Парная линейная регрессия

С тех пор как экономика стала серьезной самостоятельной наукой, исследователи пытаются дать свое представление о возможных путях экономического развития, спрогнозировать ту или иную ситуацию, предвидеть будущие значения экономических показателей, указать инструменты изменения ситуации в желательном направлении.

Поведение и значение любого экономического показателя зависят практически от бесконечного количества факторов. Учесть все факторы нереально. Обычно лишь ограниченное количество факторов действительно существенно воздействуют на исследуемый экономический показатель. Доля влияния остальных факторов столь незначительна, что их игнорирование не может привести к существенным отклонениям в предполагаемом поведении исследуемого объекта. Выделение и учет ограниченного числа реально доминирующих факторов даёт основание для качественного анализа, прогнозирования и управления ситуацией.

Любая экономическая политика заключается в регулировании экономических переменных, и она должна базироваться на знании того, как эти переменные связаны с другими переменными. В рыночной экономике нельзя непосредственно регулировать темп инфляции, но на него можно воздействовать средствами фискальной (бюджетно-налоговой) и монетарной (кредитно-денежной) политики. А значит, должна быть изучена зависимость между предложением денег и уровнем цен.

Инструментарием такого анализа являются методы статистики и эконометрики, в частности регрессионного и корреляционного анализа. Следует иметь в виду, что статистический анализ зависимостей сам по себе не вскрывает существо причинных связей между явлениями, т.е. он не решает вопроса, в силу каких причин одна переменная влияет на другую. Решение такой задачи является результатом качественного (содержательного) изучения связей, которое обязательно должно либо предшествовать статистическому анализу, либо сопровождать его.

В экономике чаще имеют не функциональные, а *корреляционные,* либо *статистические,* зависимости. Нахождение, оценка и анализ таких зависимостей, построение формул зависимостей и оценка её параметров являются одной из важнейших задач эконометрики.

Статистической называют зависимость, при которой изменение одной из величин влечёт изменение распределения другой. В частности, статистическая зависимость проявляется в том, что при изменении одной из величин изменяется среднее значение другой. Такую статистическую зависимость называют *корреляционной.*

Можно указать два варианта рассмотрения взаимосвязей между двумя переменными X и Y. В первом случае обе переменные считаются равноценными в том смысле, что они не подразделяются на первичную и вторичную (независимую и зависимую) переменные. Основным в этом случае является вопрос о наличии и силе взаимосвязи между этими переменными. Например, между ценой товара и объемом спроса на него, между урожаем картофеля и урожаем зерна, между интенсивностью движения транспорта и числом аварий. При исследовании силы линейной зависимости между такими переменными обращаются к корреляционному анализу, основной мерой которого является коэффициент корреляции.

Другой вариант рассмотрения взаимосвязей выделяет одну из величин как независимую (объясняющую), а другую как зависимую (объясняемую). В этом случае изменение первой из них может служить причиной для изменения другой. Например, рост дохода ведет к увеличению потребления; рост цены — к снижению спроса; снижение процентной ставки увеличивает инвестиции; увеличение обменного курса валюты сокращает объем чистого экспорта и т.д. Однако такая зависимость не является однозначной в том смысле, что каждому конкретному значению объясняющей переменной (набору объясняющих переменных) может соответствовать не одно, а множество значений из некоторой области. Или, каждому конкретному значению объясняющей переменной (набору объясняющих переменных) соответствует некоторое вероятностное распределение зависимой переменной (рассматриваемой как СВ).

Тогда анализируется, как объясняющая(ие) переменная(ые) влияет(ют) на зависимую переменную «в среднем». Зависимость такого типа, выражаемая соотношением

$$M\left(Y\middle|x\right)=f(x)$$

и называется *функцией регрессии* Y на X. При этом X называется *независимой (объясняющей) переменной (регрессором),* Y — *зависимой (объясняемой) переменной.* При рассмотрении зависимости двух СВ говорят о *парной регрессии.*

Зависимость нескольких переменных, выражаемая функцией

$$M\left(Y\middle|x_1,x_2,\dots,x_m\right)=f(x_1,x_2,\dots,x_m)$$

называют *множественной регрессией.*

Под *регрессией* понимается функциональная зависимость между объясняющими переменными и условным математическим ожиданием (средним значением) зависимой переменной, которая строится с целью предсказания (прогнозирования) этого среднего значения при фиксированных значениях объясняющих переменных.

Так как реальные значения зависимой переменной не всегда совпадают с её условными математическими ожиданиями, и могут быть различными при одном и том же значении объясняющей переменной (наборе объясняющих переменных), означает, что фактическая зависимость должна быть дополнена некоторым слагаемым ε, которое, в сущности, является СВ и указывает на стохастическую суть зависимости. Значит связи между зависимой и объясняющей(ими) переменными должны выражаться соотношениями

$$M\left(Y\middle|x\right) = f(x) + \varepsilon$$

$$M\left(Y\middle|x_1, x_2, \ldots, x_m\right) = f(x_1, x_2, \ldots, x_m) + \varepsilon$$

называемыми *регрессионными моделями (уравнениями)*.

Определим причины предопределяющие присутствия в регрессионных моделях случайного фактора (отклонения). Среди таких причин выделим наиболее существенные.

1. *Невключение в модель всех объясняющих переменных.*

Любая регрессионная (в частности, эконометрическая) модель является упрощением реальной ситуации. Последняя всегда представляет собой сложнейшее переплетение различных факторов, многие из которых в модели не учитываются, что порождает отклонение реальных значений зависимой переменной от ее модельных значений. Безусловно, перечислить все объясняющие переменные здесь практически невозможно. Проблема еще и в том, что никогда заранее не известно, какие факторы при создавшихся условиях действительно являются определяющими, а какими можно пренебречь. В ряде случаев учесть непосредственно какой-то фактор нельзя в силу невозможности получения по нему статистических данных. Кроме того, ряд факторов носит принципиально случайный характер (например, погода), что добавляет неоднозначности при рассмотрении некоторых моделей (например, модели сельскохозяйственной деятельности, прогнозирующие объём урожая).

2. *Неправильный выбор функциональной формы модели.*
Из-за слабой изученности исследуемого процесса либо из-за его переменчивости может быть неверно подобрана функция, его моделирующая. Это, безусловно, скажется на отклонении модели от реальности, что отразится на величине случайного члена.
3. *Агрегирование переменных.*
Во многих моделях рассматриваются зависимости между факторами, которые сами представляют сложную комбинацию других, более простых переменных. Например, при рассмотрении в качестве зависимой переменной совокупного спроса проводится анализ зависимости, в которой объясняемая переменная является сложной композицией индивидуальных спросов, оказывающих на нее определённое влияние помимо факторов, учитываемых в модели. Это может оказаться причиной отклонения реальных значений от модельных.

4. *Ошибки измерений.*

Какой бы качественной ни была модель, ошибки измерений переменных отразятся на несоответствии модельных значений эмпирическим данным, что также отразится на величине случайного члена.

5. *Ограниченность статистических данных.*

Зачастую строятся модели, выражаемые непрерывными функциями. Но для этого используется набор данных, имеющих дискретную структуру. Это несоответствие находит свое выражение в случайном отклонении.

6. *Непредсказуемость человеческого фактора.*

Эта причина может «испортить» самую качественную модель. Действительно, при правильном выборе формы модели, скрупулезном подборе объясняющих переменных все равно невозможно спрогнозировать поведение каждого индивидуума.

Решение эконометрической задачи построения качественного уравнения регрессии, соответствующего эмпирическим данным и целям исследования, является достаточно сложным и многоступенчатым процессом. Его можно разбить на три этапа:

1. выбор формулы уравнения регрессии;
2. определение параметров выбранного уравнения;
3. анализ качества уравнения и поверка адекватности уравнения эмпирическим данным, совершенствование уравнения.

Выбор формулы связи переменных называется *спецификацией* уравнения регрессии. Задача определения параметров принятого при спецификации уравнения называется *параметризации* (*идентификацией*), проверка качества уравнения регрессии, её соответствия (репрезентативности) реальности называется *верификацией.*

Если функция регрессии линейна, то говорят о *линейной регрессии.* Модель линейной регрессии (линейное уравнение) является наиболее распространенным (и простым) видом зависимости между экономическими переменными. Кроме того, построенное линейное уравнение может служить начальной точкой эконометрического анализа.

Линейная регрессия (теоретическое линейное уравнение регрессии) представляет собой линейную функцию между условным математическим ожиданием $M\left(Y\middle|X=x_i\right)$ зависимой переменной Y и одной объясняющей переменной X (x_i — значения независимой переменной в i-ом наблюдении, где $i=\overline{1,n}$),

$$M\left(Y\middle|X=x_i\right)=\alpha_0+\alpha_1 x_i$$

То, что каждое индивидуальное значение y_i отличается от соответствующего условного математического ожидания, в силу сказанного выше, необходимо ввести случайное слагаемое ε_i

$$M\left(Y\middle|X=x_i\right)=\alpha_0+\alpha_1 x_i+\varepsilon_i$$

Полученное соотношение называется *теоретической линейной регрессионной моделью;* α_0 и α_1 — *теоретическими параметрами (теоретическими коэффициентами) регрессии*; ε_i — *случайным отклонением.*

Следовательно, индивидуальные значения y_i представляются в виде суммы двух компонент — систематической $(\alpha_0 + \alpha_1 x_i)$ и случайной ε_i.
Теоретическую линейную регрессионную модель будем представлять в виде

$$Y = \alpha_0 + \alpha_1 X + \varepsilon$$

Для определения значений теоретических коэффициентов регрессии необходимо знать и использовать все значения переменных X и Y генеральной совокупности, что практически невозможно.
Таким образом, задача линейного регрессионного анализа состоят в том, чтобы по имеющимся статистическим данным $\{x_i, y_i\}_{i=1,n}$ для переменных X и Y:

1. получить наилучшие оценки неизвестных параметров α_0 и α_1;
2. проверить статистические гипотезы о параметрах модели;
3. проверить, достаточно ли хорошо принятая модель согласуется со статистическими данными (адекватность модели данным наблюдений).

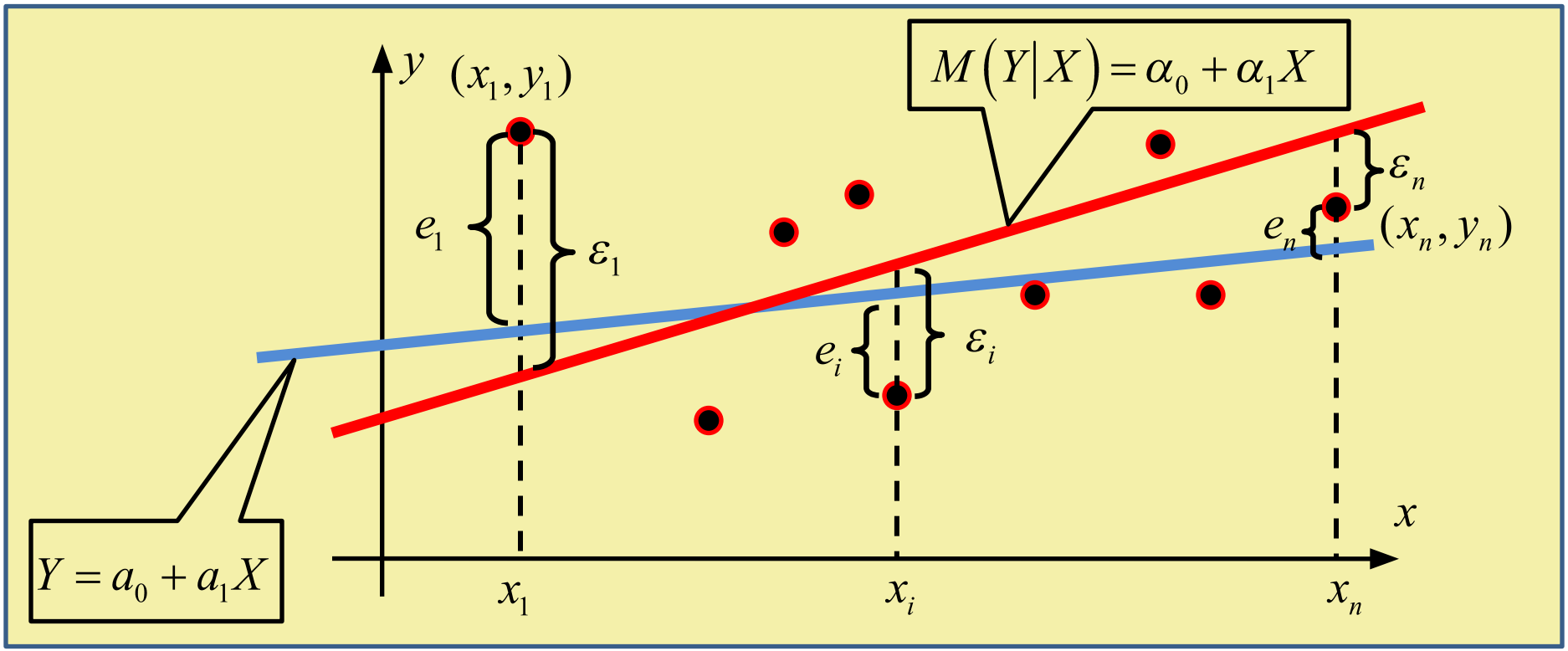
Значит, по выборке ограниченного объема можно построить лишь *эмпирическое уравнение регрессии*

$$\breve{y}_i = a_0 + a_1 x_i$$

где $\breve{y}_i$ — оценка условного математического ожидания $M\left(Y \middle| X = x_i\right)$;
a_0 и a_1 — оценки неизвестных параметров α_0 и α_1 называемые *эмпирическими коэффициентами регрессии.* Очевидно, для *i*-ого наблюдения

$$y_i = a_0 + a_1 x_i + e_i$$

где *отклонение* e_i — оценка теоретического случайного отклонения ε_i.



В силу несовпадения статистической базы для генеральной совокупности и выборки оценки a_0 и a_1 практически всегда отличаются от истинных значений коэффициентов α_0 и α_1, что влечёт несовпадение эмпирической и теоретической линий регрессии.

Задача состоит в том, чтобы по конкретной выборке $\{x_i, y_i\}_{i=\overline{1,n}}$, найти оценки a_0 и a_1 неизвестных параметров α_0 и α_1 так, чтобы построенная линия регрессии являлась бы наилучшей в определенном смысле среди всех других прямых. То есть, построенная прямая должна быть «ближайшей» к точкам наблюдений по их совокупности. Очевидно, мерами качества найденных оценок могут служить определенные композиции получаемых отклонений $e_i \ \left(i = \overline{1,n}\right)$. В частности, минимум суммы квадратов отклонений

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \tilde{y}_i\right)^2 = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i\right)^2$$

Метод определения оценок коэффициентов из условия минимизации этой суммы называется *методом наименьших квадратов (МНК)*. Этот метод оценки является наиболее простым с вычислительной точки зрения. Кроме того, оценки коэффициентов регрессии, найденные МНК при определенных предпосылках, обладают рядом оптимальных свойств.

$$\left.\begin{array}{l}\{x_i, y_i\}_{i=\overline{1,n}} \\ \tilde{y} = a_0 - a_1 x \\ \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 \ (\min)\end{array}\right\} \Rightarrow \left.\begin{array}{lll} n a_0 & + a_1 \sum x_i & = \sum y_i \\ a_0 \sum x_i & + a_1 \sum x_i^2 & = \sum y_i x_i \end{array}\right\} \Rightarrow$$

$$\left.\begin{array}{l} a_0 = \dfrac{1}{n}\sum y_i - a_1 \dfrac{1}{n}\sum x_i \\ a_1 = \dfrac{n\sum y_i x_i - \sum y_i \sum x_i}{n\sum x_i^2 - \sum {}^2 x_i} \end{array}\right\} \Rightarrow \begin{cases} a_0 = \bar{y} - a_1 \bar{x} \\ a_1 = \dfrac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \bar{x}^2} \end{cases}$$

$$a_1 = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2} = r_{xy} \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$a_1 = \frac{S_{xy}}{S_x^2} = \frac{S_{xy}}{S_x S_y}\frac{S_y}{S_x} = r_{xy}\frac{S_y}{S_x}$$

где r_{xy} выборочный коэффициент корреляции
S_x, S_y стандартные отклонения

Для полученных оценок МНК справедливо

1. Оценки МНК являются функциями от выборки.
2. Оценки МНК являются точечными оценками теоретических коэффициентов регрессии.
3. Согласно эмпирическая прямая регрессии обязательно проходит через точку $(\overline{x}, \overline{y})$.
4. Эмпирическое уравнение регрессии построено таким образом, что сумма отклонений $\sum e_i$, а также среднее значение отклонений $\overline{e}$ равны нулю.

$$\sum e_i = \overline{e} = 0$$

5. Случайные отклонения не коррелированы с наблюдаемыми значениями зависимой переменной *Y*.
6. Случайные отклонения e_i не коррелированы с наблюдаемыми значениями x_i независимой переменной *X*.

Регрессионный анализ позволяет определить оценки коэффициентов регрессии. Являясь лишь оценками, они не позволяют сделать вывод, насколько точно эмпирическое уравнение регрессии соответствует уравнению для всей генеральной совокупности, насколько близки оценки a_0 и a_1 коэффициентов к своим теоретическим прототипам α_0 и α_1, как близко оцененное значение $\tilde{y}_i$ к условному математическому ожиданию $M\left(Y \middle| X = x_i\right)$ насколько надежны найденные оценки.

Значения $\tilde{y}_i$ зависят от значений x_i и случайных отклонений ε_i. Следовательно, переменная Y является СВ, напрямую связанной с ε_i.

Истинная зависимость

$$Y = \alpha_0 + \alpha_1 X + \varepsilon \qquad \left(y_i = \alpha_0 + \alpha_1 x_i + \varepsilon_i\right)$$

$$\tilde{Y} = a_0 + a_1 X \qquad \left(\tilde{y}_i = a_0 + a_1 x_i\right)$$

Предполагаемая (линейная) зависимость

$$a_1 = \frac{S_{xy}}{S_x^2}, \quad S_{xy} = \operatorname{cov}(X,Y) = \operatorname{cov}(X, \alpha_0 + \alpha_1 X + \varepsilon)$$

Ковариация от постоянной величины равна нулю

$$= \operatorname{cov}(X,\alpha_0) + \operatorname{cov}(X,\alpha_1 X) + \operatorname{cov}(X,\varepsilon) =$$

$$= \alpha_1 S_x^2 + \operatorname{cov}(X,\varepsilon)$$

$$a_1 = \alpha_1 + \frac{S_{x\varepsilon}}{S_x^2}$$

Случайная величина

Предпосылки МНК (условия Гаусса—Маркова)

1. *Математическое ожидание случайного отклонения* ε_i *равно нулю: для всех наблюдений* $M(\varepsilon_i) = 0$.

Данное условие означает, что случайное отклонение в среднем не оказывает влияния на зависимую переменную. В каждом конкретном наблюдении случайный член может быть либо положительным, либо отрицательным, но он не должен иметь систематического смещения. Очевидно,

$$M(\varepsilon_i) = 0 \Rightarrow M\left(Y \middle| X = x_i\right) = \alpha_0 + \alpha_1 x_i$$

2. *Дисперсия случайных отклонений постоянна:* $D(\varepsilon_i) = D(\varepsilon_j) = \sigma^2$ *для любых наблюдений i и j.*

Данное условие подразумевает, что несмотря на то, что при каждом конкретном наблюдении случайное отклонение может быть либо большим, либо меньшим, но не должно быть некой причины, вызывающей большую ошибку (отклонение).

Выполнимость данной предпосылки называется *гомоскедастичностью (постоянством дисперсии отклонений)*. Невыполнимость данной предпосылки называется *гетероскедастичностью (непостоянством дисперсий отклонений)*.

Поскольку $D(\varepsilon_i) = M\left[\left(\varepsilon_i - M(\varepsilon_i)\right)^2\right] = M\left(\varepsilon_i^2\right)$, то данную предпосылку можно переписать в форме: $M\left(\varepsilon_i^2\right) = \sigma^2$.

3. *Случайные отклонения ε_i и ε_j являются независимыми друг от друга для $i \neq j$.*

Выполнимость данной предпосылки предполагает, что отсутствует систематическая связь между любыми случайными отклонениями.

Очевидно,

$$\sigma_{\varepsilon_i\varepsilon_j} = \operatorname{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} 0, & i \neq j \\ \sigma^2, & i = j \end{cases}$$

Если данное условие выполняется, то говорят об отсутствии *автокорреляции.*

С учетом выполнимости первой предпосылки получаем,

$$\left.\begin{array}{l} \sigma_{\varepsilon_i\varepsilon_j} = \operatorname{cov}(\varepsilon_i, \varepsilon_j) = \begin{cases} 0, & i \neq j \\ \sigma^2, & i = j \end{cases} \\ M(\varepsilon_i) = 0 \end{array}\right\} \Rightarrow M(\varepsilon_i\varepsilon_j) = 0$$

4. *Случайное отклонение должно быть независимо от объясняющих переменных.*

Это условие выполняется если объясняющие переменные не случайные.

Причём получаем

$$\sigma_{\varepsilon_i x_i} = \operatorname{cov}(\varepsilon_i, x_i) = M\left(\left(\varepsilon_i - M(\varepsilon_i)\right)\left(x_i - M(x_i)\right)\right) = M\left(\varepsilon_i\left(x_i - M(x_i)\right)\right) =$$

$$= M(\varepsilon_i x_i) - M(\varepsilon_i)M(x_i) = 0$$

5. Модель регрессии является линейной относительно параметров.

Теорема Гаусса-Маркова. Если предпосылки 1— 5 выполнены, то оценки, полученные по МНК, обладают следующими свойствами:

1. Оценки являются несмещенными, т.е. $M(a_0) = \alpha_0$, $M(a_1) = \alpha_1$. Это вытекает из того, что $M(\varepsilon_i) = 0$, и говорит об отсутствии систематической ошибки в определении положения линии регрессии.
2. Оценки состоятельны, так как дисперсия оценок параметров при возрастании числа наблюдений стремится к нулю:

$$D(a_0) \xrightarrow[n\to\infty]{} 0, \; D(a_1) \xrightarrow[n\to\infty]{} 0$$

Т.е. при увеличении объема выборки надежность оценок увеличивается (a_0 наверняка близко к α_0, a_1 — близко к α_1).
3. Оценки эффективны, т.е. они имеют наименьшую дисперсию по сравнению с любыми другими оценками данных параметров, линейными относительно величин y_i.

В англоязычной литературе такие оценки называются *BLUE (Best Linear Unbiased Estimators) — наилучшие линейные несмещенные оценки.*

Если предпосылки 2 и 3 нарушены, т.е. дисперсия отклонений непостоянна и (или) значения связаны друг с другом, то свойства несмещённости и состоятельности сохраняются, но свойство эффективности — нет.

С выполнимостью указанных предпосылок при построении классических линейных регрессионных моделей делаются ещё предположения, такие например, как:

- объясняющие переменные не являются СВ;
- случайные отклонения имеют нормальное распределение;
- число наблюдений существенно больше числа объясняющих переменных;
- отсутствуют ошибки спецификации;
- отсутствует совершенная мультиколлинеарность.

В силу случайного отбора элементов в выборку случайными являются также оценки a_0 и a_1 коэффициентов α_0 и α_1 теоретического уравнения регрессии. Их математические ожидания при выполнении предпосылок об отклонениях равны соответственно $M(a_0)=\alpha_0,\ M(a_1)=\alpha_1$. При этом оценки тем надежнее, чем меньше их разброс вокруг α_0 и α_1, т.е. чем меньше дисперсии $D(a_0)$ и $D(a_1)$ оценок. Надежность получаемых оценок, связана с дисперсией случайных отклонений ε_i, и является дисперсией переменной Y относительно линии регрессии (дисперсией Y, очищенной от влияния X).

$$a_1=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}=\frac{\sum(x_i-\bar{x})y_i}{\sum(x_i-\bar{x})^2}-\frac{\sum(x_i-\bar{x})\bar{y}}{\sum(x_i-\bar{x})^2}=\frac{\sum(x_i-\bar{x})y_i}{\sum(x_i-\bar{x})^2}$$

$$=\sum\frac{x_i-\bar{x}}{\sum(x_i-\bar{x})^2}y_i=\sum c_i y_i \qquad \left(c_i=\frac{x_i-\bar{x}}{\sum(x_i-\bar{x})^2}\right)$$

обозначим

$$a_0=\bar{y}-a_1\bar{x}=\frac{1}{n}\sum y_i-\bar{x}\sum c_i y_i=\sum\left(\frac{1}{n}-\bar{x}c_i\right)y_i=\sum d_i y_i, \qquad \left(d_i=\frac{1}{n}-\bar{x}c_i\right)$$

обозначим

$$D(a_1) = D\left(\sum c_i y_i\right) = \sigma^2 \sum c_i^2 = \sigma^2 \frac{1}{\sum \left(x_i - \overline{x}\right)^2}$$

$$D(a_0) = D\left(\sum d_i y_i\right) = \sigma^2 \sum d_i^2 = \sigma^2 \sum \left(\frac{1}{n} - c_i \overline{x}\right)^2 = \sigma^2 \sum \left(\frac{1}{n^2} - 2c_i \frac{1}{n} \overline{x} + c_i^2 \overline{x}^2\right) =$$

$$= \sigma^2 \left(n \frac{1}{n^2} - 2 \frac{1}{n} \overline{x} \sum c_i + \overline{x}^2 \sum c_i^2\right) =$$

$$= \sigma^2 \left(\frac{1}{n} - 2 \frac{1}{n} \overline{x} \sum \frac{x_i - \overline{x}}{\sum \left(x_i - \overline{x}\right)^2} + \overline{x}^2 \sum \left(\frac{x_i - \overline{x}}{\sum \left(x_i - \overline{x}\right)^2}\right)^2\right) =$$

$$= \sigma^2 \left(\frac{1}{n} - 2 \frac{1}{n} \overline{x} \frac{\sum (x_i - \overline{x})}{\sum \left(x_i - \overline{x}\right)^2} + \overline{x}^2 \frac{\sum (x_i - \overline{x})^2}{\left(\sum \left(x_i - \overline{x}\right)^2\right)^2}\right) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{\sum \left(x_i - \overline{x}\right)^2}\right) =$$

$$= \sigma^2 \frac{\sum \left(x_i - \overline{x}\right)^2 + n\overline{x}^2}{n \sum \left(x_i - \overline{x}\right)^2} = \sigma^2 \frac{\sum x_i^2 + n\overline{x}^2 - 2\overline{x} \sum x_i + n\overline{x}^2}{n \sum \left(x_i - \overline{x}\right)^2} = \sigma^2 \frac{\sum x_i^2}{n \sum \left(x_i - \overline{x}\right)^2}$$

$$D(a_1) = \sigma^2 \frac{1}{\sum (x_i - \bar{x})^2}; \quad D(a_0) = \sigma^2 \frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2}$$

Из полученных соотношений следует:

• Дисперсии параметров a_0 и a_1 прямо пропорциональны дисперсии отклонения σ^2. Следовательно, чем больше фактор случайности, тем менее точными будут оценки.

• Чем больше число наблюдений, тем меньше дисперсии оценок. Т.е. чем большим числом данных мы располагаем, тем вероятнее получение более точных оценок.

• Чем больше дисперсия (разброс значений $\sum (x_i - \bar{x})^2$) объясняющей переменной, тем меньше дисперсия оценок коэффициентов. Т.е, чем шире область изменений объясняющей переменной, тем точнее будут оценки (тем меньше доля случайности в их определении).

Случайные отклонения ε_i по выборке определены быть не могут, при анализе надежности оценок коэффициентов регрессии они заменяются отклонениями $e_i = y_i - a_0 - a_1 x_i$ значений y_i переменной Y от оцененной линии регрессии $\tilde{y}_i$. Дисперсия случайных отклонений $D(\varepsilon_i) = \sigma^2$ заменяется её несмещенной оценкой S^2

$$\left.\begin{aligned} D(\varepsilon_i) = \sigma^2 &= \frac{1}{n}\sum \varepsilon_i^2 = \frac{1}{n}\sum \left(y_i - \alpha_0 - \alpha_1 x_i\right)^2 \\ S^2 &= \frac{1}{n-2}\sum e_i^2 = \frac{1}{n-2}\sum \left(y_i - a_0 - a_1 x_i\right)^2 \end{aligned}\right\} \Rightarrow D(\varepsilon_i) \approx S^2$$

$$D(a_1) = \sigma^2 \frac{1}{\sum \left(x_i - \overline{x}\right)^2};$$

Тогда

$$D(a_1) \approx S_{a_1}^2 = \frac{S^2}{\sum \left(x_i - \overline{x}\right)^2};$$

$$D(a_0) = \sigma^2 \frac{\sum x_i^2}{n\sum \left(x_i - \overline{x}\right)^2}$$

$$D(a_0) \approx S_{a_0}^2 = \frac{S^2 \sum x_i^2}{n\sum \left(x_i - \overline{x}\right)^2} = \overline{x^2} S_{a_1}^2$$

$S^2 = \dfrac{\sum e_i^2}{n-2}$ – необъясненная дисперсия (мера разброса зависимой переменной вокруг линии регрессии), корень квадратный из необъясненной дисперсии, т.е. $S = \sqrt{\dfrac{\sum e_i^2}{n-2}}$, называется *стандартной ошибкой оценки (стандартной ошибкой регрессии).*

$S_{a_1} = \sqrt{\dfrac{S^2}{\sum (x_i - \overline{x})^2}}$ и $S_{a_0} = \sqrt{\dfrac{S^2 \sum x_i^2}{n \sum (x_i - \overline{x})^2}}$ — стандартные отклонения случайных величин a_1 и a_0 называемые *стандартными ошибками коэффициентов регрессии.*

Эмпирическое уравнение регрессии определяется на основе конечного числа статистических данных. Поэтому коэффициенты эмпирического уравнения регрессии являются СВ, изменяющимися от выборки к выборке. При проведении статистического анализа возникает необходимость сравнения эмпирических коэффициентов регрессии a_0 и a_1 с некоторыми теоретически ожидаемыми значениями α_0 и α_1 этих коэффициентов. Данный анализ осуществляется по схеме статистической проверки гипотез.

Для проверки гипотезы $H_0 : a_1 = \alpha_1,\ \left(H_1 : a_1 \neq \alpha_1\right)$ используется статистика

$$t = \frac{a_1 - \alpha_1}{S_{a_1}}$$

которая при справедливости H_0 имеет распределение Стьюдента с числом степеней свободы $\nu = n - 2$, где n — объем выборки. Следовательно, H_0 отклоняется на основании данного критерия, если

$$\left|T_{\text{í àáë}}\right| = \left|\frac{a_1 - \alpha_1}{S_{a_1}}\right| \geq t_{\frac{\alpha}{2},n-2}$$

где α — требуемый уровень значимости. При невыполнении этого неравенства считается, что нет оснований для отклонения H_0.

Наиболее важной на начальном этапе статистического анализа построенной модели является задача установления наличия линейной зависимости между Y и X. Эта проблема может быть решена по схеме: $H_0 : a_1 = 0,\ \left(H_1 : a_1 \neq 0\right)$ Гипотеза в такой постановке называется *гипотезой о статистической значимости коэффициента регрессии.* При этом, если принимается H_0, то есть основания считать, что величина Y не зависит от X, в этом случае говорят, что коэффициент a_1 *статистически незначим* (он слишком близок к нулю). При отклонении H_0 коэффициент a_1 считается *статистически значимым,* что указывает на наличие определенной линейной зависимости между Y и X. В этом случае рассматривается двусторонняя критическая область, так как важным является именно отличие от нуля коэффициента регрессии, и он может быть как положительным, так и отрицательным.

Поскольку полагается, что $\alpha_1 = 0$, то значимость оцениваемого коэффициента регрессии a_1 проверяется с помощью анализа отношения его величины к его стандартной ошибке S_{a_1}.

$$t = \frac{a_1}{S_{a_1}}$$

Эта дробь имеет распределение Стьюдента с числом степеней свободы $\nu = n - 2$, где n — число наблюдений, и называется *t-статистикой.* Для t-статистики проверяется нулевая гипотеза о равенстве ее нулю. Очевидно, $t = 0$ равнозначно $a_1 = 0$, поскольку t пропорциональна a_1. Фактически это свидетельствует об отсутствии линейной связи между X и Y. По аналогичной схеме на основе t-статистики проверяется гипотеза о статистической значимости коэффициента a_0:

$$t = \frac{a_0}{S_{a_0}}$$

Для парной регрессии более важным является анализ статистической значимости коэффициента a_1, так как именно в нем скрыто влияние объясняющей переменной X на зависимую переменную Y.

При оценке значимости коэффициента линейной регрессии на начальном этапе можно использовать следующее *«грубое» правило.*

Если стандартная ошибка коэффициента больше его модуля, $|t| \le 1$, то коэффициент не может быть признан значимым, так как доверительная вероятность при двусторонней альтернативной гипотезе составит менее чем 0,7.

Если $1 < |t| \le 2$, то найденная оценка может рассматриваться как относительно (слабо) значимая. Доверительная вероятность в этом случае лежит между значениями 0,7 и 0,95.

Если $2 < |t| \le 3$, то это свидетельствует о значимой линейной связи между X и Y. В этом случае доверительная вероятность колеблется от 0,95 до 0,99.

Наконец, если $|t| > 3$, то это гарантирует наличие линейной связи. Вместе с тем, в каждом конкретном случае имеет значение число наблюдений, чем их больше, тем надежнее при прочих равных условиях выводы о значимости коэффициента. Однако для $n > 10$ предложенное «грубое» правило практически всегда работает.

Ранее для коэффициентов a_0 и a_1 получили:

$$a_0 = \sum d_i y_i, \qquad \left(d_i = \frac{1}{n} - \overline{x} c_i \right)$$

$$a_1 = \sum c_i y_i \qquad \left(c_i = \frac{x_i - \overline{x}}{\sum \left(x_i - \overline{x} \right)^2} \right)$$

Т.е, a_0 и a_1 являются линейными комбинациями y_i, которые в свою очередь также является линейной комбинацией ε_i,

$$y_i = M\left(Y \middle| X = x_i \right) = \alpha_0 + \alpha_1 x_i + \varepsilon_i$$

(при этом считается, что α_0, α_1 и x_i — константы или неслучайные величины). Тогда a_0 и a_1 через y_i являются линейными функциями от ε_i, имеющими нормальное распределение. Значит, a_0 и a_1 также распределены нормально,

$$\left.\begin{aligned} & M(a_1) = \alpha_1, \;\; D(a_1) \approx S_{a_1}^2 = \frac{S^2}{\sum \left(x_i - \overline{x} \right)^2}; \\ & M(a_0) = \alpha_0, \; D(a_0) \approx S_{a_0}^2 = \frac{S^2 \sum x_i^2}{n \sum \left(x_i - \overline{x} \right)^2} = \overline{x^2} S_{a_1}^2 \end{aligned}\right\} \Rightarrow \begin{cases} a_1 \; N\left(\alpha_1, D(a_1) \right) \\ a_0 \; N\left(\alpha_0, D(a_0) \right) \end{cases}$$

Статистики

$$t_{a_0} = \frac{a_0 - \alpha_0}{S_{a_0}}, \quad t_{a_1} = \frac{a_1 - \alpha_1}{S_{a_1}}$$

имеют распределение Стьюдента с числом степеней свободы $\nu = n - 2$. Для определения $100(1-\alpha)\%$ -го доверительного интервала с помощью критических точек распределения Стьюдента по доверительной вероятности $\gamma = 1 - \alpha$ и числу степеней свободы ν определяют критическое значение $t_{\alpha/2,n-2}$, удовлетворяющее условию

$$P\left(|t| < t_{\frac{\alpha}{2},n-2}\right) = 1 - \alpha$$

или

$$\left.\begin{aligned} P\left(|t_{a_0}| < t_{\frac{\alpha}{2},n-2}\right) &= 1-\alpha \\ P\left(|t_{a_1}| < t_{\frac{\alpha}{2},n-2}\right) &= 1-\alpha \end{aligned}\right\} \Rightarrow \left\{\begin{aligned} P\left(\left|\frac{a_0 - \alpha_0}{S_{a_0}}\right| < t_{\frac{\alpha}{2},n-2}\right) &= 1-\alpha \\ P\left(\left|\frac{a_1 - \alpha_1}{S_{a_1}}\right| < t_{\frac{\alpha}{2},n-2}\right) &= 1-\alpha \end{aligned}\right.$$

$$P\left(-t_{\frac{\alpha}{2},n-2} < \frac{a_0-\alpha_0}{S_{a_0}} < t_{\frac{\alpha}{2},n-2}\right)=1-\alpha \quad P\left(a_0 - S_{a_0}t_{\frac{\alpha}{2},n-2} < \alpha_0 < a_0 + S_{a_0}t_{\frac{\alpha}{2},n-2}\right)=1-\alpha$$

$$P\left(-t_{\frac{\alpha}{2},n-2} < \frac{a_1-\alpha_1}{S_{a_1}} < t_{\frac{\alpha}{2},n-2}\right)=1-\alpha \quad P\left(a_1 - S_{a_1}t_{\frac{\alpha}{2},n-2} < \alpha_1 < a_1 + S_{a_1}t_{\frac{\alpha}{2},n-2}\right)=1-\alpha$$

$$P\left(a_0 - t_{\frac{\alpha}{2},n-2}\sqrt{\frac{S^2\sum x_i^2}{n\sum\left(x_i-\bar{x}\right)^2}} < \alpha_0 < a_0 + t_{\frac{\alpha}{2},n-2}\sqrt{\frac{S^2\sum x_i^2}{n\sum\left(x_i-\bar{x}\right)^2}}\right)=1-\alpha$$

$$P\left(a_1 - t_{\frac{\alpha}{2},n-2}\sqrt{\frac{S^2}{\sum\left(x_i-\bar{x}\right)^2}} < \alpha_1 < a_1 + t_{\frac{\alpha}{2},n-2}\sqrt{\frac{S^2}{\sum\left(x_i-\bar{x}\right)^2}}\right)=1-\alpha$$

Полученные соотношения для вероятностей определяют доверительные интервалы

$$\left(a_0 - S_{a_0}t_{\frac{\alpha}{2},n-2}; a_0 + S_{a_0}t_{\frac{\alpha}{2},n-2}\right)$$

$$\left(a_1 - S_{a_1}t_{\frac{\alpha}{2},n-2}; a_1 + S_{a_1}t_{\frac{\alpha}{2},n-2}\right)$$

которые с надежностью $(1-\alpha)$ накрывают определяемые параметры α_0 и α_1.

Одной из центральных задач эконометрического анализа является предсказание (прогнозирование) значений зависимой переменной при определенных значениях объясняющих переменных. Очевидно, можно: либо предсказать условное математическое ожидание зависимой переменной при определенных значениях объясняющих переменных (предсказание среднего значения), либо прогнозировать некоторое конкретное значение зависимой переменной (предсказание конкретного значения).

Предсказание среднего значения. Пусть построено уравнение парной регрессии $\breve{y}_i = a_0 + a_1 x_i$, на основе которого необходимо предсказать условное математическое ожидание $M\left(Y \middle| X = x_p\right)$ переменной Y при $X = x_p$, которое является оценкой y_p. Естественным является вопрос, как сильно может уклониться модельное среднее значение y_p, рассчитанное по эмпирическому уравнению регрессии, от соответствующего условного математического ожидания. Это можно получить с помощью интервальных оценок, построенных с заданной надежностью для любого конкретного значения объясняющей переменной.

Получено раннее

$$\tilde{Y}_p = a_0 + a_1 x_p$$

$$\tilde{y}_i = a_0 + a_1 x_i \Rightarrow \begin{array}{l} a_0 = \sum d_i y_i, \quad \left(d_i = \frac{1}{n} - \bar{x} c_i \right) \\ a_1 = \sum c_i y_i \quad \left(c_i = \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2} \right) \end{array} \Rightarrow \tilde{Y}_p = \sum d_i y_i + \sum c_i y_i x_p$$

$$\tilde{Y}_p = \left(\sum d_i + x_p c_i \right) y_i$$

Значит $\tilde{Y}_p$ является линейной комбинацией нормальных СВ и имеет нормальное распределение.

$$M\left(\tilde{Y}_p\right) = M\left(a_0 + a_1 x_p\right) = M(a_0) + M(a_1) x_p = \alpha_0 + \alpha_1 x_p$$

$$D\left(\tilde{Y}_p\right) = D\left(a_0 + a_1 x_p\right) = D(a_0) + D(a_1) x_p^2 + 2 x_p \operatorname{cov}(a_0, a_1)$$

Поскольку,

$$D(X+Y) = D(X) + D(Y) + 2\operatorname{cov}(X,Y)$$

$$D(cX) = c^2 D(X)$$

$$\operatorname{cov}(X, cY) = c \operatorname{cov}(X,Y)$$

$$\operatorname{cov}(a_0,a_1)=M\left[\left(a_0-M(a_0)\right)\left(a_1-M(a_1)\right)\right]=$$

$$=M\left[\left(a_0-\alpha_0\right)\left(a_1-\alpha_1\right)\right]=$$

$$=M\left[\left(\bar{y}-\bar{x}a_1-\alpha_0+\bar{x}\alpha_1\right)\left(a_1-\alpha_1\right)\right]=$$

$$=M\left[\left(-\bar{x}a_1+\bar{x}\alpha_1\right)\left(a_1-\alpha_1\right)\right]=$$

$$=-\bar{x}M\left[\left(a_1-\alpha_1\right)\left(a_1-\alpha_1\right)\right]=$$

$$=-\bar{x}M\left[\left(a_1-\alpha_1\right)^2\right]=$$

$$=-\bar{x}M\left[\left(a_1-M(a_1)\right)^2\right]=$$

$$=-\bar{x}D(a_1)=$$

$$=-\bar{x}\frac{\sigma^2}{\sum\left(x_i-\bar{x}\right)^2}=$$

$$D\left(Y_p\right)=\frac{\sigma^2\sum x_i^2}{n\sum\left(x_i-\bar{x}\right)^2}+x_p^2-2x_p\frac{\sigma^2\bar{x}}{\sum\left(x_i-\bar{x}\right)^2}=$$

$$=M\left[\left(\bar{y}-\bar{x}\alpha_1-\alpha_0+\bar{x}\alpha_1\right)\left(a_1-\alpha_1\right)\right]=$$

$$=M\left[\left(a_0-\alpha_0\right)\left(a_1-\alpha_1\right)\right]=\operatorname{cov}(a_0,a_1)=M\left[\left(a_0-M(a_0)\right)\left(a_1-M(a_1)\right)\right]=$$

$$=M\left[\left(\bar{y}-\bar{x}a_1-\alpha_0+\bar{x}\alpha_1\right)\left(a_1-\alpha_1\right)\right]=$$

$$=M\left[\left(-\bar{x}a_1+\bar{x}\alpha_1\right)\left(a_1-\alpha_1\right)\right]=$$

$$=-\bar{x}M\left[\left(a_1-\alpha_1\right)^2\right]=$$

$$=-\bar{x}M\left[\left(a_1-M(a_1)\right)^2\right]=$$

$$=-\bar{x}D(a_1)=$$

$$=-\bar{x}\frac{\sigma^2}{\sum\left(x_i-\bar{x}\right)^2}$$

$$D(a_1)=\sigma^2\frac{1}{\sum\left(x_i-\bar{x}\right)^2}=$$

$$D(a_0)=\sigma^2\frac{\sum x_i^2}{n\sum\left(x_i-\bar{x}\right)^2}=$$

$$\sum x_i^2=n\sigma_x^2+n\bar{x}^2=\sum\left(x_i-\bar{x}\right)^2+n\bar{x}^2$$

$$=\sigma^2\left(\frac{1}{n}+\frac{\left(\bar{x}-x_p\right)^2}{\sum\left(x_i-\bar{x}\right)^2}\right)$$

$$D\left(Y_p\right)=\frac{\sigma^2}{n}+\frac{\sigma^2\left(\bar{x}-x_p\right)^2}{\sum\left(x_i-\bar{x}\right)^2}=\sigma^2\left(\frac{1}{n}+\frac{\left(\bar{x}-x_p\right)^2}{\sum\left(x_i-\bar{x}\right)^2}\right)$$

$$D\left(Y_p\right)=\frac{\sigma^2\sum x_i^2}{n\sum\left(x_i-\bar{x}\right)^2}+x_p^2-2x_p\frac{\sigma^2\bar{x}}{\sum\left(x_i-\bar{x}\right)^2}=$$

$$=\frac{\sigma^2}{n}+\frac{\sigma^2\bar{x}^2}{\sum\left(x_i-\bar{x}\right)^2}+x_p^2-2x_p\frac{\sigma^2\bar{x}}{\sum\left(x_i-\bar{x}\right)^2}=$$

$$=\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\sum\left(x_i-\bar{x}\right)^2}+\frac{x_p^2}{\sigma^2}-2x_p\frac{\bar{x}}{\sum\left(x_i-\bar{x}\right)^2}\right)=$$

$$=\sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2+x_p^2-2x_p\bar{x}}{\sum\left(x_i-\bar{x}\right)^2}\right)=$$

$$=\sigma^2\left(\frac{1}{n}+\frac{\left(\bar{x}-x_p\right)^2}{\sum\left(x_i-\bar{x}\right)^2}\right)$$

Получено ранее

$$\left.\begin{aligned} D\left(\tilde{Y}_p\right) &= \sigma^2\left(\frac{1}{n}+\frac{\left(\bar{x}-x_p\right)^2}{\sum\left(x_i-\bar{x}\right)^2}\right) \\ \sigma^2 \approx S^2 &= \frac{\sum e_i^2}{n-2} \end{aligned}\right\} \Rightarrow S^2\left(\tilde{Y}_p\right) = \frac{\sum e_i^2}{n-2}\left(\frac{1}{n}+\frac{\left(\bar{x}-x_p\right)^2}{\sum\left(x_i-\bar{x}\right)^2}\right)$$

Выборочная исправленная дисперсия

Статистика

$$T(\tilde{Y}_p) = \frac{\tilde{Y}_p - (\alpha_0 + \alpha_1 x_p)}{S(\tilde{Y}_p)}$$

имеют распределение Стьюдента с числом степеней свободы $\nu = n - 2$. Для определения $100(1-\alpha)\%$ -го доверительного интервала с помощью критических точек распределения Стьюдента по доверительной вероятности $\gamma = 1 - \alpha$ и числу степеней свободы ν определяют критическое значение $t_{\alpha/2,n-2}$, удовлетворяющее условию

$$P\left(\left|T(\tilde{Y}_p)\right| < t_{\frac{\alpha}{2},n-2}\right) = 1-\alpha \quad \text{или} \quad P\left(\left|\frac{\tilde{Y}_p - (\alpha_0 + \alpha_1 x_p)}{S(\tilde{Y}_p)}\right| < t_{\frac{\alpha}{2},n-2}\right) = 1-\alpha$$

После преобразований получим

$$P\left(a_0 + a_1 x_p - S(\tilde{Y}_p) t_{\frac{\alpha}{2},n-2} < \alpha_0 + \alpha_1 x_p < a_0 + a_1 x_p + S(\tilde{Y}_p) t_{\frac{\alpha}{2},n-2}\right) = 1-\alpha$$

Доверительный интервал для $M\left(Y\middle|X=x_p\right)=\alpha_0+\alpha_1 x_p$ имеет вид:

$$\left(a_0+a_1x_p-t_{\frac{\alpha}{2},n-2}\sqrt{\frac{\sum e_i^2}{n-2}\left(\frac{1}{n}+\frac{\left(\bar{x}-x_p\right)^2}{\sum\left(x_i-\bar{x}\right)^2}\right)};\ a_0+a_1x_p+t_{\frac{\alpha}{2},n-2}\sqrt{\frac{\sum e_i^2}{n-2}\left(\frac{1}{n}+\frac{\left(\bar{x}-x_p\right)^2}{\sum\left(x_i-\bar{x}\right)^2}\right)}\right)$$

который с надежностью $(1-\alpha)$ накрывают прогнозируемое среднее значение, условное математическое ожидание $M\left(Y\middle|X=x_p\right)$.

Для проверки гипотезы

$$H_0 : M(Y|X = x_p) = y_p$$

$$H_1 : M(Y|X = x_p) \neq y_p$$

используется статистика:

$$T = \frac{M(Y|X = x_p) - y_p}{\sqrt{\frac{\sum e_i^2}{n-2}\left(\frac{1}{n} + \frac{(\bar{x} - x_p)^2}{\sum (x_i - \bar{x})^2}\right)}}$$

имеющая распределение Стьюдента с числом степеней свободы $\nu = n - 2$. Поэтому H_0 отклоняется, если $|T_{\text{íàáë}}| \geq t_{\frac{\alpha}{2}, n-2}$ (α — требуемый уровень значимости).

Предсказание индивидуальных значений зависимой переменной. Иногда более важно знать дисперсию *Y,* чем её средние значения (доверительные интервалы для условных математических ожиданий). Что позволяет определить допустимые границы для конкретного значения *Y*.

Пусть y_0 есть некоторое возможное значение переменной Y при определенном значении x_p объясняющей переменной *X*. Предсказанное по уравнению регрессии значение Y при $X = x_p$ составляет y_p. Если рассматривать значение y_0 как СВ Y_0, а y_p — как СВ Y_p,то

имеют нормальное распределение

$$Y_0 \sim N\left(\alpha_0 + \alpha_1 x_p, \sigma^2\right)$$

$$Y_p \sim N\left(\alpha_0 + \alpha_1 x_p, \sigma^2\left(\frac{1}{n} + \frac{\left(\bar{x} - x_p\right)^2}{\sum\left(x_i - \bar{x}\right)^2}\right)\right)$$

СВ Y_0 и Y_p являются независимыми, а следовательно, СВ $U = Y_0 - Y_p$ имеет нормальное распределение

$$M(U) = 0, \quad D(U) = \sigma^2 \left(\frac{1}{n} + \frac{\left(\overline{x} - x_p\right)^2}{\sum \left(x_i - \overline{x}\right)^2} \right)$$

Тогда статистика

$$\frac{U}{S_U} = \frac{Y_0 - Y_p}{S\sqrt{1 + \frac{1}{n} + \frac{\left(\overline{x} - x_p\right)^2}{\sum \left(x_i - \overline{x}\right)^2}}}, \qquad \left(S = \sqrt{\frac{\sum e_i^2}{n-2}} \right)$$

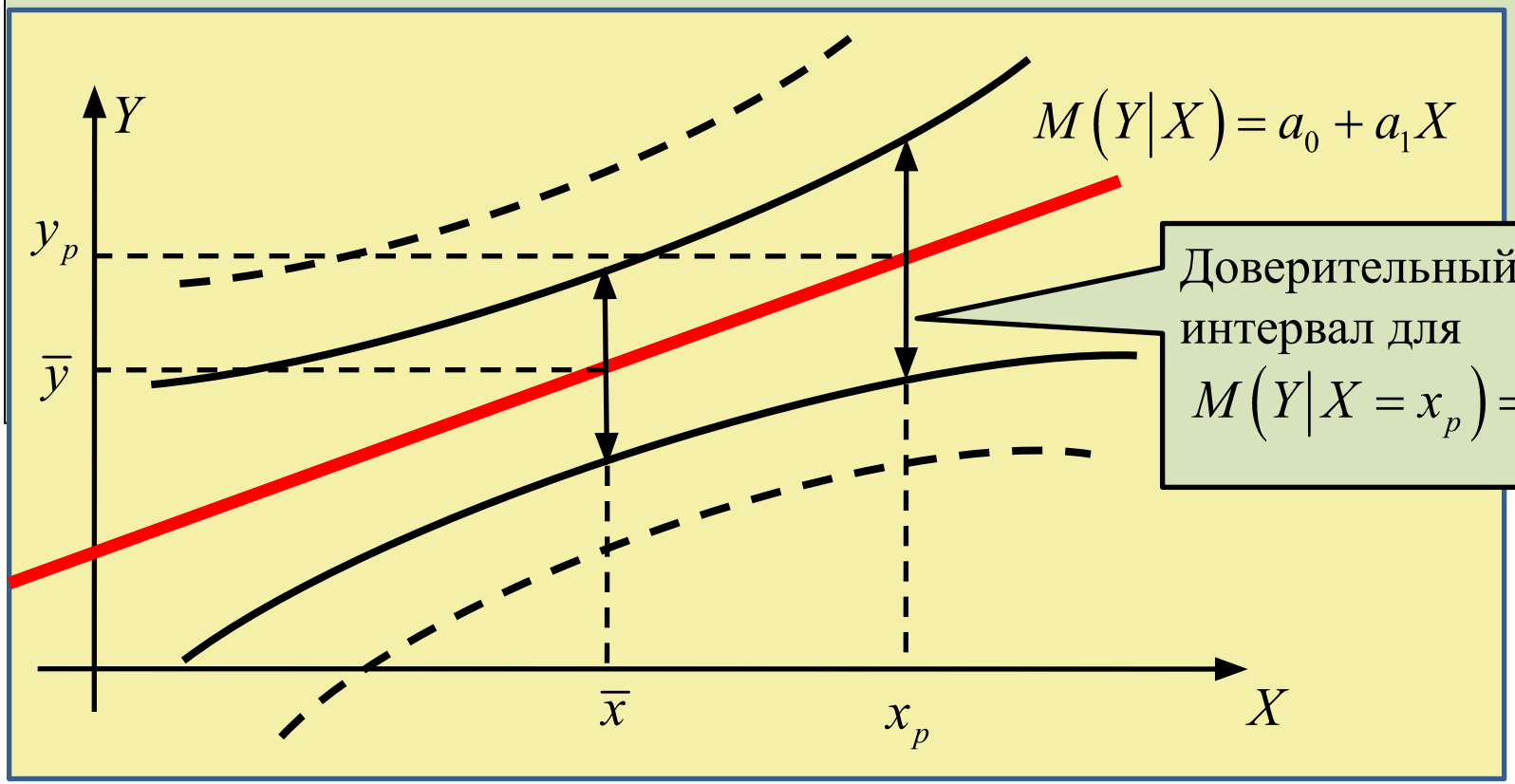
имеет распределение Стьюдента с $\nu = n - 2$ степенями свободы. Значит

$$P\left(-t_{\frac{\alpha}{2},n-2} \le \frac{Y_0 - Y_p}{S\sqrt{1 + \frac{1}{n} + \frac{\left(\overline{x} - x_p\right)^2}{\sum \left(x_i - \overline{x}\right)^2}}} \le t_{\frac{\alpha}{2},n-2} \right) = 1 - \alpha$$

Интервал

$$\left(a_0 + a_1 x_p \pm t_{\frac{\alpha}{2},n-2} S \sqrt{1 + \frac{1}{n} + \frac{\left(\overline{x} - x_p\right)^2}{\sum \left(x_i - \overline{x}\right)^2}} \right)$$

определяет границы, за пределами которых могут оказаться не более $100\alpha\%$ точек наблюдений при $X = x_p$. Этот интервал шире доверительного интервала для условного математического ожидания.



Мерой качества уравнения регрессии (соответствия уравнения регрессии статистическим данным) является *коэффициент детерминации* R^2. В случае парной регрессии коэффициент детерминации будет совпадать с квадратом коэффициента корреляции. Коэффициент детерминации рассчитывается по формуле

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \overline{y})^2}$$

Уравнение линейной регрессии, построенное на основании эмпирических (наблюдаемых) парных данных $\{x_i, y_i\}_{i=\overline{1,n}}$, имеет вид

$$\tilde{y} = a_0 + a_1 x$$

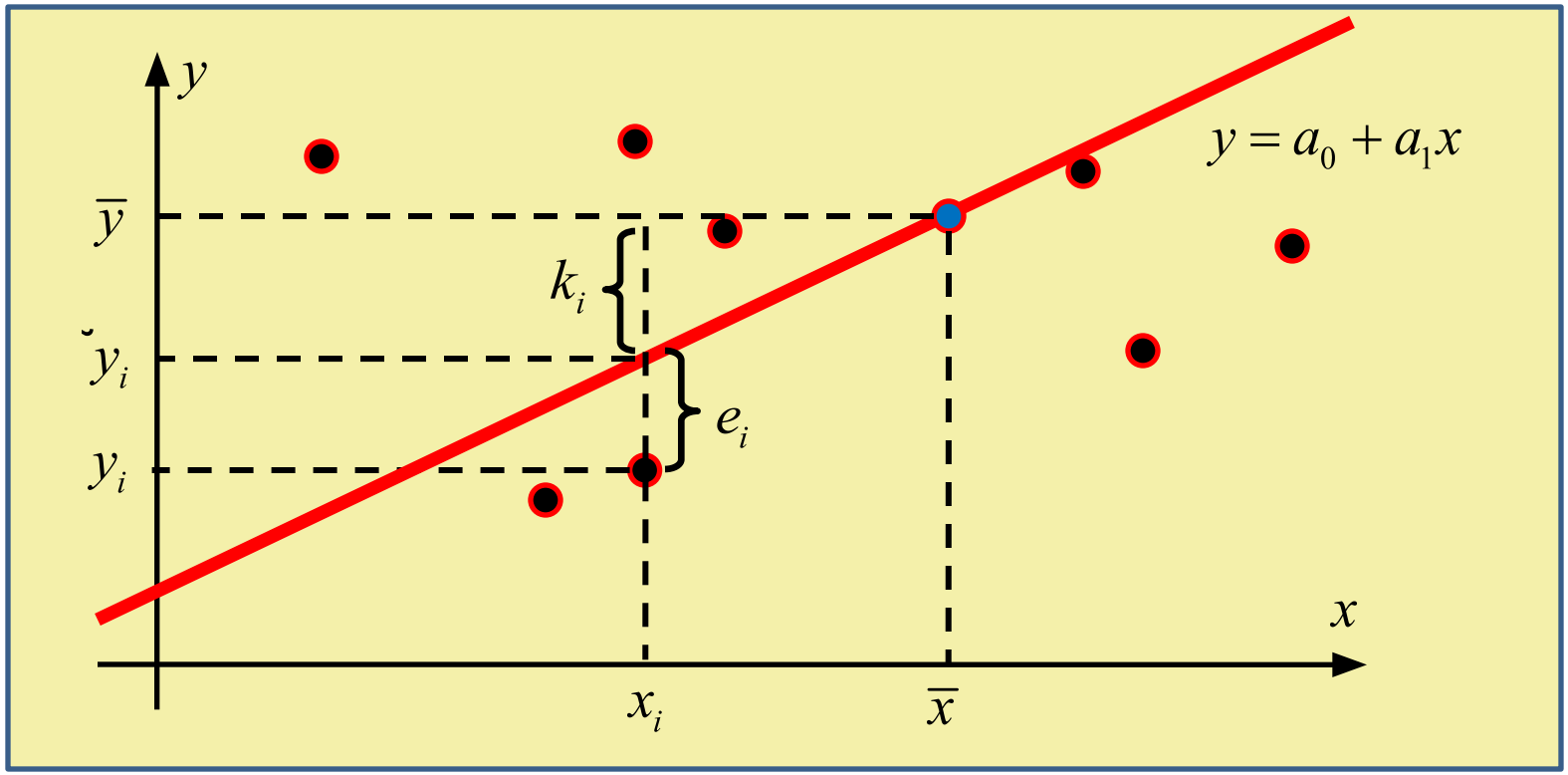
Наблюдаемые (реальные) значения y_i, отличаются от модельных $\tilde{y}_i = a_0 + a_1 x_i$ на величину e_i:

$$y_i - \tilde{y}_i = e_i$$

Очевидно,

$$y_i - \overline{y} = (y_i - \overline{y}) + (y_i - \tilde{y}_i) \Rightarrow y_i - \overline{y} = k_i + e_i$$

Где $(y_i - \overline{y})$ есть отклонение i–ого наблюдаемого значения от среднего значения $\overline{y}$ зависимой переменной, k_i – отклонение i–ого значения на прямой регрессии от среднего значения зависимой переменной, e_i есть отклонение i–ого значения y_i от расчётного (модельного) значения $\breve{y}_i$ определяемого регрессией.



$$y_i - \bar{y} = (y_i - \bar{y}) + (y_i - \tilde{y}_i) \Rightarrow \begin{cases} y_i - \bar{y} = k_i + e_i \\ k_i = \tilde{y}_i - \bar{y} \\ e_i = y_i - \tilde{y}_i \end{cases}$$

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \bar{y})^2 + \sum (y_i - \tilde{y}_i)^2 + 2\sum (y_i - \bar{y})(y_i - \tilde{y}_i)$$

$$\sum (y_i - \bar{y})^2 = \sum k_i^2 + \sum e_i^2 + 2\sum (y_i - \bar{y})e_i$$

$$\sum (y_i - \bar{y})^2 = \sum k_i^2 + \sum e_i^2$$

$$\sum (y_i - \bar{y})e_i = \sum \tilde{y}_i e_i - \bar{y}\sum e_i = \sum \tilde{y}_i e_i =$$

$$= \sum (a_0 + a_1 x_i)e_i =$$

$$= a_0 \sum e_i + a_1 \sum x_i e_i = a_1 \sum x_i e_i = 0$$

Очевидно, $\sum (y_i - \bar{y})^2$ — *общая (полная) сумма квадратов* может интерпретироваться как мера общего разброса (рассеивания) переменной y относительно $\bar{y}$. $\sum k_i^2 = \sum (y_i - \bar{y})^2$ — *объясненная сумма квадратов,* интерпретируемая как мера разброса, объяснимого с помощью регрессии. Сумма $\sum e_i^2 = \sum (y_i - \bar{y})^2$ — *остаточная (необъясненная) сумма квадратов,* являющаяся мерой остаточного, не объясненного уравнением регрессии разброса (разброса точек вокруг линии регрессии).

$$\sum (y_i - \bar{y})^2 = \sum k_i^2 + \sum e_i^2$$

$$1 = \frac{\sum k_i^2}{\sum (y_i - \bar{y})^2} + \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} \qquad R^2 = \frac{\sum k_i^2}{\sum (y_i - \bar{y})^2}$$

обозначим

$$1 = R^2 + \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} \qquad R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}$$

Коэффициент детерминации $R^2 = \frac{\sum k_i^2}{\sum (y_i - \bar{y})^2}$ определяет долю разброса зависимой переменной, объяснимую регрессией Y на X.

Дробь $\frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}$ определяет долю разброса зависимой переменной, не объясненную регрессией Y на X.

Очевидно для коэффициента детерминации справедливо $0 \le R^2 \le 1$

Коэффициент детерминации R^2 является мерой, позволяющей определить, в какой степени найденная прямая регрессии дает лучший результат для объяснения поведения зависимой переменной Y, чем горизонтальная прямая $Y = \bar{y}$. Чем теснее линейная связь между X и Y, тем ближе коэффициент детерминации к единице, чем слабее такая связь, тем R^2 ближе к нулю

$$R^2 = \frac{\sum k_i^2}{\sum (y_i - \bar{y})^2} = \frac{\sum (\tilde{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = \frac{\sum (a_0 + a_1 x_i - (a_0 + a_1 \bar{x}))^2}{\sum (y_i - \bar{y})^2} =$$

$$= \frac{\sum (a_1 x_i - a_1 \bar{x})^2}{\sum (y_i - \bar{y})^2} = a_1^2 \frac{\sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2} =$$

$$= \left(\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \right)^2 \frac{\sum (x_i - \bar{x})^2}{\sum (y_i - \bar{y})^2} = \frac{\left(\sum (x_i - \bar{x})(y_i - \bar{y}) \right)^2}{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2} =$$

$$= \left(\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \right)^2 = r_{xy}^2$$

Коэффициент детерминации равен квадрату коэффициента корреляции