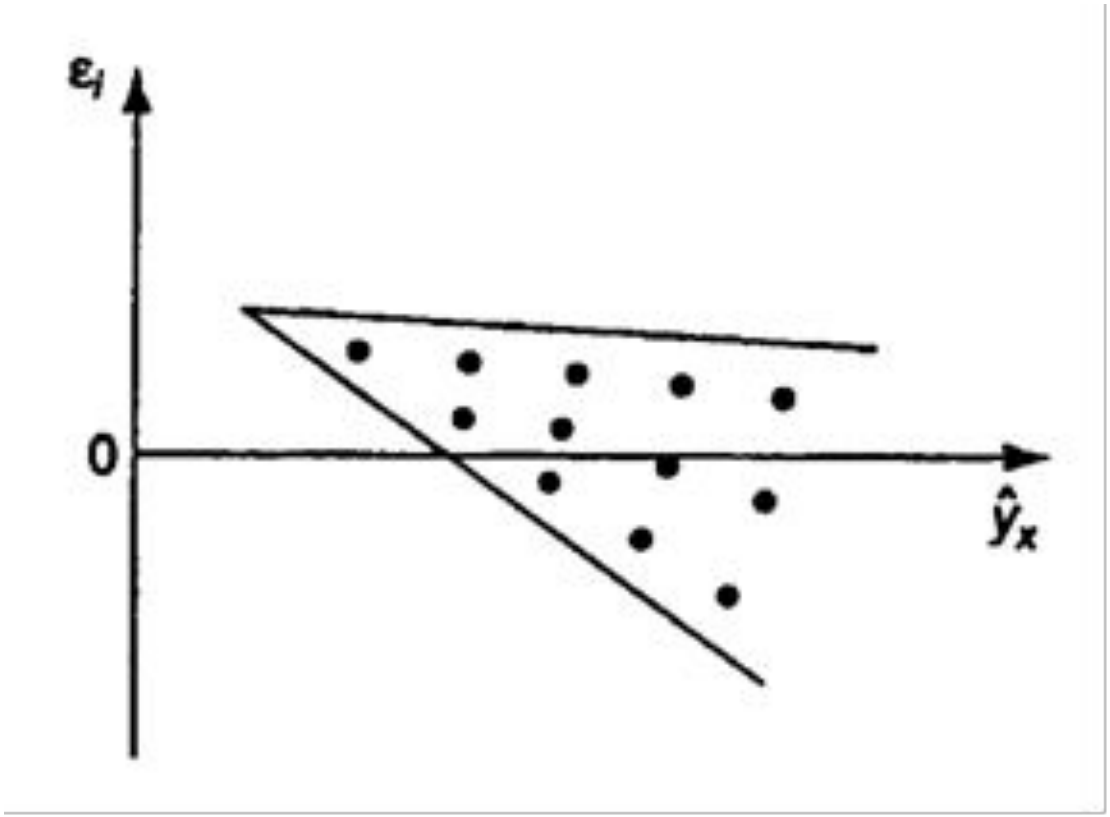
Возможны следующие случаи, если ε_i зависит от $\hat{y}_x$ то:

1) остатки ε_i не случайны;



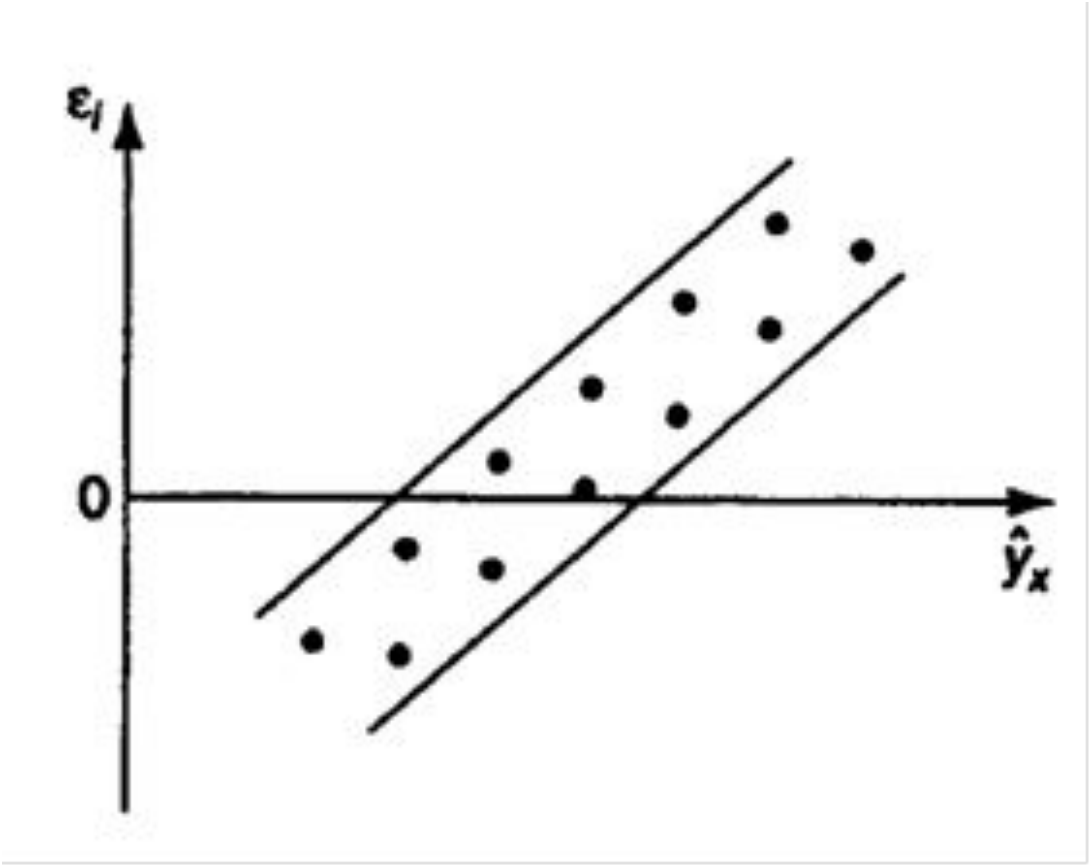
Исследования остатков

2) остатки ε_i не имеют постоянной дисперсии;



Исследования остатков

3) остатки ε_i носят систематический характер.



Исследования остатков

В этих случаях необходимо либо применять другую функцию, либо вводить дополнительную информацию и заново строить уравнение регрессии до тех пор, пока остатки не будут случайными величинами.

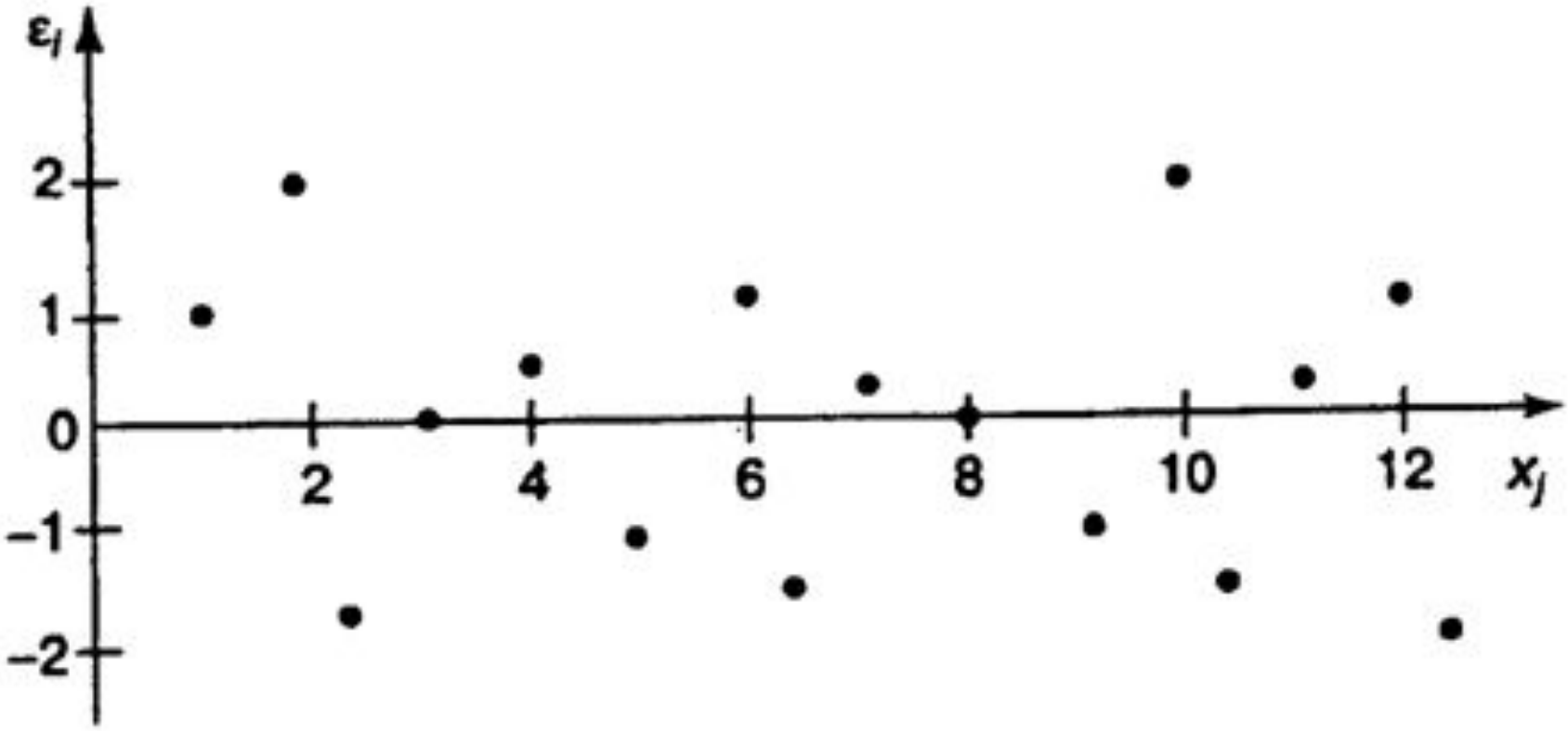
Исследования остатков

Вторая предпосылка МНК относительно нулевой средней величины остатков означает, что $\sum(y - \hat{y}_x) = 0$. Это выполнимо для линейных моделей и моделей, нелинейных относительно включаемых переменных.

Исследования остатков

Вместе с тем, несмещенность оценок коэффициентов регрессии, полученных МНК, зависит от независимости случайных остатков и величин x, что также исследуется в рамках соблюдения второй предпосылки МНК. С этой целью наряду с изложенным графиком зависимости остатков ε_i от теоретических значений результативного признака $\hat{y}_x$ строится график зависимости случайных остатков ε_i от факторов, включенных в регрессию x_j.

Исследования остатков



Исследования остатков

Если остатки на графике расположены в виде горизонтальной полосы, то они независимы от значений x_j. Если же график показывает наличие зависимости ε_i и x_j, то модель неадекватна.

Исследования остатков

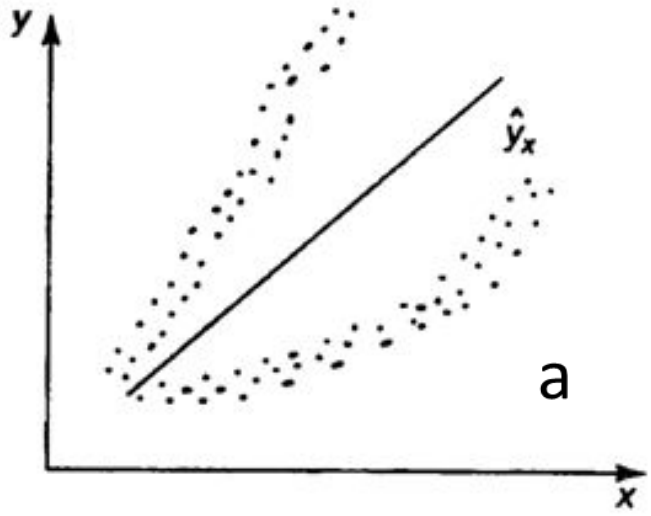
Предпосылка о нормальном распределении остатков позволяет проводить проверку параметров регрессии и корреляции с помощью F- и t-критериев. Вместе с тем, оценки регрессии, найденные с применением МНК, обладают хорошими свойствами даже при отсутствии нормального распределения остатков, т.е. при нарушении пятой предпосылки МНК.

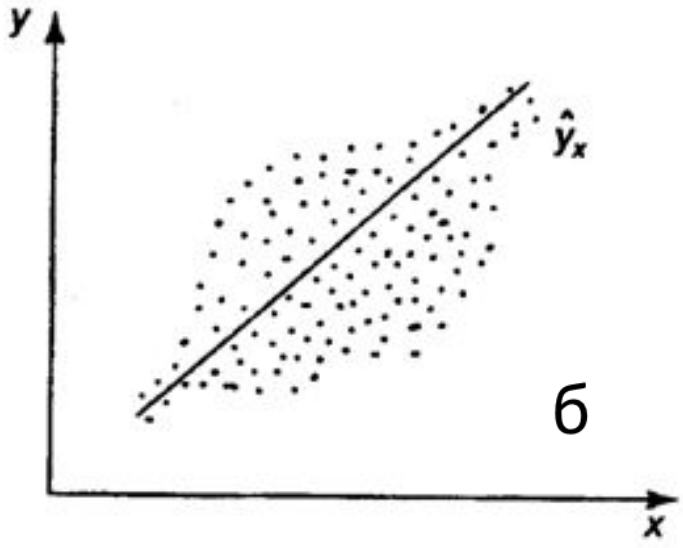
Исследования остатков

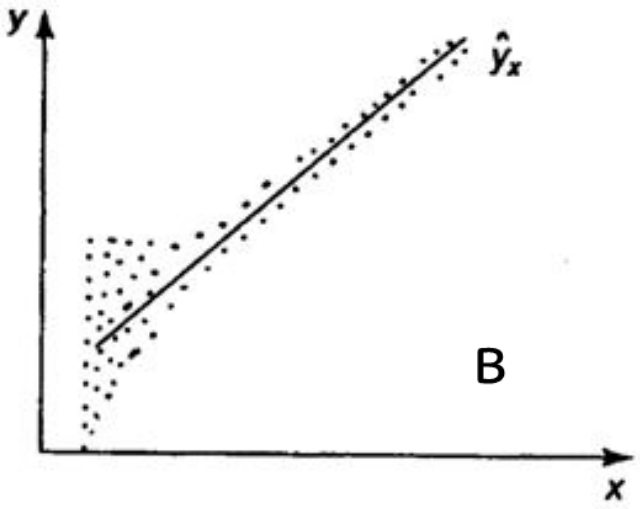
Совершенно необходимым для получения по МНК состоятельных оценок параметров регрессии является соблюдение третьей и четвертой предпосылок.

В соответствии с третьей предпосылкой МНК требуется, чтобы дисперсия остатков была *гомоскедастичной*. Это значит, что для каждого значения фактора x_j и остатки ε_i имеют одинаковую дисперсию. Если это условие применения МНК не соблюдается, то имеет место *гетероскедастичность*. Наличие гетероскедастичности можно наглядно видеть из поля корреляции.

Исследования остатков. Примеры гетероскедастичности







Исследования остатков. Примеры гетероскедастичности

На рис. изображено:

а – дисперсия остатков растет по мере увеличения x;

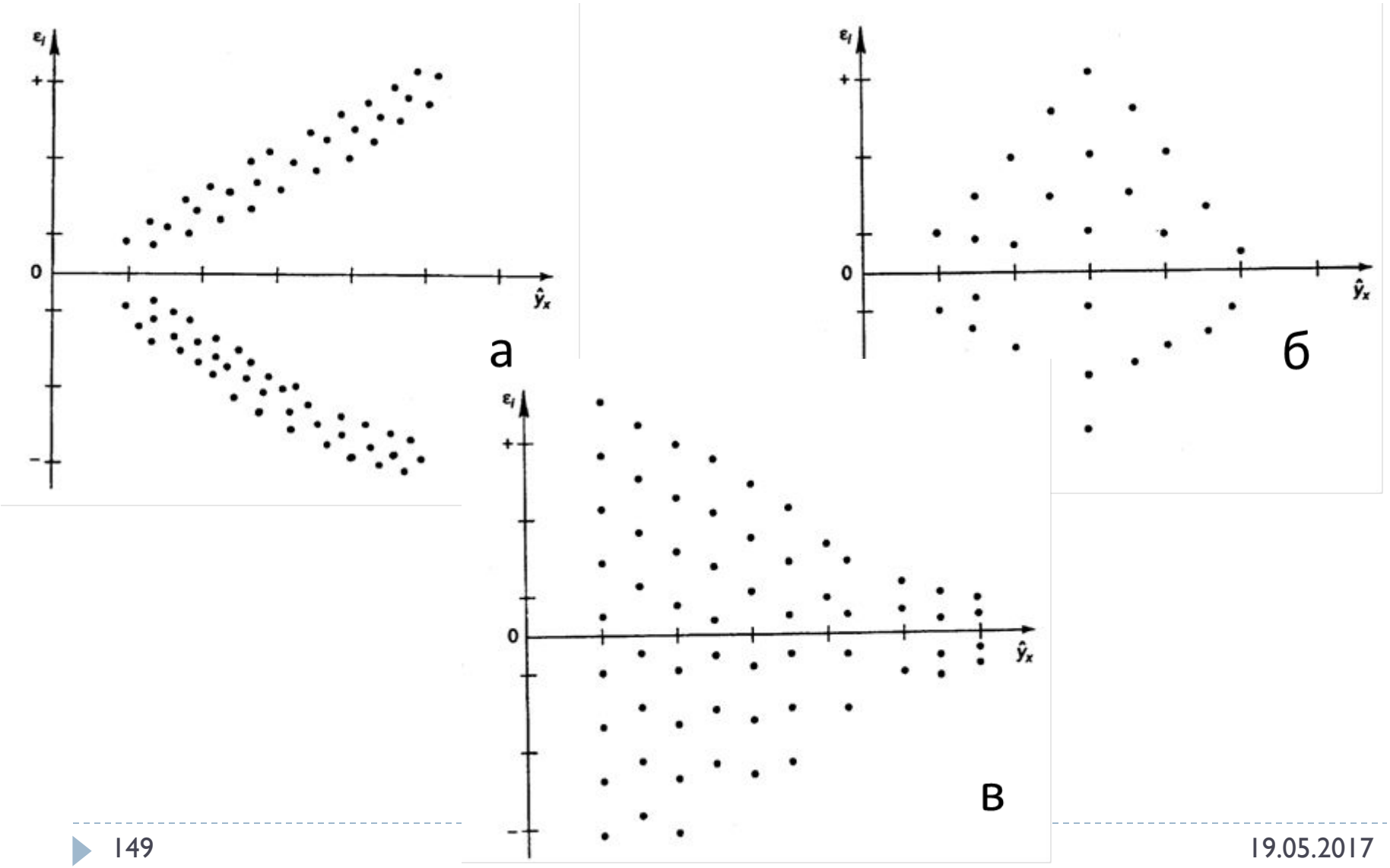
б – дисперсия остатков достигает максимальной величины при средних значениях переменной и уменьшается при минимальных и максимальных значениях x;

в – максимальная дисперсия остатков при малых значениях и дисперсия остатков однородна по мере увеличения значений x.

Исследования остатков. Примеры гетероскедастичности

Наличие гомоскедастичности или гетероскедастичности можно видеть и по рассмотренному выше графику зависимости остатков ε_i от теоретических значений результативного признака $\hat{y}_x$

Исследования остатков. Примеры гетероскедастичности



Исследования остатков.

Для множественной регрессии данный вид графиков является наиболее приемлемым визуальным способом изучения гомо- и гетероскедастичности.

Исследования остатков. Автокорреляция.

При построении регрессионных моделей чрезвычайно важно соблюдение четвертой предпосылки МНК – отсутствие автокорреляции остатков, т.е. значения остатков ε_i, распределены независимо друг от друга. Автокорреляция остатков означает наличие корреляции между остатками текущих и предыдущих (последующих) наблюдений

Исследования остатков. Автокорреляция.

Коэффициент корреляции между ε_i и ε_j , где ε_i – остатки текущих наблюдений, ε_j – остатки предыдущих наблюдений (например, $j = i - 1$)), может быть определен как

$$r_{\varepsilon_i \varepsilon_j} = \frac{\mathrm{cov}(\varepsilon_i,\ \varepsilon_j)}{\sigma_{\varepsilon_i} \cdot \sigma_{\varepsilon_j}}$$

т.е. по обычной формуле линейного коэффициента корреляции. Если этот коэффициент окажется существенно отличным от нуля, то остатки автокоррелированы.

Исследования остатков. Автокорреляция.

При несоблюдении основных предпосылок МНК приходится корректировать модель, изменяя ее спецификацию, добавлять (исключать) некоторые факторы, преобразовывать исходные данные для того, чтобы получить оценки коэффициентов регрессии, которые обладают свойством несмещенности, имеют меньшее значение дисперсии остатков и обеспечивают в связи с этим более эффективную статистическую проверку значимости параметров регрессии.

Обобщенный метод наименьших квадратов (ОМНК)

самостоятельно. См. пособие стр. 73

Регрессионные модели с переменной структурой (фиктивные переменные) самостоятельно. См. пособие стр. 80

Спасибо за внимание!