Линейные регрессионные модели

1. Парная линейная регрессия

1.1. Взаимосвязи переменных

В естественных науках большей частью имеют дело со строгими (функциональными) зависимостями, при которых каждому значению одной переменной соответствует единственное значение другой. Однако в подавляющем большинстве случаев между экономическими переменными таких зависимостей нет. Например, нет строгой зависимости между доходом и потреблением, ценой и спросом, производительностью труда и стажем работы и т.д. Это связано с целым рядом причин и, в частности, с тем, что, во-первых, при анализе влияния одной переменной на другую не учитывается целый ряд других факторов, влияющих на нее. Во-вторых, это влияние может быть не прямым, а проявляться через цепочку других факторов; в-третьих, многие такие воздействия носят случайный характер и т.д. Поэтому в эконометрике говорят не о функциональных, а о корреляционных, либо статистических зависимостях. *Статистической* называют зависимость, при которой изменение одной из величин влечет 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*.

В эконометрике при рассмотрении взаимосвязей между двумя переменными X и Y выделяют 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, как правило, к увеличению потребления, рост цены – к снижению спроса и т.д. Однако, как было указано выше, 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. Другими словами, каждому конкретному значению объясняющей переменной (набору объясняющих переменных) соответствует некоторое вероятностное распределение зависимой переменной (рассматриваемой как случайная величина). Поэтому анализируют, как объясняющая(ие) переменная(ые) влияет(ют) на зависимую переменную «в среднем». Зависимость такого типа, выражаемая соотношением $M(Y \mid x) = f(x)$, называется *регрессией* Y на X. При рассмотрении зависимости двух случайных величин говорят о парной регрессии. Зависимость нескольких переменных, выражаемая функцией $M(Y \mid x_1, x_2, ..., x_m) = f(x_1, x_2, ..., x_m)$ называют *множественной регрессией*.

Для отражения того факта, что реальные значения зависимой переменной не всегда совпадают с ее условными математическими ожиданиями и могут быть различными при одном и том же значении объясняющей переменной (наборе объясняющих переменных), фактическая зависимость должна быть дополнена слагаемым ε, которое, по существу, является случайной величиной и указывает на статистическую суть зависимости. Из этого следует, что связи между зависимой и объясняющей(ими) переменными выражаются соотношениями $Y = M(Y \mid x) + \varepsilon$ и $Y = M(Y \mid x_1, x_2, ..., x_m) + \varepsilon$, называемыми *регрессионными* моделями (*уравнениями*).

Задача построения качественного уравнения регрессии, соответствующего выборочным (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 проверка адекватности уравнения эмпирическим данным, совершенствование уравнения.

Выбор формулы связи переменных называется *спецификацией* уравнения регрессии. В случае парной регрессии выбор формулы обычно осуществляется по графическому изображению реальных статистических данных в виде точек в декартовой системе координат, которое называется *корреляционным полем* (*диаграммой рассеивания*).

Пример 1. Диаграмма рассеяния на рис. 1 соответствует данным о годовом располагаемом доходе *X* и годовых расходах *Y* на личное потребление (в 1999 г., в условных единицах) 20 семей. Эти данные представлены в таблице.

x_i	2508	2572	2408	2522	2700	2531	2390	2595	2524	2685	2435	2354
y_i	2406	2464	2336	2281	2641	2385	2297	2416	2460	2549	2311	2278

x_i	2404	2381	2581	2529	2562	2624	2407	2448
y_i	2240	2183	2408	2379	2378	2554	2232	2356

По расположению точек на корреляционном поле естественно предположить, что зависимость между X и Y близка к линейной. Для построения диаграммы рассеяния в Excel можно воспользоваться командой меню *Вставка* $\Rightarrow$ *Точечнаядиаграмма*.

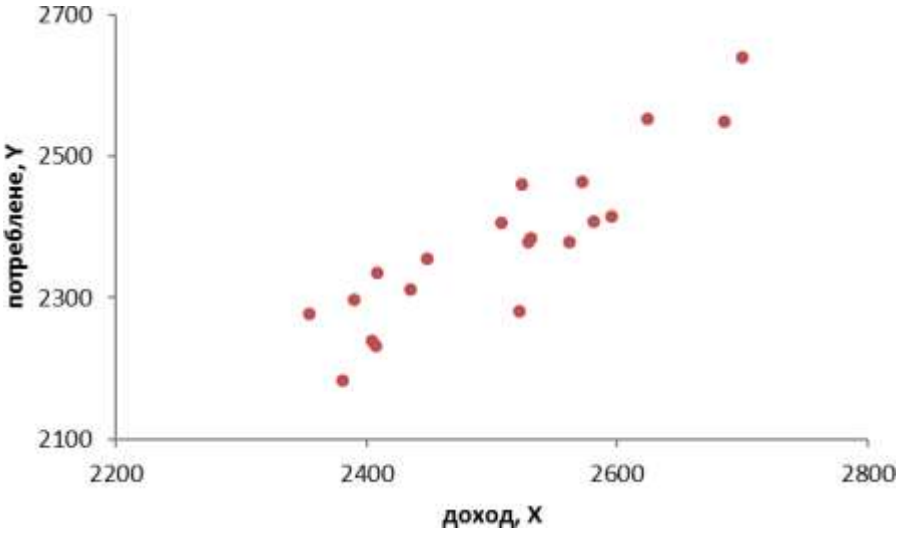


Рис. 1

Более точной мерой зависимости между величинами является коэффициент корреляции r_{xy}, $-1 \le r_{xy} \le 1$, который является безразмерной величиной и показывает *степень линейной связи* двух переменных. Величина $|r_{xy}|$, близкая к 1, указывает, что зависимость между данными случайными величинами почти линейная. Для вычисления коэффициента корреляции по выборке в Excel можно активировать Мастер функций f_x, в окне Категория выбрать *Статистические,* в окне Функция выбрать *Коррел*. Для данного примера находим $r_{xy} = 0{,}895$, что позволяет сделать вывод о том, что связь между рассматриваемыми переменными близка к линейной. Это также подтверждается расположением точек на корреляционном поле.

1.2. Метод наименьших квадратов

Если регрессия линейна, то речь ведут о *линейной регрессии.* Модель линейной регрессии является наиболее распространенным (и простым) уравнением зависимости между экономическими переменными. Кроме того, построенное линейное уравнение может быть начальной точкой эконометрического анализа.

Из предыдущих рассуждений ясно, что *парная линейная регрессия* (*теоретическое линейное уравнение регрессии*) представляет собой линейную функцию между условным математическим ожиданием $M(Y \mid X = x)$ зависимой переменной Y и одной переменной X:

$$M(Y \mid X = x) = \beta_0 + \beta_1 \cdot x. \qquad (1.1)$$

Отметим, что принципиальной в данном случае является линейность по параметрам β_0 и β_1 уравнения. Для отражения того факта, что каждое индивидуальное значение y отклоняется от соответствующего условного математического ожидания, необходимо ввести в соотношение (1.1) случайное слагаемое ε:

$$Y = M(Y \mid X = x) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon. \qquad (1.2)$$

Соотношение (1.2) называется *теоретической линейной регрессионной моделью*; β_0 и β_1 – *теоретическими параметрами* (*теоретическими коэффициентами*) регрессии; ε – *случайным отклонением (случайной ошибкой)*.

Для определения значений теоретических коэффициентов регрессии необходимо знать и использовать все значения переменных X и Y генеральной совокупности, что практически невозможно. Таким образом, нужно уметь оценивать коэффициенты β_0 и β_1 на основе статистических данных (выборки) $(x_i, y_i), i = 1,2,..., n$, переменных X и Y. Тогда по выборке ограниченного объема мы сможем построить так называемое *эмпирическое* (*выборочное*) *уравнение регрессии*

$$\hat{y} = b_0 + b_1 x, \qquad (1.3)$$

где $\hat{y}$ – оценка условного математического ожидания $M(Y \mid X = x)$; b_0 и b_1 – оценки неизвестных параметров β_0 и β_1, называемые *эмпирическими* (*выборочными*) *коэффициентами регрессии*. В каждом конкретном наблюдении выборки имеем $y_i = b_0 + b_1 x_i + e_i$, где отклонение e_i – оценка теоретического случайного отклонения ε_i.

В силу несовпадения статистической базы для генеральной совокупности и выборки оценки b_0 и b_1 практически всегда отличаются от истинных значений коэффициентов β_0 и β_1, что приводит к несовпадению эмпирической и теоретической линий регрессии. Различные выборки из одной и той же генеральной совокупности обычно приводят к определению отличающихся друг от друга оценок. Задача состоит в том, чтобы по конкретной выборке (x_i, y_i), $i = 1,2...,n$, найти оценки b_0 и b_1 неизвестных параметров β_0 и β_1 так, чтобы построенная линия регрессии являлась бы наилучшей в определенном смысле среди всех других прямых. Другими словами, построенная прямая $\hat{y} = b_0 + b_1 x$ должна быть «ближайшей» к точкам наблюдений по их совокупности.

Рассмотрим задачу «наилучшей» аппроксимации набора наблюдений (x_i, y_i), $i = 1,2...,n$ линейным уравнением (1.3). На рис. 2 приведены диаграмма рассеяния наблюдений из примера 1 и линия регрессии. Величина $\hat{y}_i$ описывается как расчетное значение переменной y, которое соответствует x_i. Наблюдаемые значения y_i не лежат в точности на линии регрессии, т.е. не совпадают с расчетными значениями $\hat{y}_i$. Разность $y_i - \hat{y}_i$ обозначим e_i.

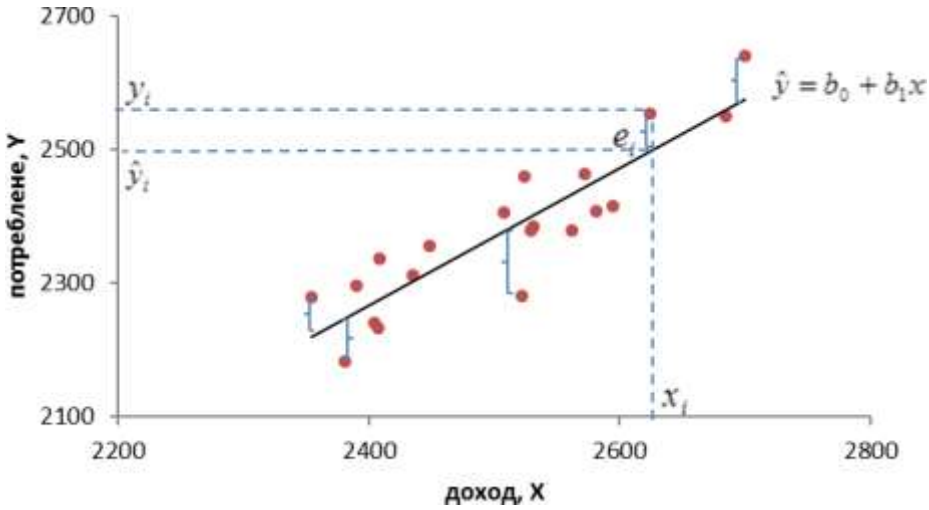


Рис. 2

Коэффициенты β_0 и β_1 будем оценивать по выборке с помощью *метода наименьших квадратов* (*МНК*, именно этот метод используется в Microsoft Excel). Суть МНК заключается в минимизации суммы квадратов отклонений наблюдаемых значений объясняемой переменной от ее расчетных значений, т.е. необходимо найти минимум квадратичной функции двух переменных b_0 и b_1:

$$F(b_0,b_1)=\sum_{i=1}^{n}e_i=\sum_{i=1}^{n}\left(y_i-\hat{y_i}\right)^2=\sum_{i=1}^{n}\left(y_i-b_0-b_1x_i\right)^2 .$$

Необходимым условием существования минимума функции двух переменных является равенство нулю ее частных производных по переменным b_0 и b_1:

$$\begin{cases}\dfrac{\partial F}{\partial b_0}=-2\sum\left(y_i-b_0-b_1x_i\right)=0\\ \dfrac{\partial F}{\partial b_1}=-2\sum\left(y_i-b_0-b_1x_i\right)=0.\end{cases}$$

Разрешив полученную систему относительно b_0 и b_1, получим формулы для вычисления оценок парной линейной регрессии:

$$\begin{cases}b_1=\dfrac{\overline{xy}-\bar{x}\cdot\bar{y}}{\overline{x^2}-\bar{x}^2},\\ b_0=\bar{y}-b_1\bar{x}\end{cases} \qquad (1.4)$$

где $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$, $\overline{x^2}=\frac{1}{n}\sum_{i=1}^{n}x_i^2$, $\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$, $\overline{xy}=\frac{1}{n}\sum_{i=1}^{n}x_iy_i$.

Замечания.

1) Нетрудно доказать, что имеет место соотношение $b_1=r_{xy}\cdot\frac{S_y}{S_x}$, где $r_{xy}=\dfrac{\overline{xy}-\bar{x}\cdot\bar{y}}{S_x\cdot S_y}$ – выборочный коэффициент корреляции; $S_x=\sqrt{\overline{x^2}-\bar{x}^2}$ и $S_y=\sqrt{\overline{y^2}-\bar{y}^2}$ – стандартные отклонения. Таким образом, коэффициент регрессии пропорционален коэффициенту корреляции.
2) Линия регрессии проходит через точку $(\bar{x},\bar{y})$, и выполняются равенства $\bar{e}=0$, $\bar{y}=\bar{\hat{y}}$.
3) В настоящее время оценки b_0 и b_1 обычно не вычисляют «вручную» по формулам (6). Для этого можно воспользоваться, например, настройкой «Пакет анализа» в приложении *Excel* (см. пример ниже).

1.3. Предпосылки регрессионного анализа

Линейная регрессионная модель с двумя переменными имеет вид

$$y = \beta_0 + \beta_1 x + \varepsilon ,$$

где Y – объясняемая переменная, X – объясняющая переменная, ε – случайная ошибка. Для того чтобы регрессионный анализ, основанный на МНК, давал наилучшие из всех возможных результаты, должны выполняться определенные условия (*условия Гаусса – Маркова*).

1. *Математическое ожидание случайной ошибки в любом наблюдении должно быть равно нулю, т.е.*

$$M(\varepsilon_i) = 0 \qquad (i = 1,2,...,n).$$

2. *Дисперсия случайной ошибки должна быть постоянной для всех наблюдений, т. е.*

$$D(\varepsilon_i) = M(\varepsilon_i^2) = \sigma^2 \qquad (i = 1,2,...,n).$$

3. *Случайные ошибки должны быть статистически независимы (некоррелированы) между собой, т.е.*

$$M(\varepsilon_i \cdot \varepsilon_j) = 0 \qquad (i \neq j).$$

4. *Объясняющая переменная x_i есть величина неслучайная.*

При выполнении условий Гаусса-Маркова модель называется классической нормальной линейной регрессионной моделью. Наряду с условиями Гаусса-Маркова обычно предполагается, что *случайная ошибка распределена нормально*, т. е. $\varepsilon_i \approx N(0,\sigma^2)$. Заметим, что в этом случае требование некоррелированности случайных ошибок эквивалентно их независимости.

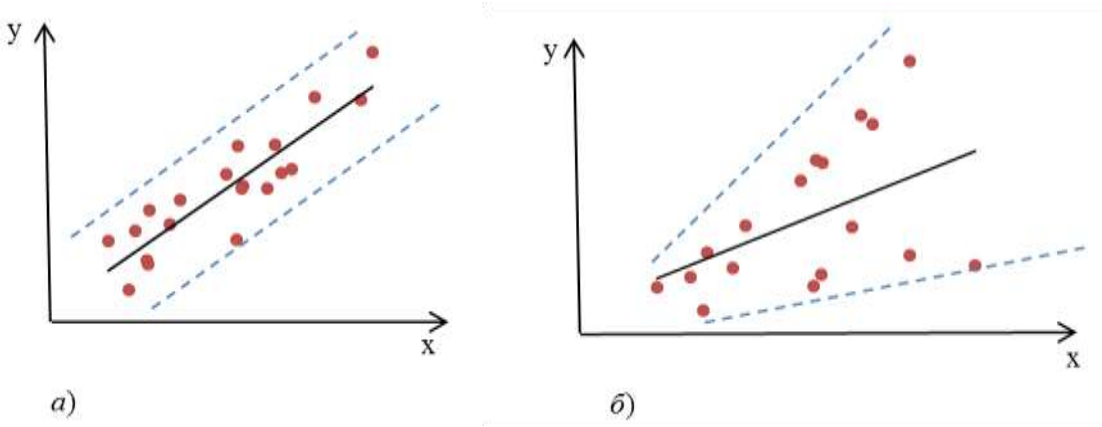
Рассмотрим подробнее условия и предположения, лежащие в основе регрессионного анализа.

Первое условие означает, что случайная ошибка не должна иметь систематического смещения. Если постоянный член включен в уравнение регрессии, то это условие выполняется автоматически.

Второе условие означает, что дисперсия случайной ошибки в каждом наблюдении имеет только одно значение. Под дисперсией σ^2 имеется в виду возможное поведение случайной ошибки до того, как была сделана выборка. Величина σ^2 неизвестна, и одна из задач регрессионного анализа состоит в ее оценке. Условие *независимости* дисперсии случайной ошибки от номера наблюдения называется *гомоскедастичностью* (что означает одинаковый разброс). *Зависимость* дисперсии случайной ошибки от номера наблюдения называется *гетероскедастичностью*. Таким образом,

- $D(\varepsilon_i)=\sigma^2$ $(i=1,2,...,n)$ – гомоскедастичность,
- $D(\varepsilon_i)=\sigma_i^2$ $(i=1,2,...,n)$ – гетероскедастичность.

Характерные диаграммы рассеяния гомоскедастичности и гетероскедастичности показаны на рис. 3, *а* и *б* соответственно.



а) *б*)

Рис. 3

Если условие гомоскедастичности не выполняется, то оценки коэффициентов регрессии будут *неэффективными*, хотя и *несмещенными*. Существуют специальные методы диагностирования и устранения гетероскедастичности.

Третье условие указывает на некоррелированность случайных отклонений для разных наблюдений. Это условие часто нарушается, когда данные являются временными рядами. В случае когда третье условие не выполняется, говорят об *автокорреляции остатков*. Типичный вид данных при наличии автокорреляции показан на рис. 4. Если условие независимости случайных ошибок не выполняется, то оценки коэффициентов регрессии, полученные по МНК, оказываются *неэффективными*, хотя и *несмещенными*. Существуют методы диагностирования и устранения автокорреляции.

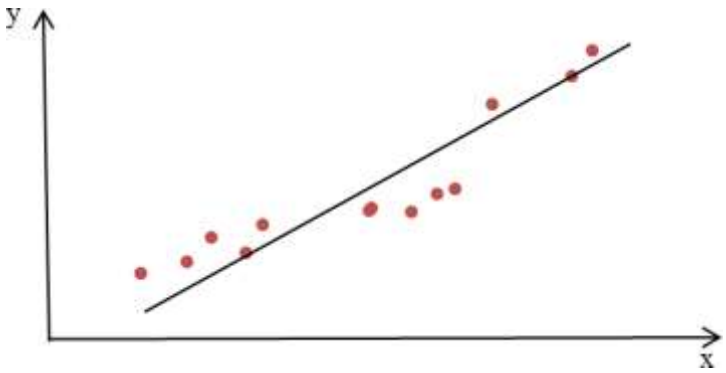


Рис. 4

Четвертое условие является особенно важным. Если условие о неслучайности объясняющей переменной не выполняется, то оценки коэффициентов регрессии будут *смещенными* и *несостоятельными.* Нарушение этого условия может быть связано с ошибками измерения объясняющих переменных или с использованием лаговых переменных.

Предположение о нормальности распределения случайной ошибки необходимо для проверки значимости параметров регрессии и для интервального оценивания.

Теорема Гаусса-Маркова. Если условия 1 – 4 регрессионного анализа выполняются, то оценки b_0 и b_1, сделанные с помощью МНК, являются наилучшими линейными несмещенными оценками, т.е. обладают следующими свойствами:

1) несмещенности: $M(b_0) = \beta_0$, $M(b_1) = \beta_1$,

(означает отсутствие систематической ошибки в положении линии регрессии);

2) *эффективности*: имеют наименьшую дисперсию в классе всех линейных несмещенных оценок, равную

$$D(b_0) = \frac{\overline{x^2} \cdot \sigma^2}{n \cdot S_x^2}, \quad D(b_1) = \frac{\sigma^2}{n \cdot S_x^2};$$

3) *состоятельности*: $\lim\limits_{n\to\infty} D(b_0) = 0$, $\lim\limits_{n\to\infty} D(b_1) = 0$.

1.4. Расчет стандартных ошибок коэффициентов регрессии

Теоретические дисперсии $D(b_0)$, $D(b_1)$ зависят от дисперсии σ^2 случайной ошибки. По данным выборки отклонения ε_i, а, следовательно, и их дисперсия σ^2 неизвестны, поэтому они заменяются наблюдаемыми остатками e_i и их выборочной дисперсией. Несмещенной оценкой дисперсии σ^2 является величина (остаточная дисперсия) $S^2 = \frac{\sum e_i^2}{n-2}$, которая служит мерой разброса зависимой переменной вокруг линии регрессии. Величина $S = \sqrt{S^2}$ называется *стандартной ошибкой регрессии*. Заменив в теоретических дисперсиях неизвестную σ^2 на оценку S^2, получим оценки дисперсии:

$$S_{b_0}^2 = \frac{\overline{x^2} \cdot S^2}{n \cdot \left(\overline{x^2} - \bar{x}^2\right)}, \quad S_{b_1}^2 = \frac{S^2}{n \cdot \left(\overline{x^2} - \bar{x}^2\right)}. \qquad (1.5)$$

Величины $S_{b_0} = \sqrt{S_{b_0}^2}$, $S_{b_1} = \sqrt{S_{b_1}^2}$ есть стандартные отклонения случайных величин b_0 и b_1, называемые *стандартными ошибками коэффициентов регрессии*. Объяснение данных соотношений имеет наглядную графическую интерпретацию. В знаменателях дробей (1.5) стоит сумма $\sum(x_i - \bar{x})_2$ квадратов отклонений x_i от среднего значения $\bar{x}$. Эта сумма велика (следовательно, вся дробь мала, и выборочные дисперсии оценок меньше), если регрессия определяется на широком диапазоне значений переменной X. Например, на рис. 5 через пары точек (1, 3) и (2, 3) проведена одна и та же прямая. Но диапазон (1, 3) шире диапазона (2,3). Если вместо точки 3 рассмотреть либо точку 3а, либо 3б (т.е. при случайном изменении выборки), то наклон прямой для пары (1, 3) изменится значительно меньше, чем для пары (2, 3).

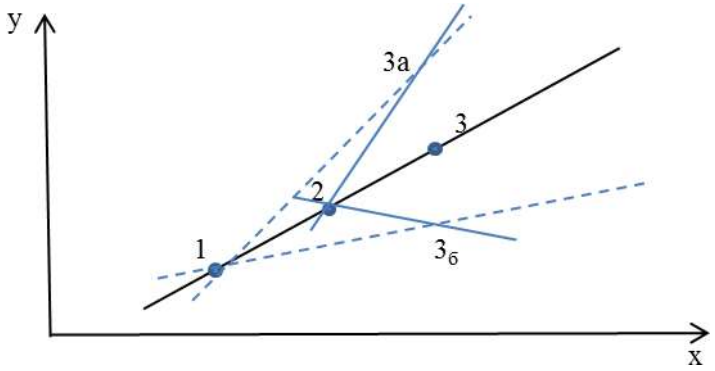


Рис. 5

Выборочная дисперсия свободного члена уравнения регрессии $S_{b_0}^2 = S_{b_1}^2 \cdot \frac{\sum x_i^2}{n}$ пропорциональна дисперсии $S_{b_1}^2$. Действительно, чем сильнее меняется наклон прямой, проведенной через данную точку $(\bar{x}, \bar{y})$, тем больше разброс значений свободного члена, характеризующего точку пересечения этой прямой с осью OY.

Кроме того, разброс значений свободного члена тем больше, чем больше средняя величина $\overline{x^2}$. Это связано с тем, что при больших по модулю значениях X даже небольшое изменение наклона регрессионной прямой может вызвать большое изменение оценки свободного члена, поскольку в этом случае в среднем велико расстояние от точек наблюдений до оси OY. На рис. 6 через пары точек (1, 2) и (3, 4) проходит одна и та же прямая, пересекающая ось OY в точке (0, b_0). Для второй из этих пар значения переменной X больше по абсолютной величине (при одинаковом диапазоне изменений X и Y), чем для первой. Если в этих парах точки 1 и 3 изменить на одну и ту же величину (новые точки 1_a, 3_a), то углы наклона новых прямых (1_a, 2) и (3_a, 4) будут одинаковы. Но свободный член b_{01} для первой прямой будет существенно меньше отличаться от b_0, чем свободный член b_{02} для второй прямой.

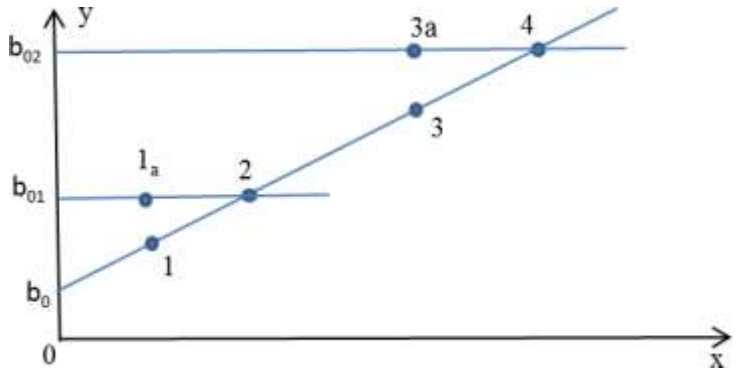


Рис. 6

1.5. Проверка гипотез относительно коэффициентов линейного уравнения регрессии

Эмпирическое уравнение регрессии определяется на основе конечного числа статистических данных. Поэтому коэффициенты эмпирического уравнения регрессии являются случайными величинами, изменяющимися от выборки к выборке. При проведении статистического анализа перед исследователем зачастую возникает необходимость сравнения теоретических значений β_0 и β_1 с некоторыми заданными значениями. Данный анализ осуществляется по схеме статистической проверки гипотез. Пусть по выборочным данным получена оценка b_1. Для проверки гипотезы

$$H_0 : \beta_1 = \beta_1^*,$$
$$H_1 : \beta_1 \neq \beta_1^*$$

используется статистика

$$t = \frac{b_1 - \beta_1^*}{S_{b_1}},$$

которая при справедливости H_0 имеет распределение Стьюдента с $\nu = n - 2$ степенями свободы. По выборке вычисляется значение t-статистики – $t_{набл.}$. По таблице критических точек распределения Стьюдента по заданному уровню значимости α и числу степеней свободы $\nu = n - 2$ находят критическую точку $t_{кр.}$. Тогда

- если $|t_{набл.}| < t_{кр.}$, то нет оснований для отклонения H_0;
- если $|t_{набл.}| > t_{кр.}$, то H_0 отвергается в пользу H_1.

Замечание. В экономических исследованиях проверку гипотез осуществляют, как правило, при 5%-ном и 1%-ном уровнях значимости. Если нулевая гипотеза отклоняется при 1%-ном уровне значимости, то она автоматически отклоняется и при 5%-ном уровне, а если не отвергается при 5%-ном уровне, то не отвергается и при 1%-ном уровне. Если при 5%-ном уровне значимости гипотеза отклоняется, а при 1%-ном – не отвергается, то результаты проверки гипотезы приводятся при двух уровнях значимости.

Результаты оценивания регрессии совместимы не только с конкретной гипотезой $H_0 : \beta_1 = \beta_1^*$, но и с некоторым множеством

значений, совместимых с оценкой b_1. Точнее, значение β_1^* совместимо с b_1, если H_0 не отвергается, т.е. выполняется условие

$$\left|\frac{b_1-\beta_1^*}{S_{b_1}}\right| < t_{кр.}, \quad \text{или} \quad -t_{кр.} < \frac{b_1-\beta_1^*}{S_{b_1}} < t_{кр.}.$$

Разрешив это неравенство относительно β_1^* и учитывая, что гипотеза $H_0: \beta_1 = \beta_1^*$ не отвергается, получим *доверительный интервал* для неизвестного коэффициента β_1:

$$b_1 - t_{кр.} \cdot S_{b_1} < \beta_1 < b_1 + t_{кр.} \cdot S_{b_1}.$$

Посредине интервала лежит величина b_1. Границы интервала симметричны относительно b_1, зависят от выбора уровня значимости и являются случайными числами. Доверительный интервал покрывает значение β_1 с заданной вероятностью $(1-\alpha)$, т.е.

$$P(b_1 - t_{кр.} \cdot S_b < \beta_1 < b_1 + t_{кр.} \cdot S_b) = 1-\alpha.$$

Наиболее важной на начальном этапе статистического анализа построенной модели все же является задача установления наличия линейной зависимости между Y и X. Эта проблема может быть решена по той же схеме:

$$H_0: \beta_1 = 0,$$

$$H_1: \beta_1 \neq 0.$$

Гипотеза в такой постановке называется *гипотезой о статистической значимости коэффициента регрессии*. При этом, если H_0 не отвергается, то есть основания считать, что величина Y не зависит от X (точнее связь между этими двумя переменными далека от линейной зависимости). В этом случае говорят, что коэффициент b_1 *статистически незначим* (он слишком близок к нулю). При отклонении H_0 коэффициент b_1 считается *статистически значимым*, что указывает на наличие определенной линейной зависимости между Y и X. В данном случае рассматривается двусторонняя критическая область, т.к. важным является именно отличие от нуля коэффициента регрессии, и он может быть как положительным, так и отрицательным.

Поскольку в данном случае полагается, что $\beta_1 = 0$, то формально значимость оцененного коэффициента регрессии b_1 проверяется с помощью анализа отношения его величины к его стандартной ошибке $S_{b_1} = \sqrt{S_{b_1}^2}$. В случае выполнения исходных предпосылок модели эта дробь

имеет распределение Стьюдента с числом степеней свободы $\nu = n - 2$, где n – число наблюдений. Данное отношение называется t-статистикой:

$$t = \frac{b_1}{S_{b_1}} = \frac{b_1}{\sqrt{S^2_{b_1}}}.$$

Для t-статистики проверяется нулевая гипотеза о равенстве ее нулю. Очевидно, что $t = 0$ равнозначно $b_1 = 0$. Фактически это свидетельствует об отсутствии линейной связи между Y и X. По аналогичной схеме на основе t-статистики проверяется *гипотеза о статистической значимости коэффициента* b_0:

$$t = \frac{b_0}{S_{b_0}} = \frac{b_0}{\sqrt{S^2_{b_0}}}.$$

Отметим, что для парной регрессии более важным является анализ статистической значимости коэффициента b_1, т.к. именно в нем скрыто влияние объясняющей переменной X на зависимую переменную Y.

1.6. Проверка общего качества уравнения регрессии. Коэффициент детерминации.

После проверки значимости каждого коэффициента регрессии обычно проверяется общее качество уравнения, которое оценивается по тому, как хорошо эмпирическое уравнение регрессии согласуется со статистическими данными. Другими словами, насколько широко рассеяны точки наблюдений относительно линии регрессии.

Пусть на основе выборочных наблюдений построено уравнение регрессии $\hat{y}$, тогда значение зависимой переменной y в каждом наблюдении можно разложить на две составляющие:

$$y_i = \hat{y}_i + e_i,$$

где остаток e_i – та часть зависимой переменной y, которую невозможно объяснить с помощью уравнения регрессии. Разброс значений зависимой переменной характеризуется выборочной дисперсией S_y^2. Разложим дисперсию S_y^2:

$$S_y^2 = S_{\hat{y}+e}^2 = S_{\hat{y}}^2 + S_e^2 + 2\mathrm{cov}(\hat{y}, e),$$

где $\mathrm{cov}(\hat{y}, e)$ – выборочная ковариация переменных $\hat{y}$ и e. Поскольку $\mathrm{cov}(\hat{y}, e) = 0$, то $S_y^2 = = S_{\hat{y}}^2 + S_e^2$ (такое разложение возможно, если только константа b_0 включена в уравнение регрессии). Таким образом, дисперсия S_y^2 разложена на две части:

- $S_{\hat{y}}^2$ – часть, объясненная регрессионным уравнением;
- S_e^2 – необъясненная часть.

Коэффициентом детерминации R^2 называется отношение

$$R^2 = \frac{S_{\hat{y}}^2}{S_y^2} = 1 - \frac{S_e^2}{S_y^2},$$

характеризующее долю вариации (разброса) зависимой переменной, *объясненную* с помощью уравнения регрессии. Отношение $\frac{S_e^2}{S_y^2}$ есть доля *необъясненной* дисперсии. Если $R^2 = 1$, то подгонка точная: $S_y^2 = = S_{\hat{y}}^2$, $S_e^2 = 0$, $y_i = \hat{y}_i$, $i = 1,2,...,n$, т.е. все точки наблюдения лежат на регрессионной прямой. Если $R^2 = 0$, то регрессия ничего не дает (т.е. переменная x не улучшает качества предсказания y по сравнению с горизонтальной прямой $\hat{y} = \bar{y}$):

$$S_y^2 = S_e^2, \quad S_{\hat{y}}^2 = 0, \quad \hat{y_i} = \bar{y}, \quad (i = 1,2,...,n).$$

Чем ближе к единице R^2, тем лучше качество подгонки, т.е. $\hat{y}$ более точно аппроксимирует y.

Как правило $0 \le R^2 \le 1$, но в исключительных случаях возможно нарушение неравенства $0 \le R^2$. Последнее происходит обычно для линейных уравнений, в которых отсутствует свободный член b_0. Оценивая такое уравнение по МНК, мы вынуждены рассматривать лишь те прямые (гиперплоскости), которые проходят через начало координат. Значение R^2 получается отрицательным тогда, когда разброс значений зависимой переменной вокруг линии $Y = \bar{y}$ меньше, чем вокруг любой из прямых, проходящих через начало координат.

Для определения статистической значимости коэффициента детерминации R^2 проверяется гипотеза

$$H_0 : R_2 = 0,$$

$$H_1 : R_2 > 0.$$

Для проверки нулевой гипотезы используется статистика:

$$F = \frac{R_2(n-2)}{1-R^2},$$

которая при справедливости H_0 имеет распределение Фишера (F-распределение) с $\nu_1 = 1, \ \nu_2 = n-2$ степенями свободы. Вычисленный критерий F сравнивается с критическим значением $F_{\text{кр.}}$:

– если $F_{набл.} < F_{кр.}$, то нет оснований для отклонения H_0, т.е. R^2 статистически незначим (говорят еще, что уравнение регрессии незначимо в целом);

– если $F_{набл.} > F_{кр.}$, то нет оснований для отклонения H_0, т.е. R^2 статистически значим (уравнение регрессии значимо в целом).

Замечание. В случае парной регрессии коэффициент детерминации есть квадрат коэффициента корреляции переменных x и y: $R^2 = r_{x,y}^2$.

1.7. Доверительные интервалы для зависимой переменной

Одной из центральных задач моделирования является предсказание (прогнозирование) значений зависимой переменной при определенных значениях объясняющих переменных. Здесь возможен двоякий подход: 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 $\hat{y} = b_0 + b_1 x$, на основе которого необходимо предсказать условное математическое ожидание $M(Y|X = x_0)$ переменной Y при X = x_0. В данном случае значение $\hat{y}_0 = b_0 + b_1 x_0$ является оценкой $M(Y|X = x_0)$. Тогда естественным является вопрос, как сильно может уклониться модельное среднее значение $\hat{y}_0$, рассчитанное по эмпирическому уравнению регрессии, от соответствующего условного математического ожидания. Ответ на этот вопрос дается на основе интервальных оценок, построенных с заданной надежностью $(1-\alpha)$ при любом конкретном значении x_0 объясняющей переменной.

Доверительный интервал для $M(Y|X = x_0) = \beta_0 + \beta_1 x_0$ имеет вид:

$$\left(b_0 + b_1 x_0 - t_{\frac{\alpha}{2};n-2} \cdot S \cdot \sqrt{\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{n(\overline{x^2} - \bar{x}^2)}}; b_0 + b_1 x_0 + t_{\frac{\alpha}{2};n-2} \cdot S \cdot \sqrt{\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{n(\overline{x^2} - \bar{x}^2)}} \right) \qquad (1.6)$$

Предсказание индивидуальных значений зависимой переменной

На практике иногда более важно знать дисперсию Y, чем ее средние значения или доверительные интервалы для условных математических ожиданий.

Пусть нас интересует некоторое возможное значение y_0 переменной Y при определенном значении x_0 объясняющей переменной X. Предсказанное по уравнению регрессии значение Y при X = x_0 составляет $\hat{y}_0$. Тогда интервал

$$\left(\hat{y_0} - t_{\frac{\alpha}{2};n-2} \cdot S \cdot \sqrt{1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{n(\overline{x^2} - \bar{x}^2)}}; \hat{y_0} + t_{\frac{\alpha}{2};n-2} \cdot S \cdot \sqrt{1 + \frac{1}{n} + \frac{(\bar{x} - x_0)^2}{n(\overline{x^2} - \bar{x}^2)}} \right)$$

определяет границы, за пределами которых могут оказаться не более 100α% точек наблюдений при X = x_0. Заметим, что данный интервал шире доверительного интервала для условного математического ожидания (на рис. 6 границы этого интервала отмечены пунктирной линией).

Проводя анализ построенных интервалов, несложно заметить, что наиболее узкими они будут при X = x_0. По мере удаления x_0 от среднего значения доверительные интервалы расширяются (см. рис. 7). Поэтому необходимо достаточно осторожно экстраполировать полученные результаты на прогнозные области. С другой стороны, с ростом числа наблюдений n эти интервалы сужаются к линии регрессии при $n \to \infty$.

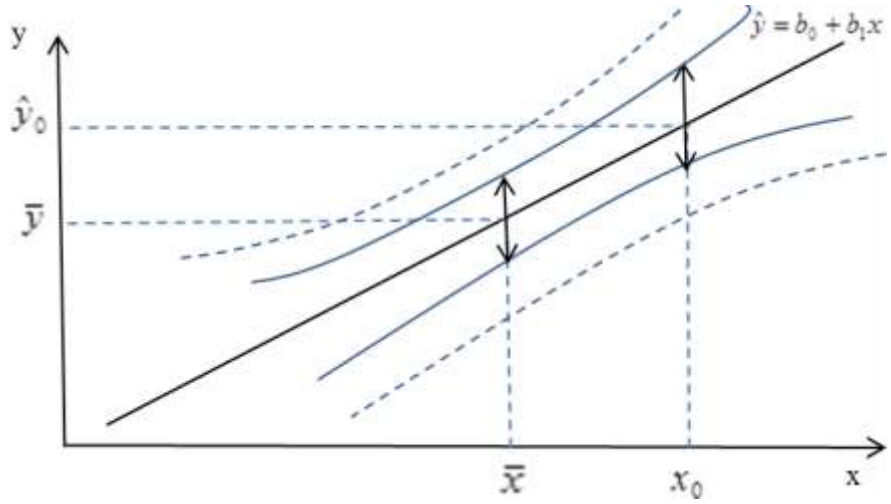


Рис. 7

1.8. Реализация типовых задач в MS Excel

Для расчета параметров уравнения линейной регрессии и проверки его адекватности исследуемому процессу Microsoft Excel располагает функцией *Регрессия*. Для вызова этой функции необходим пакет статистического анализа. *Пакет анализа* представляет собой надстройку, т.е. программу, которая доступна при установке Microsoft Office или Excel. Чтобы использовать эту надстройку, необходимо сначала загрузить ее. Для этого:

– на вкладке *Файл* выберите элемент *Параметры*, затем пункт *Надстройки*;

– нажмите кнопку *Перейти*;

– В окне *Доступные надстройки* установите флажок *Пакет анализа*, а затем нажмите кнопку *ОК*.

Для вызова функции *Регрессия* необходимо выбрать команду меню *Данные → Анализ данных*. На экране раскроется диалоговое окно *Анализ данных*, в котором следует выбрать значение *Регрессия*, в результате чего на экране появится диалоговое окно *Регрессия*, представленное на рис. 8.

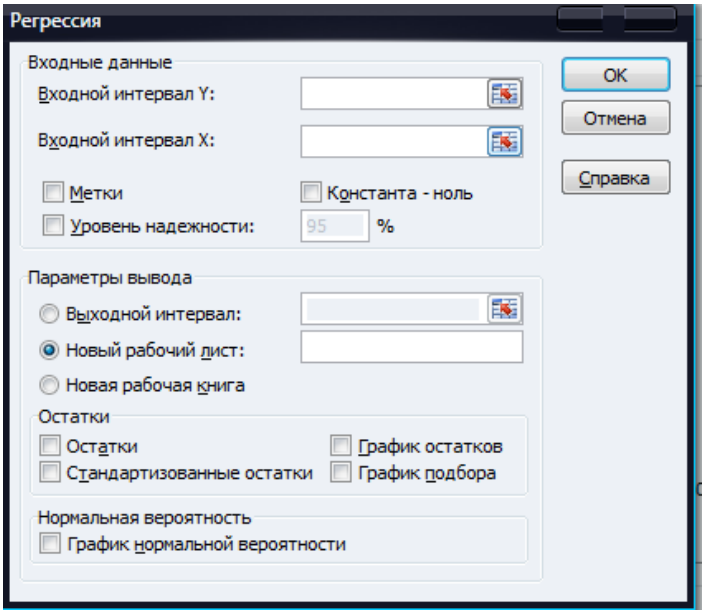


Рис. 8

В диалоговом окне *Регрессия* задаются следующие параметры.

1. В поле *Входной интервал Y* вводится диапазон ячеек, содержащих исходные данные по результативному признаку. Диапазон должен состоять из одного столбца.

2. В поле *Входной интервал X* вводится диапазон ячеек, содержащих исходные данные факторного признака. Максимальное число входных диапазонов (столбцов) равно 16.
3. Флажок *Метки* устанавливается в том случае, если первая строка во входном диапазоне содержит заголовок. Если заголовок отсутствует, этот флажок следует сбросить. В последнем случае для данных выходного диапазона будут автоматически созданы стандартные названия.
4. Флажок опции *Уровень надежности* устанавливается в том случае, если в расположенное рядом с флажком поле необходимо ввести уровень надежности, отличный от уровня 95%, применяемого по умолчанию. Установленный в данном поле уровень надежности используется для проверки значимости коэффициента детерминации и коэффициентов регрессии. Если данный флажок сброшен, в таблице параметров уравнения регрессии генерируются две одинаковые пары столбцов для границ доверительных интервалов.
5. Флажок *Константа-ноль* устанавливается в том случае, когда требуется, чтобы линия регрессии прошла через начало координат (т.е. $b_0 = 0$).
6. Переключатель в группе *Параметры вывода* может быть установлен в одно из трех положений, определяющих, где должны быть размещены результаты расчета: *Выходной интервал*, *Новый рабочий лист* или *Новая рабочая книга*.
7. Флажок опции *Остатки* устанавливается в том случае, если в диапазон ячеек с выходными данными требуется включить столбец остатков.
8. Флажок опции *Стандартизированные остатки* устанавливается в том случае, если в диапазон ячеек с выходными данными требуется включить столбец стандартизированных остатков.
9. Флажок опции *График остатков* должен быть установлен, если на рабочий лист требуется вывести графики зависимости остатков от факторных признаков x_i.
10. Флажок опции *График подбора* должен быть установлен, если на рабочий лист требуется вывести точечные графики зависимости теоретических результативных значений $\hat{y}$ от факторных признаков x_i.

11. Флажок опции *График нормальной вероятности* должен быть установлен, если на рабочий лист требуется вывести точечный график зависимости наблюдаемых значений y от автоматически формируемых интервалов персентелей.

Пример. Для исследования зависимости годового объема производства Y от основных фондов X получены данные по 20-ти предприятиям.

X	12,5	17,5	17,5	17,5	22,5	22,5	22,5	22,5	22,5	27,5	27,5
Y	20,5	21,5	21,5	22,5	22,5	22,5	23,5	23,5	23,5	23,5	23,5

27,5	27,5	27,5	27,5	27,5	27,5	27,5	27,5	27,5
23,5	24,5	24,5	24,5	24,5	24,5	24,5	24,5	24,5

Результаты решения данной задачи с помощью функции Регрессия представлены на рис. 9 – 11.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,923584
R-квадрат	0,853006
Нормированный R-квадрат	0,84484
Стандартная ошибка	0,47647
Наблюдения	20

Рис. 9. Результаты расчета: регрессионная статистика

На рис. 9 представлены результаты расчета регрессионной статистики. Эти результаты соответствуют следующим статистическим показателям:

• Множественный R – коэффициент корреляции R;
• R-квадрат – коэффициент детерминации R^2;
• Нормированный R – нормированное значение коэффициента корреляции;
• Стандартная ошибка – стандартное отклонение для остатков;
• Наблюдения – число исходных наблюдений.

На рис. 10 представлены результаты расчета дисперсионного анализа, которые используются для проверки значимости коэффициента детерминации R^2.

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	23,71358	23,71358	104,4544	6,38E-09
Остаток	18	4,08642	0,227023		
Итого	19	27,8			

Рис. 10. Результаты расчета: дисперсионный анализ

Значения в столбцах на рис. 10 имеют следующую интерпретацию.

• Столбец *df* – число степеней свободы. Для строки *Регрессия* число степеней свободы определяется количеством факторных признаков *m*, для строки *Остаток* – числом наблюдений *n* и количеством переменных в уравнении регрессии *m* + 1: *n* – (*m* + 1), а для строки *Итого* – суммой степеней свободы для строк *Регрессия* и *Остаток* и, следовательно, равно *n* – 1.

• Столбец *SS* – сумма квадратов отклонений. Для строки *Регрессия* значение определяется как сумма квадратов отклонений расчетных данных от среднего:

$$SS_1^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 .$$

Для строки *Остаток* это сумма квадратов отклонений фактических данных от теоретических:

$$SS_2^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

Для строки *Итого* это сумма квадратов отклонений расчетных данных от среднего:

$$SS_3^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{или} \quad SS_2^2 = SS_1^2 + SS_2^2 .$$

• Столбец *MS* содержит значения дисперсии, которые рассчитываются по формуле:

$$MS = \frac{SS}{df} .$$

Для строки *Регрессия* это факторная дисперсия $\sigma_{\phi}^2 = \dfrac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{n}$.

Для строки *Остаток* это остаточная дисперсия $\sigma_{ост.}^2 = \dfrac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - (m+1)}$.

• Столбец F содержит расчетное значение F-критерия Фишера F_p, вычисляемое по формуле:

$$F_p = \frac{MS(Регрессия)}{MS(Остатки)}.$$

• Столбец *Значимость F* содержит значение уровня значимости, соответствующее вычисленному значению F_p.

На рис. 11 представлены полученные значения коэффициентов регрессии b_i и их статистические оценки.

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	17,593	0,578	30,430	6,23E-17	16,378	18,807
X	0,242	0,024	10,220	6,38E-09	0,192	0,292

Рис. 11. Результаты расчета: коэффициенты уравнения регрессии и их статистические оценки

Столбцы на рис. 11 содержат следующие значения:

• Коэффициенты – значение коэффициентов b_i.

• Стандартная ошибка – стандартные ошибки коэффициентов b_i.

• t-статистика – расчетные значения t-критерия, вычисляемые по формуле:

$$t-статистика = \frac{коэффициенты}{стандартная\ ошибка}.$$

• P-значение – значения уровней значимости, соответствующие вычисленным значениям t_p.

• Нижние 95% и Верхние 95% – нижние и верхние границы доверительных интервалов для коэффициентов регрессии b_i.

Переходя к анализу полученных расчетных данных, можно построить уравнение регрессии с вычисленными коэффициентами, которое будет выражать зависимость годового объема производства от основных фондов:

$$\hat{y} = 17{,}593 + 0{,}242x.$$

Выборочный коэффициент детерминации $R^2 = 0{,}853$ (рис. 9) показывает, что 85,3% разброса зависимой переменной y объясняется построенной регрессией $\hat{y}$. Рассчитанный уровень значимости (показатель *Значимость F* рис. 10) $\alpha_p = 6{,}38 \cdot 10^{-9} < 0{,}05$ подтверждает

статистическую значимость величины R^2 (т.е. гипотеза $H_0 : R^2 = 0$ отвергается в пользу $H_1 : R_2 > 0$ при уровне значимости $\alpha = 0{,}05$). В этом случае говорят еще, что уравнение регрессии значимо в целом при $\alpha = 0{,}05$.

Следующим этапом является проверка значимости коэффициентов регрессии b_0 и b_1. При парном сравнении коэффициентов и их стандартных ошибок (см. рис. 11) можно сделать вывод, что вычисленные коэффициенты являются статистически значимыми (т.е. гипотезы $H_0 : \beta_0 = 0$ и $H_0 : \beta_1 = 0$ отвергаются). Этот вывод подтверждается величинами *P*-значений коэффициентов, которые меньше уровня значимости $\alpha = 0{,}05$. Доверительные интервалы с уровнем надежности $\gamma = 1 - \alpha = 1 - 0{,}05 = 0{,}95$ для теоретических коэффициентов β_0 и β_1 равны соответственно (16,378; 18,807) и (0,192; 0,292). Последнее означает, что, основываясь на выборочных данных, можно утверждать о попадании неизвестных параметров β_0 и β_1 в указанные интервалы с вероятностью 0,95. Заметим также, что значение 0 не принадлежит никакому из этих интервалов, откуда можно сделать вывод о том, что гипотезы $H_0 : \beta_0 = 0$ и $H_0 : \beta_1 = 0$ отвергаются при уровне значимости $\alpha = 0{,}05$, как и было сказано выше.

Проверка значимости коэффициента детерминации R^2 и коэффициентов регрессии b_0 и b_1 при факторном признаке подтверждает адекватность полученного уравнения.

Дадим экономическую интерпретацию. Коэффициент регрессии $b_1 = 0{,}242$ показывает, что при увеличении размера основных фондов на 1 у.е., годовой объем производства возрастает в среднем на 0,242 у.е. Коэффициент регрессии $b_0 = 17{,}593$ означает, что при нулевом размере основных фондов годовой объем производства ожидается (в среднем) на уровне 17,593 у.е.

Замечание. К экономической интерпретации коэффициента b_0 следует относиться с известной долей осторожности, сообразуясь со здравым смыслом, поскольку выборочные данные находятся достаточно далеко от нуля. В ряде случаев ограничиваются интерпретацией коэффициента при объясняющей переменной.

Дадим точечный и интервальный прогнозы среднего размера годового объема производства при размере основных фондов 25 у.е.

Подставив в выборочное уравнение регрессии значение $x = 25$, получим точечный прогноз:

$$\hat{y}(x = 25) = 17{,}593 + 0{,}242 \cdot 25 = 26{,}642 .$$

Таким образом, при размере основных фондов на уровне 25 у.е., годовой объем производства ожидается (в среднем) на уровне 26,642 у.е.

Для построения доверительного интервала для прогнозного среднего значения воспользуемся формулой (1.6):

$$\left(\hat{y}(x = x_0) - t_{\frac{\alpha}{2};n-2} \cdot S \cdot \sqrt{\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{n(\overline{x^2} - \bar{x}^2)}} ; \hat{y}(x = x_0) + t_{\frac{\alpha}{2};n-2} \cdot S \cdot \sqrt{\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{n(\overline{x^2} - \bar{x}^2)}} \right). \qquad (1.6)$$

Имеем: $n = 20$; $\hat{y}(x = 25) = 26{,}642$; $S = 0{,}476$ (рис. 8); $\bar{x} = \frac{1}{n}\sum_{1=1}^{20} x_i = 24$;

$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{20} x_i^2 = 596{,}25$; $t_{\frac{\alpha}{2};n-2} = t_{\frac{0{,}05}{2};20-2} = 2{,}12$; (из таблиц критических точек распределения Стьюдента или Excel – f_x – статистические – Стьюдент.обр.2х). Подставив полученные значения в формулу (1.6), получим 95%-ный доверительный интервал для прогнозного среднего значения результативного признака Y при $X = 25$: (26,642 – 0,231; 26,642 + 0,231), откуда находим, что в интервал (26,411; 26,873) среднее значение годового объема производства при размере основных фондов, равным 25 у.е., попадает с вероятностью 0,95 (если ориентироваться на выборочные данные).

2. Множественная линейная регрессия

2.1. Определение параметров уравнения регрессии

На любой экономический показатель практически всегда оказывают влияние не один, а несколько факторов. Например, спрос на некоторое благо определяется не только ценой данного блага, но и ценами на заменяющие и дополняющие блага, доходом потребителей и многими другими факторами. В этом случае вместо парной регрессии $M(Y|x) = f(x)$ рассматривается *множественная регрессия*

$$M(Y \mid x_1, x_2, ..., x_m) = f(x_1, x_2, ..., x_m).$$

Задача оценки статистической взаимосвязи переменных Y и X_1, X_2, …, X_m формулируется аналогично случаю парной регрессии. *Уравнение множественной регрессии* может быть представлено в виде

$$Y = f(\beta, X) + \varepsilon,$$

где X = (X_1, X_2, …, X_m) – вектор *независимых (объясняющих) переменных*; β – вектор *параметров* (подлежащих определению); ε – *случайная ошибка (отклонение)*; Y – *зависимая (объясняемая) переменная*.

Рассмотрим самую употребляемую и наиболее простую из моделей множественной регрессии – модель множественной линейной регрессии. *Теоретическое линейное уравнение* регрессии имеет вид:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m + \varepsilon$$

или для индивидуальных наблюдений i, $i = 1,2,\ldots,n$:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_m x_{im} + \varepsilon_i.$$

Здесь $\beta = (\beta_0, \beta_1, ...\beta_m)$ – вектор размерности $(m+1)$ неизвестных параметров. Коэффициент β_j, j=1,2,…,m, называется j-м теоретическим коэффициентом регрессии. Он характеризует чувствительность величины Y к изменению X_j. Другими словами, он отражает влияние на условное математическое ожидание $M(Y \mid x_1, x_2, ..., x_m)$ зависимой переменной Y объясняющей переменной X_j при условии, что все другие объясняющие переменные остаются постоянными. Коэффициент β_0 определяет значение Y, в случае, когда все объясняющие переменные X_j равны нулю. Случайная ошибка ε удовлетворяет тем же предпосылкам, что и в модели с парной регрессией. Предполагается, что объясняющие переменные *некоррелированы* друг с другом (в модели отсутствует мультиколлинеарность).

На основе n наблюдений оценивается выборочное уравнение регрессии

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_m x_m ,$$

где $b_0, b_1, \ldots, b_m$ – оценки параметров $\beta_0, \beta_1, \ldots, \beta_m$.

Для оценки параметров регрессии используется метод наименьших квадратов. В соответствии с МНК минимизируется сумма квадратов остатков:

$$F(b_0, b_1, b_2, \ldots, b_m) = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_m x_{im})^2 \to \min .$$

Необходимым условием ее минимума является равенство нулю всех ее частных производных. В результате приходим к системе из $(m + 1)$ линейного уравнения с $(m + 1)$ неизвестными, называемой *системой нормальных уравнений*. Ее решение в явном виде обычно записывается в матричной форме, иначе оно становится слишком громоздким. Оценки параметров модели и их теоретические дисперсии в матричной форме определяются выражениями

$$b = (X^T X)^{-1} X^T Y , \qquad D(b_i) = (X^T X)^{-1}_{ii} \cdot \sigma^2 ,$$

где b – вектор с компонентами $b_0, b_1, \ldots, b_m$; X – матрица значений объясняющих переменных; Y – вектор значений зависимой переменной; σ^2 – дисперсия случайной ошибки. Несмещенной оценкой σ^2 является величина S^2 (остаточная дисперсия):

$$S^2 = \frac{1}{n - m - 1} \sum e_i^2 .$$

Величина S называется стандартной ошибкой регрессии. Заменяя в теоретических дисперсиях неизвестную дисперсию σ^2 на ее оценку S^2 и извлекая квадратный корень, получим стандартные ошибки оценок коэффициентов регрессии

$$S_{b_i} = S \sqrt{(X^T X)^{-1}_{ii}} .$$

Если предпосылки относительно случайной ошибки выполняются, оценки параметров множественной регрессии являются *несмещенными, состоятельными и эффективными*.

2.2. Интервальные оценки коэффициентов теоретического уравнения регрессии и зависимой переменной

По аналогии с парной регрессией после определения точечных оценок b_i коэффициентов β_i (i = 0, 1, 2,…, m) теоретического уравнения регрессии могут быть рассчитаны интервальные оценки указанных коэффициентов. Для построения интервальной оценки коэффициента β_i строится t-статистика $t = \dfrac{b_i - \beta_i}{S_{b_i}}$, имеющая распределение Стьюдента с числом степеней свободы $\nu = n - m - 1$ (n – объем выборки, m – количество объясняющих переменных в модели). Пусть необходимо построить 100(1 – α)%-ный доверительный интервал для коэффициента β_i. Тогда по таблице критических точек распределения Стьюдента по требуемому уровню значимости α и числу степеней свободы ν находят критическую точку $t_{\frac{\alpha}{2};n-m-1}$, удовлетворяющую условию

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

Подставив в это условие t-статистику $t = \dfrac{b_i - \beta_i}{S_{b_i}}$, получим

$$P\left(-t_{\frac{\alpha}{2};n-m-1} < \frac{b_i - \beta_i}{S_{b_i}} < t_{\frac{\alpha}{2};n-m-1}\right) = 1-\alpha ,$$

или после преобразования

$$P\left(b_i - t_{\frac{\alpha}{2};n-m-1} < \beta_i < b_i + t_{\frac{\alpha}{2};n-m-1}\right) = 1-\alpha .$$

Таким образом, доверительный интервал, накрывающий с надежностью (1 – α) неизвестное значение параметра β_i, определяется неравенством

$$b_i - t_{\frac{\alpha}{2};n-m-1} \cdot S_{b_i} < \beta_i < b_i + t_{\frac{\alpha}{2};n-m-1} \cdot S_{b_i} .$$

Не вдаваясь в детали, отметим, что по аналогии с парной регрессией может быть построена интервальная оценка для среднего предсказания

$$\hat{Y_0} - t_{\frac{\alpha}{2};n-m-1} \cdot S\left(\hat{Y}_0\right) < M\left(Y \middle| X_0^T = (1, x_{01}, x_{02}, ..., x_{0m})^T\right) < \hat{Y}_0 + t_{\frac{\alpha}{2};n-m-1} \cdot S\left(\hat{Y_0}\right), \qquad (2.1)$$

где $\hat{Y_0} = b_0 + b_1 x_{01} + b_2 x_{02} + ... + b_m x_{0m}$; $S\left(\hat{Y_0}\right) = S\sqrt{X_0^T \left(X^T X\right)^{-1} X_0}$.

2.3. Анализ качества выборочного уравнения множественной линейной регрессии

Как и в случае парной линейной регрессии, статистическая значимость коэффициентов множественной линейной регрессии с m объясняющими переменными проверяется на основе t-статистики $t = \frac{b_i}{S_{b_i}}$, имеющей в данной ситуации распределение Стьюдента с числом степеней свободы $\nu = n - m - 1$ (n – объем выборки). При требуемом уровне значимости α наблюдаемое значение t-статистики сравнивается с критической точкой $t_{\frac{\alpha}{2};n-m-1}$ распределения Стьюдента. Если $|t| > t_{\frac{\alpha}{2};n-m-1}$, то коэффициент b_i считается статистически значимым. В противном случае ($|t| < t_{\frac{\alpha}{2};n-m-1}$) коэффициент b_i считается статистически незначимым (статистически близким к нулю). Это означает, что фактор X_i фактически линейно не связан с зависимой переменной Y. Его наличие среди объясняющих переменных не оправдано со статистической точки зрения. Последовательный отсев несущественных факторов составляет основу многошагового регрессионного анализа. Однако по коэффициентам регрессии нельзя определить, какой из факторов оказывает наибольшее влияние на зависимую переменную, так как коэффициенты регрессии между собой несопоставимы (они измерены разными единицами). Различия в единицах измерения факторов устраняют с помощью *частных коэффициентов эластичности,* рассчитываемых по формуле

$$Э_i = b_i \frac{\bar{x}_i}{\bar{y}}, \qquad (2.2)$$

где x_i – среднее значение фактора. Частные коэффициенты эластичности показывают, на сколько процентов в среднем изменяется зависимая переменная с изменением на 1% каждого фактора при фиксированном значении других факторов.

После проверки значимости каждого коэффициента регрессии обычно проверяется общее качество уравнения регрессии. Для этой цели, как и в случае парной регрессии, используется коэффициент детерминации R^2, который в общем случае рассчитывается по формуле

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

Суть данного коэффициента – доля общего разброса значений зависимой переменной, объясненного уравнением регрессии. Как отмечалось ранее $0 \le R^2 \le 1$. Чем ближе этот коэффициент к единице, тем больше уравнение регрессии объясняет поведение Y. Поэтому естественно желание построить регрессию с наибольшим R^2 . Для множественной регрессии коэффициент детерминации является неубывающей функцией числа объясняющих переменных. Добавление новой объясняющей переменной никогда не уменьшает значение R^2 . Для компенсации такого увеличения R^2 вводится *скорректированный коэффициент детерминации* с поправкой на число степеней свободы:

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

Если увеличение доли объясненной регрессии при добавлении переменной мало, то скорректированный коэффициент детерминации может уменьшиться, следовательно, добавлять переменную нецелесообразно. Доказано, что $\overline{R}^2$ увеличивается при добавлении новой объясняющей переменной тогда и только тогда, когда t-статистика для этой переменной по модулю больше единицы.

Для определения статистической значимости коэффициента детерминации R^2 проверяется гипотеза

$$H_0 : R_2 = 0,$$

$$H_1 : R_2 > 0.$$

Для проверки данной гипотезы используется статистика:

$$F = \frac{R^2\left(n-2\right)}{1-R^2},$$

которая, при справедливости H_0 и выполнении предпосылок МНК, имеет распределение Фишера (F-распределение) с $\nu_1 = m$, $\nu_2 = n - m - 1$ степенями свободы. Так же, как и в случае парной регрессии, вычисленный критерий F сравнивается с критическим значением $F_{\text{кр.}}$.

2.4. Мультиколлинеарность

Мультиколлинеарность это коррелированность двух или нескольких объясняющих переменных в уравнении регрессии. При наличии мультиколлинеарности МНК-оценки формально существуют, но обладают рядом недостатков:

1) небольшое изменение исходных данных приводит к существенному изменению оценок регрессии;
2) оценки, как правило, имеют большие стандартные ошибки, малую значимость, в то время как модель в целом является значимой (высокое значение R^2).

Если при оценке уравнения регрессии несколько факторов оказались незначимыми, то нужно выяснить, нет ли среди них сильно коррелированных между собой. При наличии мультиколлинеарности для ее устранения или уменьшения имеется ряд методов, в частности пошаговые процедуры отбора наиболее информативных переменных. Например, на первом шаге рассматривается лишь одна объясняющая переменная, имеющая с зависимой переменной Y наибольший коэффициент детерминации. На втором шаге включается в регрессию новая объясняющая переменная, которая вместе с первоначально отобранной образует пару объясняющих переменных, имеющую с Y наиболее высокий (скорректированный) коэффициент детерминации. На третьем шаге вводится в регрессию еще одна объясняющая переменная, которая вместе с двумя первоначально отобранными образует тройку объясняющих переменных, имеющую с Y наибольший (скорректированный) коэффициент детерминации, и т.д. Процедура введения новых переменных продолжается до тех пор, пока будет увеличиваться соответствующий (скорректированный) коэффициент детерминации $\overline{R}^2$. В большинстве случаев получаемые с помощью пошаговой процедуры наборы переменных оказываются оптимальными или близкими к оптимальным.

Пример. Используя данные Федеральной службы государственной статистики России (за двенадцать месяцев) требуется:

1) Оценить влияние факторов (X_k, $\forall k = \overline{1,6}$) на изучаемый показатель (Y) и друг на друга с помощью коэффициентов линейной корреляции.

2) Используя процедуру выбора факторов, предложить и построить подходящую линейную регрессионную модель изучаемого показателя.
3) Дать экономическую интерпретацию с использованием коэффициентов эластичности. Получить точечные и интервальные прогнозы изучаемого показателя на следующий месяц.

В % к предыдущему периоду	Оборот розничной торговли непродовольственными товарами	Располагаемые денежные доходы	Реальная заработная плата	Индексы цен товаров и услуг населению	Индексы цен продовольственных товаров	Индексы цен непродовольственных товаров	Индексы цен платных услуг населению
	Y	X1	X2	X3	X4	X5	X6
Июнь	100,6	107,4	106,5	100,6	100,7	100,3	100,9
Июль	102,9	100,3	99,5	100,5	100,3	100,4	100,9
Август	104,4	97	100,6	99,9	99	100,5	100,8
Сентябрь	101,3	106,8	102,2	100,3	99,3	101,1	100,9
Октябрь	103,8	99,1	98,2	100,6	100,4	100,7	100,7
Ноябрь	100,8	101,8	101,8	100,7	100,9	100,6	100,6
Декабрь	117,3	142	125,7	100,8	101,1	100,5	100,8
Январь 2006г.	75,2	52,1	76,8	102,4	102,0	100,4	106,2
Февраль	100,4	121,9	101	101,7	103,0	100,5	101,0
Март	109,9	108,9	106,2	100,8	101,2	100,4	100,7
Апрель	103	105,5	99	100,4	100,3	100,3	100,6
Май	100	99,1	103,7	100,5	100,5	100,4	100,6

Построим линейную регрессионную модель с использованием всех шести объясняющих переменных с помощью функции *Регрессия* (заметим, что во входной интервал X следует вводить сразу весь набор значений объясняющих переменных):

ВЫВОД ИТОГОВ

Регрессионная статистика

Множественный R	0,964746
R-квадрат	0,930735
Нормированный R-квадрат	0,847618
Стандартная ошибка	3,785404
Наблюдения	12

Дисперсионный анализ

	df	*SS*
Регрессия	6	962,7403
Остаток	5	71,6464
Итого	11	1034,387

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	635,5103686	684,2110492	0,928822137	0,395611	-1123,31	2394,331
X1	0,131664673	0,239250121	0,550322286	0,605779	-0,48335	0,746677
X2	0,4591524	0,377586527	1,216019024	0,27825	-0,51146	1,429769
X3	53,11332128	67,25470889	0,789733866	0,465468	-119,77	225,9971
X4	-24,01925153	28,81047116	-0,83369867	0,442444	-98,078	50,04042
X5	-21,09145229	24,61310441	-0,85691962	0,430635	-84,361	42,17855
X6	-13,901573	15,58480592	-0,89199525	0,413259	-53,963	26,16045

Анализируя выходные данные, приходим к выводу, что все коэффициенты регрессии незначимы при уровне значимости 0,05 (все *P*-значения больше 0,05). С другой стороны, высокое значение R^2 и значимость уравнения в целом (*F*-значение, равное 0,008986, меньше 0,05), указывают на то, что в модели присутствуют значимые переменные.

1) Для отбора факторов в модель регрессии и оценки их мультиколлинеарности, найдем матрицу парных коэффициентов корреляции. Расчет корреляционной матрицы предусмотрен функцией *Корреляция* в пакете *Анализ данных*. Для вызова функции *Корреляция* необходимо выбрать команду меню *Данные* → *Анализ данных*. На экране раскроется диалоговое окно *Анализ данных*, в котором следует выбрать значение *Корреляция*. Тогда на экране появится диалоговое окно *Корреляция*, представленное на рис. 12.

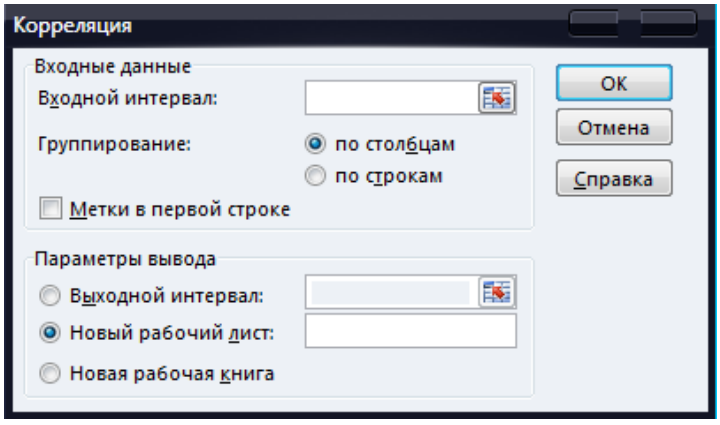


Рис.12

Во входной интервал вводим числовые данные всех переменных модели вместе с буквами, не забыв поставить флажок на метку. Задав выходной интервал (или оставив по умолчанию новый рабочий лист), получим матрицу парных коэффициентов корреляции:

	Y	*X1*	*X2*	*X3*	*X4*	*X5*	*X6*
Y	1						
X1	0,896244	1					
X2	0,913353	0,929783	1				
X3	-0,65744	-0,41067	-0,49785	1			
X4	-0,2807	0,014257	-0,1398	0,852351	1		
X5	0,079024	0,074599	0,045947	-0,21036	0,38206	1	
X6	-0,8566	-0,76876	-0,72354	0,801879	0,38871	0,12163	1

Анализируя вышеуказанную матрицу, замечаем, что наиболее существенное влияние на фактор Y оказывают переменные X2 ($r_{Y,X2} \approx 0{,}91$), X1 ($r_{Y,X1} \approx 0{,}896$), X6 ($r_{Y,X6} \approx -0{,}857$). Кроме этого, существует тесная корреляционная связь между переменными X1 и X2 ($r_{X1,X2} \approx 0{,}93$), X3 и X6 ($r_{X3,X6} \approx 0{,}802$), X3 и X4 ($r_{X3,X4} \approx 0{,}852$). Поэтому при построении регрессии с использованием всех объясняющих переменных будет иметь место мультиколлинеарность. Для устранения мультиколлинеарности применим процедуру пошагового отбора наиболее информативных переменных.

2) *1-й шаг.* Из объясняющих переменных X1 – X6 выделяется переменная X2, имеющая с зависимой переменной Y наибольший коэффициент детерминации $R^2_{Y,j}$ (равный для парной модели квадрату коэффициента корреляции $r^2_{Y,j}$). Воспользуемся функцией *Регрессия* для получения парной регрессии с участием переменных Y и X2. Ограничимся при этом выводом *Регрессионной статистики*:

ВЫВОД ИТОГОВ	
Регрессионная статистика	
Множественный R	0,913353
R-квадрат	0,834213
Нормированный R-квадрат	**0,817635**
Стандартная ошибка	4,141107
Наблюдения	12

Скорректированный коэффициент детерминации равен 0,818.

2-й шаг. Среди всевозможных пар объясняющих переменных X2, X*j*, *j* = 1, 3, 4, 5, 6, выбирается пара (X2, X6), имеющая с зависимой переменной Y наиболее высокий скорректированный коэффициент детерминации, равный 0,896. Результаты расчетов приводятся ниже.

X2,X1 ВЫВОД ИТОГОВ		X2,X3 ВЫВОД ИТОГОВ		X2,X4 ВЫВОД ИТОГОВ	
Регрессионная статистика		*Регрессионная статистика*		*Регрессионная статистика*	
Множественный R	0,922243	Множественный R	0,942792	Множественный R	0,926332
R-квадрат	0,850532	R-квадрат	0,888857	R-квадрат	0,858091
Нормированный R-квадрат	**0,817317**	Нормированный R-квадрат	**0,864158**	Нормированный R-квадрат	**0,826556**
Стандартная ошибка	4,144712	Стандартная ошибка	3,574056	Стандартная ошибка	4,038544
Наблюдения	12	Наблюдения	12	Наблюдения	12

X2,X6 ВЫВОД ИТОГОВ		X2,X5 ВЫВОД ИТОГОВ	
Регрессионная статистика		*Регрессионная статистика*	
Множественный R	0,956363	Множественный R	0,914106
R-квадрат	0,91463	R-квадрат	0,835589
Нормированный R-квадрат	**0,895659**	Нормированный R-квадрат	**0,799054**
Стандартная ошибка	3,13236	Стандартная ошибка	4,346955
Наблюдения	12	Наблюдения	12

3-й шаг. Среди всевозможных троек объясняющих переменных (X2, X6, X*j*), *j* = 1, 3, 4, 5, наиболее информативной оказалась тройка (X2, X6, X4), имеющая максимальный скорректированный коэффициент детерминации, равный 0,885. Результаты расчетов:

X1,X2,X6 ВЫВОД ИТОГОВ		X2,X3,X6 ВЫВОД ИТОГОВ	
Регрессионная статистика		*Регрессионная статистика*	
Множественный R	0,956621	Множественный R	0,957054
R-квадрат	0,915125	R-квадрат	0,915953
Нормированный R-квадрат	**0,883296**	Нормированный R-квадрат	**0,884435**
Стандартная ошибка	3,31274	Стандартная ошибка	3,296542
Наблюдения	12	Наблюдения	12

X2,X4,X6 ВЫВОД ИТОГОВ		X2,X5,X6 ВЫВОД ИТОГОВ	
Регрессионная статистика		*Регрессионная статистика*	
Множественный R	0,957151	Множественный R	0,956363
R-квадрат	0,916138	R-квадрат	0,914631
Нормированный R-квадрат	**0,884689**	Нормированный R-квадрат	**0,882618**
Стандартная ошибка	3,292908	Стандартная ошибка	3,322358
Наблюдения	12	Наблюдения	12

Так как скорректированный коэффициент детерминации на 3-м шаге не увеличился, то в регрессионной модели достаточно ограничиться лишь двумя отобранными ранее объясняющими переменными X2 и X6. Построим эту линейную регрессионную модель с помощью функции *Регрессия*:

ВЫВОД ИТОГОВ	
Регрессионная статистика	
Множественный R	0,956363154
R-квадрат	0,914630482
Нормированный R-квадрат	0,895659478
Стандартная ошибка	3,132359619
Наблюдения	12

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	946,0816	473,0408	48,21202	1,55E-05
Остаток	9	88,30509	9,811677		
Итого	11	1034,387			

	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%
Y-пересечение	302,8787507	98,34281602	3,079825888	0,013144	80,41184	525,3457
X2	0,556121622	0,127354024	4,36673772	0,001806	0,268027	0,844216
X6	-2,547239382	0,874833084	-2,91168615	0,017264	-4,52625	-0,56823

Оцененное уравнение имеет вид:

$$\hat{y} = 302{,}879 + 0{,}556x_2 - 2{,}547x_6 .$$

Нетрудно убедиться в том, что теперь все коэффициенты регрессии значимы при уровне значимости 0,05 (все *P*-значения меньше 0,05).

Кроме рассмотренной выше пошаговой процедуры *присоединения* объясняющих переменных используются также пошаговые процедуры *присоединения – удаления* и процедура *удаления* объясняющих переменных, изложенные, например, в [5]. Следует отметить, что какая бы пошаговая процедура ни использовалась, она не гарантирует определения оптимального (в смысле получения максимального коэффициента детерминации) набора объясняющих переменных. Однако в большинстве случаев получаемые с помощью пошаговых процедур наборы переменных оказываются оптимальными или близкими к оптимальным.

3) Дадим экономическую интерпретацию найденного уравнения с использованием коэффициентов эластичности.

Согласно формулам (2.2) $Э_i = b_i \frac{\overline{x}_i}{\overline{y}}$. По условию $\overline{x_2} = 101{,}767$, $\overline{x_6} = 101{,}223$, $\overline{y} = 101{,}663$. Тогда $Э_2 = 0{,}556 \cdot \frac{101{,}767}{101{,}663} = 0{,}557$; $Э_6 = -2{,}547 \cdot \frac{101{,}223}{101{,}663} = -2{,}536$.

Коэффициент $Э_2 = 0{,}557$ означает, что при увеличении реальной заработной платы на 1% оборот розничной торговли непродовольственными товарами вырастет в среднем на 0,557%. Коэффициент $Э_6 = -2{,}536$ означает, что при увеличении индексов цен платных услуг населению на 1% оборот розничной торговли непродовольственными товарами упадет в среднем на 2,536%.

Получим теперь точечные и интервальные прогнозы изучаемого показателя на следующий месяц, если реальная заработная плата в июне предполагается на уровне 108%, а индексы цен платных услуг населению на уровне 100,7% по отношению к майским показателям. Подставив указанные значения в полученное уравнение, получим точечную оценку среднего оборота розничной торговли непродовольственными товарами в июне:

$$\hat{y} = 302{,}879 + 0{,}556 \cdot 108 - 2{,}547 \cdot 100{,}7 = 106{,}44.$$

Для построения интервальной оценки изучаемого показателя воспользуемся формулой (2.1). Для наглядности приведем ряд промежуточных результатов:

$$X = \begin{pmatrix} 1 & 106,5 & 100,9 \\ 1 & 99,5 & 100,9 \\ 1 & 100,6 & 100,8 \\ 1 & 102,2 & 100,9 \\ 1 & 98,2 & 100,7 \\ 1 & 101,8 & 100,6 \\ 1 & 125,7 & 100,8 \\ 1 & 76,8 & 106,2 \\ 1 & 101 & 101 \\ 1 & 106,2 & 100,7 \\ 1 & 99 & 100,6 \\ 1 & 103,7 & 100,6 \end{pmatrix}; \quad X_0 = \begin{pmatrix} 1 \\ 108 \\ 100,7 \end{pmatrix}; \quad X^T \cdot X = \begin{pmatrix} 12 & 1221,2 & 1214,68 \\ 1221,2 & 125547 & 123480,2 \\ 1214,68 & 123480,2 & 122980,9 \end{pmatrix};$$

$$\left(X^T \cdot X\right)^{-1} = \begin{pmatrix} 985,694 & -0,999 & -8,732 \\ -0,999 & 0,002 & 0,008 \\ -8,732 & 0,008 & 0,078 \end{pmatrix}; \qquad X_0^T \cdot \left(X^T \cdot X\right)^{-1} \cdot X_0 \approx 0,115.$$

При расчетах были использованы математические функции пакета *Мастера функций* Excel (категория математические) *МУМНОЖ* (возвращает матричное произведение двух массивов) и *МОБР* (возвращает обратную матрицу). В результате вычислений имеем: $S\left(\hat{Y}_0\right) = 3,132 \cdot \sqrt{0,115} = 1,062$. Из таблиц критических точек распределения Стьюдента (см. также Excel – f_x – статистические – стьюдент.обр.2х) находим: $t_{\frac{\alpha}{2};n-2} = t_{\frac{0,05}{2};12-2-1} = 2,262$. Подставив эти значения в формулу (2.1), получим 95%-ный доверительный интервал для прогнозного среднего значения результативного признака Y при $X_2 = 108$, $X_6 = 100,7$:

$$106,44 - 2,262 \cdot 1,062 < M\left(Y \middle| X_0^T = (1;108;100,7)^T\right) < 106,44 + 2,262 \cdot 1,062,$$

$$104,038 < M\left(Y \middle| X_0^T = (1,108;100,7)^T\right) < 108,842.$$

Основываясь на выборочных данных, можно утверждать, что средний оборот розничной торговли непродовольственными товарами в июне будет находиться в найденном доверительном интервале с вероятностью 0,95.