15. РЕГРЕССИОННЫЙ АНАЛИЗ

Задача любого исследования состоит в установлении причинных зависимостей. Только знание истинных причин явлений позволяет правильно объяснять наблюдаемые закономерности. Корреляционный анализ не вскрывает причинного характера связи. Корреляция дает лишь оценку силы, или тесноты связи. Для вскрытия причинного характера связи между явлениями (переменными, признаками) используют регрессионный анализ (см. также [116]).

15.1. Понятие «регрессия»

Понятие «регрессия» связано с Фрэнсисом Гальтоном. В 1885 году был издан его научный труд «*Регрессия в направлении к общему среднему размеру при наследовании роста*». В этой работе он пришел к выводу, что признаки родителей не полностью наследуются детьми, и чем отдаленнее предок, тем в меньшей мере сказываются его свойства на потомке. Гальтон показал, что дети очень высоких или очень низких родителей в среднем имеют менее высокий или соответственно менее низкий рост. Кроме того, отклонение роста детей не так велико, как отклонение роста их родителей от среднего роста исследованных лиц. Это движение назад в направлении к среднему Гальтон назвал *регрессией* (to regress - движение в обратном направлении). Гальтон писал: «*Закон регрессии веско свидетельствует против полного наследования какого-либо признака. Из большого числа детей только немногие будут уклоняться от среднего уровня по сравнению с уклонением одного из родителей, отличающегося своими природными качествами. Чем ярче талант одного из родителей, тем реже родители имеют счастье видеть, что природа также щедро одарила их сыновей, и еще реже бывает, чтобы одаренность передавалась в последующие поколения. Закон беспристрастен и объективен. Он равномерно распределяет наследование хороших и плохих признаков. Он разрушает чрезмерные иллюзии одного одаренного родителя, лелеющего мечту, что его дети унаследуют все его способности. Закон устраняет также преувеличенные опасения относительно того, что детям передадуться все слабости, недостатки и болезни родителей. Разумеется, эти утверждения не находятся*

в противоречии с общей теорией, согласно которой дети талантливых родителей имеют бóльшую вероятность обладать какими-либо дарованиями, чем дети родителей со средними способностями. Наши рассуждения выражают только тот факт, что самый одаренный из всех детей немногих высокоодаренных родительских пар не так будет талантлив, как самый одаренный из всех детей очень многих родительских пар со средними способностями.»

В статистической трактовке *регрессией* называют изменение функции в зависимости от изменений одного или нескольких аргументов. Под *функцией* понимают переменную, которая зависит от другой переменной - *аргумента* (независимая переменная). Регрессия - это односторонняя статистическая зависимость. При простой корреляции изучают зависимость между изменчивостью двух переменных X и Y. С помощью регрессии ставится дополнительная задача: установить, как количественно меняется одна переменная при изменении другой (или других) на единицу. Если исследуют зависимость переменной Y от X, то устанавливают регрессию Y на X. Если же изучают зависимость переменной X от Y, то определяют регрессию X на Y. Цель регрессионного анализа - по значениям одной переменной, выбранной в качестве аргумента, предсказать соответствующее значение другой (функции). В этом заключается первое отличие метода регрессии от метода корреляции. Второе отличие состоит в том, что степень и характер регрессии можно установить и при небольшом числе пар значений зависимой и независимой переменных.

15.2. Задачи регрессионного анализа

В исследованиях по животноводству регрессионный анализ используют для решения следующих задач:

1. *Установления формы зависимости между переменными* (линейная-нелинейная, отрицательная-положительная и т.д.).
2. *Определения функции регрессии*. Важно выяснить, каково было бы действие на зависимую переменную главных факторов, если бы прочие факторы не изменялись и если бы были исключены случайные элементы.

3. *Прогностической оценки неизвестных значений зависимой переменной*. С помощью функции регрессии можно воспроизвести значения зависимой переменной внутри интервала заданных значений независимых переменных (*интерполяция*) или оценить течение процесса вне заданного интервала (*экстраполяция*).

15.3. Виды регрессии

Относительно числа учитываемых признаков регрессия может быть: *простой* - между двумя переменными, и *множественной* (или *частной*) - между зависимой переменной Y и несколькими независимыми (объясняющими) переменными: X_1, X_2, ...X_m; относительно формы зависимости - *линейной* и *нелинейной*; относительно направления связи - *положительной* и *отрицательной*.

По характеру отношений между зависимой и независимыми переменными регрессия может быть *непосредственной* (причина оказывает прямое воздействие на следствие), *косвенной* (независимая переменная действует через какую-то третью или ряд других причин на зависимую переменную) и *ложной* (нонсенс-регрессия - возникает при формальном подходе без уяснения причин, которые обуславливают данную связь).

15.4. Простая линейная регрессия

Под простой линейной регрессией понимают одностороннюю линейную статистическую зависимость признака только от одной независимой переменной. Анализируемый признак чаще называют *зависимой* или *результативной* переменной и обозначают символом «*y*», а фактор-причину - *независимой* или *объясняющей* переменной и обозначают символом «*x*» (в случае множественной регрессии - x_k, где $k = 1,...,m$ факторов).

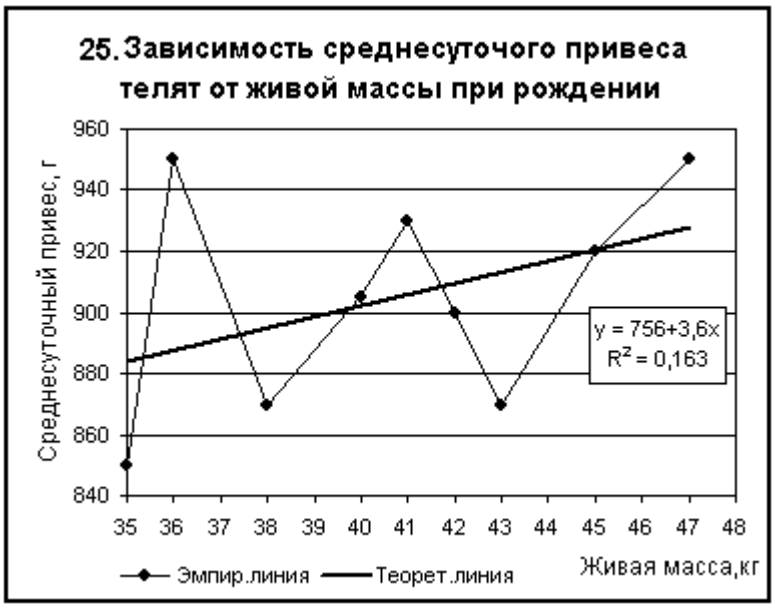
Простая линейная регрессия может быть выражена:

- эмпирической линией регрессии;
- уравнением регрессии и теоретической линией регрессии;
- коэффициентом регрессии.

15.4.1. Эмпирическая линия регрессии

Для построения линии регрессии необходимо иметь два ряда данных. На горизонтальной оси *x* системы координат отмечают значения независимой переменной. На вертикальной оси *y* - значения зависимой переменной, соответствующие значениям независимой переменной. Соединяющая все точки линия представляет собой линию регрессии Y по X (см. рис. 25)

Живая масса,кг **x**	Привес, г/сутки **y**
35	850
36	950
38	870
40	905
41	930
42	900
43	870
45	920
47	950



15.4.2. Уравнение подбора прямой регрессии

Эмпирическая линия регрессии обычно представляет собой более или менее ломаную линию. Несмотря на наглядность характера связи между X и Y, она не дает возможности точно определить любое значение Y по заданному значению X. Для этой цели используют уравнение регрессии, которое в общем виде можно записать так:

$$y_i - \bar{y} = b(x_i - \bar{x}) + e_i,$$

где y_i - значение i-го наблюдения зависимой переменной $(i = 1,...,n)$; x_i - значение соответствующей независимой переменной; $\bar{x}$ и $\bar{y}$ - средние по n наблюдениям; b – коэффициент пропорциональности; e_i - ошибка.

Уравнение выражает определенную зависимость: вслед за отклонением x_i от среднего по переменной X происходит и отклонение y_i от среднего по переменной Y. Показатель b

является коэффициентом *пропорциональности*, т.е. мерой, которая *в среднем* указывает на количественное изменение Y при изменении X на определенную величину.

Перенеся $\bar{y}$ в правую часть равенства, получим

$$y_i = \bar{y} + b(x_i - \bar{x}) + e_i .$$

Если $\bar{x}$ приравнять нулю, то $\bar{y}$ будет являться первоначальным значением Y, с которого надо начинать при построении линии регрессии, когда x_i=0. Поэтому его обычно обозначают через b_0 или *a*. Тогда уравнение линейной регрессии принимает вид:

$$y_i = b_0 + b_1 x_i + e_i$$

или

$$y_i = a + b x_i + e_i .$$

Это уравнение для простой линейной регрессии, где x_i - независимая переменная (фактор-причина); *a* (или b_0) и *b* (или b_1) являются параметрами регрессии, которые подлежат оценке.

a (b_0) - это константа регрессии. Она определяет точку пересечения прямой регрессии с осью ординат. *a* (b_0) является *средним* значением Y в точке x_i=0. Поэтому биологическая интерпретация *a* (или b_0) часто бывает затруднительной или даже невозможной. Константа выполняет в уравнении регрессии функцию выравнивания*. Благодаря ей функция регрессии является *несмещенной*.

b (b_1) - коэффициент пропорциональности, который характеризует наклон прямой к оси абсцисс; является мерой *влияния* переменной X на переменную Y, или мерой *зависимости* переменной Y от переменной X. Он указывает *среднюю* величину изменения переменной Y при изменении X на *одну* единицу. Знак при *b* (b_1) определяет направление этого изменения. Положительное значение означает поступательный характер изменения зависимой переменной при увеличении значений

* Константу можно представить в виде коэффициента при фиктивной переменной, принимающей для всех i=1, …, n значение 1; фиктивная переменная обычно не записывается, но иногда с математической точки зрения ее удобно включить в уравнение.

аргумента. При отрицательном b (b_1) имеет место отрицательная регрессия - с увеличением x_i значения переменной Y убывают.

Параметры регрессии - не безразмерные величины. Константа уравнения регрессии, a (b_0), имеет размерность зависимой переменной. Размерность b (b_1) есть отношение размерности зависимой переменной к размерности независимой переменной.

Параметры уравнения регрессии неизвестны. Различным значениям a и b будут соответствовать различные прямые регрессии. Поэтому задача регрессионного анализа состоит в нахождении таких оценок этих параметров (*подборе* прямой), которые бы наиболее хорошо согласовывались с фактическими данными. Для этого используют метод *наименьших квадратов* (Least Squares, LS).

Система нормальных уравнений LS-метода для простой линейной регрессии имеет вид:

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}.$$

Решение системы дает несмещенные оценки b и a:

$$\hat{b} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2},$$

$$\hat{a} = \bar{y} - \hat{b}\,\bar{x}.$$

Эти оценки включают в уравнение для подбора прямой регрессии Y на X:

$$\hat{y}_i = \hat{a} + \hat{b}\,x_i.$$

По данному уравнению для каждого значения независимой переменной x_i (i=1,…,n) рассчитывают функцию регрессии - $\hat{y}_i$. Значения функции регрессии $\hat{y}_i$ (i=1,…,n) называют *предсказанными* или *рассчетными* значениями переменной Y для фиксированных x_i. Тогда наблюдаемое значение (y_i) можно представить как

$$\begin{aligned} y_i &= \hat{y}_i + e_i \\ &= \hat{a} + \hat{b}\,x_i + e_i, \end{aligned}$$

где e_i - возмущающая переменная или остаток, включающая влияние неучтенных факторов (интерпретируется как ошибка).

Прогностические оценки ($\hat{y}_i$) являются наилучшими линейными приближениями (аппроксимацией) к фактическим (эмпирическим) значениям, y_i, т.к. их стандартная ошибка сведена LS-методом к минимуму. Совокупность предсказанных значений образует теоретическую линию регрессии (см. рис.25 и 26).

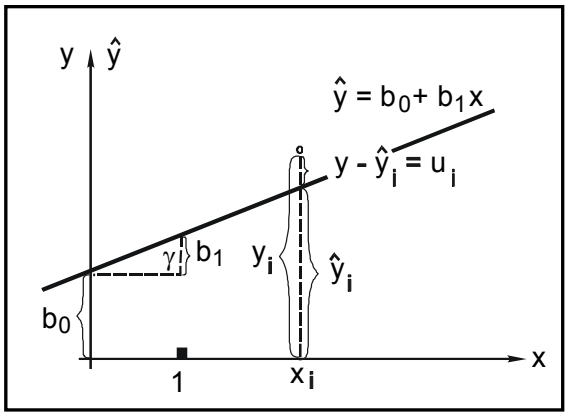


Рис. 26. Прямая регрессия и ее параметры

Из-за модифицирующего влияния неучтенных факторов для каждого значения x_i может наблюдаться несколько эмпирических значений y_i. Значения функции регрессии ($\hat{y}_i$) являются, таким образом, оценками *средних* значений переменной Y для каждого фиксированного значения переменной X.

Пример 15.1. Требуется описать линейной зависимостью соотношение живой массы при рождении (x_i) и скорости роста (y_i) 10 бычков:

x_i	40	42	35	36	45	47	40	43	41	38
y_i	1000	900	850	950	920	950	810	870	930	870

Решение:

$$\sum x_i = 40 + 42 + ... + 41 + 38 = 407;$$

$$\sum y_i = 1000 + 900 + ... + 930 + 870 = 9050;$$

$$\sum x_i^2 = 40^2 + 42^2 + ... + 41^2 + 38^2 = 16693;$$

$$\sum x_i y_i = 40 \times 1000 + 42 \times 900 + ... + 41 \times 930 + 38 \times 870 = 368800;$$

Система уравнений для линейной регрессии Y на X:

$$\begin{bmatrix} 10 & 407 \\ 407 & 16693 \end{bmatrix} \begin{bmatrix} a \\ b_{yx} \end{bmatrix} = \begin{bmatrix} 9050 \\ 368800 \end{bmatrix}.$$

Оценки параметров регрессии:

$$\hat{b}_{yx} = \frac{10 \times 368800 - 407 \times 9050}{10 \times 16693 - 407^2} = \frac{+4650}{1281} = +3{,}63 \text{ г/кг};$$

$$\hat{a} = 905 - 3{,}63 \times 40{,}7 = 757{,}3 \text{ г.}$$

Функция регрессии: $\hat{y}_i = 757{,}3 + 3{,}63 x_i$,

или, если $y_i = \bar{y} + b_{yx}(x_i - \bar{x}) + e_i$, то

$$\hat{y}_i = 905 + 3{,}63(x_i - \bar{x}).$$

С помощью этих уравнений вычисляют прогностические оценки привесов ($\hat{y}_i$) для живой массы бычков при рождении (x_i). Например, для живой массы бычка № 1 равной 40 кг, прогностическая оценка среднесуточного привеса составит

$$\hat{y}_1 = 757{,}3 + 3{,}63 \times 40 = 902{,}5 \text{г} \quad \text{или}$$
$$\hat{y}_1 = 905 + 3{,}63(40 - 40{,}7) = 902{,}5 \text{г}.$$

Это прогнозируемое (теоретическое) значение представляет собой наилучшее, в смысле LS-метода, линейное приближение (аппроксимацию) к фактическому (эмпирическому) значению, y_1=1000 г, т.к. стандартная ошибка прогноза минимизирована. По прогностическим значениям построена теоретическая прямая регрессии (рис. 25).

Можно прогнозировать среднесуточный привес для любого значения живой массы бычка при рождении. Так, при живой массе 30 кг ожидаемый среднесуточный привес составит:

$$\hat{y} = 757{,}3 + 3{,}63 \times 30 = 866 \text{ грамм.}$$

Пример 15.2. Эффективным средством для определения и решения уравнений линейной регрессии являются матрицы и матричная алгебра.

Относительно данных **примера 15.1**, уравнение регрессии для i-го животного можно записать в следующем виде:

$$y_i = \mu + b_{yx}(x_i - \bar{x}) + e_i,$$

где $\mu = \bar{y}$ и $b_{yx} = b$.

Тогда для животного № 1 уравнение регрессии запишется так:

$$1000 = \mu + b_{yx}(-0{,}7) + e_1,$$

где $$-0{,}7 = x_i - \bar{x} = 40 - 40{,}7.$$

Уравнения для всех животных могут быть представлены в следующем матричном виде (для константы μ введена фиктивная переменная, равная 1 для всех i=1, …, n):

$$\begin{bmatrix}1000\\900\\850\\950\\920\\950\\810\\870\\930\\870\end{bmatrix}=\mu\begin{bmatrix}1\\1\\1\\1\\1\\1\\1\\1\\1\\1\end{bmatrix}+b_{yx}\begin{bmatrix}-0,7\\1,3\\-5,7\\-4,7\\4,3\\6,3\\-0,7\\2,3\\0,3\\-2,7\end{bmatrix}+\begin{bmatrix}e_1\\e_2\\e_3\\e_4\\e_5\\e_6\\e_7\\e_8\\e_9\\e_{10}\end{bmatrix}$$

или

$$\begin{bmatrix}1000\\900\\850\\950\\920\\950\\810\\870\\930\\870\end{bmatrix}=\begin{bmatrix}1&-0,7\\1&1,3\\1&-5,7\\1&-4,7\\1&4,3\\1&6,3\\1&-0,7\\1&2,3\\1&0,3\\1&-2,7\end{bmatrix}\begin{bmatrix}\mu\\b_{yx}\end{bmatrix}+\begin{bmatrix}e_1\\e_2\\e_3\\e_4\\e_5\\e_6\\e_7\\e_8\\e_9\\e_{10}\end{bmatrix}.$$

Умножение матриц дает уравнение для каждого животного в наборе данных. Теперь определим:

- y - вектор наблюдений

$$y' = [1000\ 900\ 850\ 950\ 920\ 950\ 810\ 870\ 930\ 870];$$

- X - матрица инцидентности (плана)

$$X' = \begin{bmatrix}1&1&1&1&1&1&1&1&1&1\\-0,7&1,3&-5,7&-4,7&4,3&6,3&-0,7&2,3&0,3&-2,7\end{bmatrix};$$

- b - вектор оцениваемых параметров

$$b = \begin{bmatrix}\mu\\b_{yx}\end{bmatrix};$$

- e - вектор рандомизированной ошибки

$$e' = [e_1\ e_2\ e_3\ e_4\ e_5\ e_6\ e_7\ e_8\ e_9\ e_{10}].$$

Запись уравнения регрессии в матричном виде:

$$y = Xb + e.$$

Оценки наименьших квадратов могут быть получены решением следующих нормальных уравнений:

$$(X'X)\hat{b} = X'y,$$

где

$$X'X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -0,7 & 1,3 & -5,7 & -4,7 & 4,3 & 6,3 & -0,7 & 2,3 & 0,3 & -2,7 \end{bmatrix} \begin{bmatrix} 1 & -0,7 \\ 1 & 1,3 \\ 1 & -5,7 \\ 1 & -4,7 \\ 1 & 4,3 \\ 1 & 6,3 \\ 1 & -0,7 \\ 1 & 2,3 \\ 1 & 0,3 \\ 1 & -2,7 \end{bmatrix} = \begin{bmatrix} 10 & 0 \\ 0 & 128,1 \end{bmatrix};$$

$$\hat{b} = \begin{bmatrix} \hat{\mu} \\ \hat{b}_{yx} \end{bmatrix};$$

$$X'y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -0,7 & 1,3 & -5,7 & -4,7 & 4,3 & 6,3 & -0,7 & 2,3 & 0,3 & -2,7 \end{bmatrix} \begin{bmatrix} 1000 \\ 900 \\ 850 \\ 950 \\ 920 \\ 950 \\ 810 \\ 870 \\ 930 \\ 870 \end{bmatrix} = \begin{bmatrix} 9050 \\ 465 \end{bmatrix}.$$

Умножение обеих частей уравнения $(X'X)\hat{b} = X'y$ на $(X'X)^{-1}$ дает:

$$(X'X)^{-1}(X'X)\hat{b} = (X'X)^{-1}X'y$$

$$I\,\hat{b} = (X'X)^{-1}X'y.$$

Таким образом,

$$\hat{b} = (X'X)^{-1}X'y =$$

$$= \begin{bmatrix} 10 & 0 \\ 0 & 128,1 \end{bmatrix}^{-1} \begin{bmatrix} 9050 \\ 465 \end{bmatrix} =$$

$$= \begin{bmatrix} 1/10 & 0 \\ 0 & 1/128,1 \end{bmatrix} \begin{bmatrix} 9050 \\ 465 \end{bmatrix} =$$

$$= \begin{bmatrix} (1/10)(9050) + (0)(465) \\ (0)(9050) + (1/128,1)(465) \end{bmatrix} =$$

$$= \begin{bmatrix} 905,0 \\ 3,63 \end{bmatrix}.$$

Полученные значения идентичны оценкам среднего значения ($\bar{y}$) и коэффициента пропорциональности (b), рассчитанным по алгоритму скалярной алгебры.

15.4.3. Коэффициент регрессии

В биозоотехнических исследованиях часто интерес представляет не сама *прямая регрессии*, а влияние, которое оказывает одна переменная на другую. В таких случаях рассчитывают коэффициент регрессии. *Коэффициент регрессии* - это отношение ковариансы между независимой и зависимой переменными к варианс независимой переменной:

$$\hat{b}_{yx} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2}.$$

Ранее было показано, что

$$\hat{\sigma}_{xy} = \frac{SP_{xy}}{n-1}$$

и

$$\hat{\sigma}_x^2 = \frac{SS_x}{n-1},$$

где SP_{xy} - сумма произведений отклонений от средних: $\sum(x_i - \bar{x})(y_i - \bar{y})$;

SS_x - сумма квадратов отклонений от средней: $\sum(x_i - \bar{x})^2$.

Тогда оценку коэффициента регрессии по выборке можно рассчитать по следующим формулам:

$$\hat{b}_{yx} = \frac{SP_{xy}/(n-1)}{SS_x/(n-1)} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} =$$

$$= \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} =$$

$$= \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} =$$

$$= \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{\sum x_i^2 - n\bar{x}^2}.$$

Таким образом, коэффициент регрессии представляет собой не что иное, как коэффициент пропорциональности ($\hat{b}$) в уравнении регрессии:

$$\hat{y}_i = a + \hat{b}\, x_i .$$

Как и коэффициент пропорциональности, коэффициент регрессии является мерой *зависимости* переменной Y от переменной X. Он показывает среднюю величину изменения переменной Y при изменении X на одну единицу. Знак при коэффициенте регрессии определяет направление этого изменения. Положительное значение означает поступательный характер изменения зависимой переменной при увеличении значений аргумента. При отрицательном коэффициенте регрессии имеет место негативный характер изменений, при котором с увеличением X значения переменной Y убывают.

В примере 15.1 коэффициент регрессии указывает на то, что с увеличением живой массы теленка на 1 кг можно ожидать повышения среднесуточного привеса на 3,63 грамма; аналогично, с уменьшением живой массы на 1 кг - снижение среднесуточного привеса на 3,63 грамма. Вместе с тем следует отметить, что по коэффициенту регрессии нельзя определить насколько сильна связь между двумя переменными.

Регрессионный анализ дает более широкую информацию, чем корреляционный. Он позволяет установить зависимость как переменной Y по переменной X, так и наоборот - X по Y. Поэтому коэффициентов регрессии может быть два:

- изменение переменной Y при изменении переменной X

$$\hat{b}_{yx} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2};$$

- изменение переменной X при изменении переменной Y

$$\hat{b}_{xy} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_y^2}.$$

Пример 15.3. Для данных примера 15.1 получим:

$$\sum y_i^2 = 1000^2 + 900^2 + ... + 930^2 + 870^2 = 8218700;$$

$$\hat{b}_{xy} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_y^2} = \frac{\sum x_i y_i - n\bar{x}\,\bar{y}}{\sum y_i^2 - n\bar{y}^2} =$$

$$= \frac{368800 - 10 \times 40{,}7 \times 905}{8218700 - 10 \times 905^2} = \frac{+465}{28450} = +0{,}016 \text{ кг/г.}$$

Результат свидетельствует о том, что с каждым повышением среднесуточного привеса на 1 грамм живая масса бычков при рождении будет возрастать на 0,016 кг. С биологической точки зрения зависимость живой массы при рождении от среднесуточного привеса лишена смысла. Здесь эта зависимость дана лишь для иллюстрации расчетов.

15.5. Нулевая гипотеза и доверительный интервал

Отклонение прогностической оценки ($\hat{y}_i$) от фактического значения зависимой переменной (y_i) называют *остатком* или *ошибкой* прогноза:

$$y_i - \hat{y}_i = \hat{e}_i \qquad \text{при } i=1, \ldots, n.$$

Ошибки прогноза имеют свое распределение со средним равным нулю и вариансой σ_e^2. Эту остаточную вариансу оценивают по формуле:

$$\hat{\sigma}_e^2 = \frac{SS_e}{df} = \frac{\sum (y_i - \hat{y}_i)^2}{n - (m+1)} = \frac{\sum \hat{e}_i^2}{n - (m+1)},$$

где SS_e - сумма квадратов остатков; df - число степеней свободы; n - объем выборки; m - число независимых переменных.

Физический смысл остаточной вариансы - это та часть общей вариансы зависимой переменной Y ($\hat{\sigma}_y^2$), которую нельзя объяснить зависимостью переменной Y от переменной X.

Остаточную вариансу используют для расчета вариансы, обусловленной регрессией или независимой переменной ($\hat{\sigma}_{\hat{b}}^2$):

$$\hat{\sigma}_{\hat{b}}^2 = \frac{\hat{\sigma}_e^2}{SS_x} = \frac{\hat{\sigma}_e^2}{\sum (x_i - \bar{x})^2} = \frac{n\ \hat{\sigma}_e^2}{\sum x_i^2 - n\bar{x}^2} = \frac{\hat{\sigma}_e^2}{\hat{\sigma}_x^2 (n-1)},$$

и ошибки оценки коэффициента регрессии ($m_{\hat{b}}$)

$$m_{\hat{b}} = \sqrt{\hat{\sigma}_{\hat{b}}^2} = \frac{\hat{\sigma}_e}{\hat{\sigma}_x \sqrt{n-1}},$$

где $\hat{\sigma}_e$ - стандартное отклонение остатков; SS_x и $\hat{\sigma}_x$ - сумма квадратов и стандартное отклонение независимой переменной соответственно.

Таким образом, стандартная ошибка выборочной оценки коэффициента регрессии прямопропорциональна рассеянию остатков и обратнопропорциональна рассеянию значений независимой переменной (аргумента) и объему выборки.

Для проверки нулевой гипотезы (H_0: $\hat{b}=0$) рассчитывают t-статистику:

$$t_{\hat{b}}=\frac{|\hat{b}-0|}{m_{\hat{b}}}.$$

Если $t_{\hat{b}} \geq t_{\alpha;df}$ при числе степеней свободы df=n-m-1, то нулевую гипотезу при уровне значимости α отклоняют; оценку коэффициента регрессии считают статистически значимой.

Доверительный интервал для истинного значения коэффициента регрессии:

$$(\hat{b}-t_{\alpha;\,df}\,m_{\hat{b}}) < b < (\hat{b}+t_{\alpha;\,df}\,m_{\hat{b}}).$$

Можно утверждать, что с доверительной вероятностью $P=1-\alpha$ параметр в генеральной совокупности (истинное значение b) не выйдет за пределы этих границ.

Пример 15.4. Проверка нулевой гипотезы и построение доверительного интервала иллюстрируются для коэффициента регрессии среднесуточного привеса бычков на их живую массу при рождении (пример 15.1).

По функции регрессии

$$\hat{y}_i = 757{,}3 + 3{,}63\,x_i$$

рассчитаны прогностические оценки $\hat{y}_i$ (например, для бычка № 1 $\hat{y}_1 = 757{,}3 + 3{,}63 \times 40 = 902{,}5$ г) и их ошибки ($e_1 = 1000 - 902{,}5 = 97{,}5$):

№	x_i	y_i	$\hat{y}_i$	$\hat{e}_i$	$\hat{e}_i^2$
1	40	1000	902,5	97,5	9506,25
2	42	900	909,7	-9,7	94,09
3	35	850	884,3	-34,3	1176,49
4	36	950	887,9	62,1	3856,41
5	45	920	920,6	-0,6	0,36
6	47	950	927,9	22,1	488,41
7	40	810	902,5	-92,5	8556,25
8	43	870	913,3	-43,3	1874,89
9	41	930	906,1	23,9	571,21
10	38	870	895,2	-25,2	635,04
		-	-	-	26759,4

По прогностическим оценкам проводят теоретическую линию регрессии (см. рис. 25).

Варианса и стандартное отклонение остатков:

$$\hat{\sigma}_e^2 = \frac{SS_e}{n-(m+1)} = \frac{26759,4}{10-(1+1)} = 3345;$$

$$\hat{\sigma}_e = \sqrt{3345} = 57,8.$$

Расчет SS_x, $\hat{\sigma}_x^2$ и $\hat{\sigma}_x$:

$$\sum x_i = 40+42+...+41+38 = 407;$$

$$\bar{x} = 407/10 = 40,7;$$

$$\sum x_i^2 = 40^2+42^2+...+41^2+38^2 = 16693;$$

$$SS_x = \sum (x_i-\bar{x})^2 = \sum x_i^2 - N\bar{x}^2 =$$

$$= 16693 - 10\times 40,7^2 = 128,1;$$

$$\hat{\sigma}_x^2 = \frac{128,1}{10-1} = 14,2;$$

$$\hat{\sigma}_x = \sqrt{14,2} = 3,77;$$

Варианса и ошибка регрессии:

$$\hat{\sigma}_{\hat{b}}^2 = \frac{\hat{\sigma}_e^2}{SS_x} = \frac{3345}{128,1} = 26,1;$$

$$m_{\hat{b}} = \sqrt{\hat{\sigma}_{\hat{b}}^2} = \sqrt{26,1} = 5,1 \quad \text{или}$$

$$= \frac{\hat{\sigma}_e}{\hat{\sigma}_x\sqrt{n-1}} = \frac{57,8}{3,77\sqrt{10-1}} = 5,1.$$

Таким образом,

$$\hat{b} \pm m_{\hat{b}} = 3,63 \pm 5,1 \text{ г/кг.}$$

Фактический t-критерий:

$$t_{\hat{b}} = \frac{\hat{b}_{yx}}{m_{\hat{b}}} = \frac{+3,63}{5,1} = 0,71.$$

При $\alpha = 5\%$ и числе степеней свободы df=10-(1+1)=8 критическое значение $t_{0,05;8}$=2,31 (табл. А.8 Приложения А). Так как $t_{\hat{b}} < t_{0,05;8}$, то нулевая

гипотеза остается в силе. Истинное значение коэффициента регрессии с доверительной вероятностью 95% находится в диапозоне оценок

$$(\hat{b}-t_{0,05;8}\,m_{\hat{b}}) < b < (\hat{b}+t_{0,05;8}\,m_{\hat{b}}),\quad \text{т.е.}$$

от (3,63-2,306×5,1) до (3,63+2,306×5,1) или
от -8,1 до +15,4 г/кг

и может принимать нулевое значение.

Пример 15.5. Регрессионный анализ многоплодия (X) и молочности (Y) 8 свиноматок иллюстрирует иную технику расчета.

Исходные данные и расчет сумм квадратов и произведений:

x_i	y_i	$(x_i-\bar{x})$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
10	53	+0,25	0,0625	+3	9	+0,75
7	49	-2,75	7,5625	-1	1	+2,75
9	54	-0,75	0,5625	+4	16	-3,00
6	36	-3,75	14,0625	-14	196	+52,5
14	55	+4,25	18,0625	+5	25	+21,25
8	52	-1,75	3,0625	+2	4	-3,5
11	50	+1,25	1,5625	0	0	0
13	51	+3,25	10,5625	+1	1	+3,25
Σ 78	400	0,00	SS_x =55,5	0	SS_y =252	SP_{xy} =+74,00

$$\bar{x}=\frac{\sum x_i}{n}=\frac{78}{8}=9{,}75;\qquad \hat{\sigma}_x^2=\frac{SS_x}{n-1}=\frac{55{,}5}{8-1}=7{,}93;$$

$$\bar{y}=\frac{\sum y_i}{n}=\frac{400}{8}=50{,}0;\qquad \hat{\sigma}_y^2=\frac{SS_y}{n-1}=\frac{252}{8-1}=36;$$

$$\hat{\sigma}_{xy}=\frac{SP_{xy}}{n-1}=\frac{+74}{8-1}=+10{,}57;$$

Коэффициенты регрессий:

$$\hat{b}_{yx}=\frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2}=\frac{+10{,}57}{7{,}93}=+1{,}33\ \text{кг/гол};$$

$$\hat{b}_{xy}=\frac{\hat{\sigma}_{xy}}{\hat{\sigma}_y^2}=\frac{+10{,}57}{36}=+0{,}294\ \text{гол/кг}.$$

Коэффициенты регрессий показывают, что при увеличении многоплодия на одну голову молочность свиноматок повышалась в среднем на 1,3 кг. С другой стороны, увеличение молочности свиноматок на 1 кг повышало их многоплодие в среднем на 0,3 головы.

Константы уравнений регрессий:

$$\hat{a}_y = \bar{y} - \hat{b}_{yx}\ \bar{x} = 50 - 1{,}33 \times 9{,}75 = 37{,}0 \text{ кг},$$

$$\hat{a}_x = \bar{x} - \hat{b}_{xy}\ \bar{y} = 9{,}75 - 0{,}294 \times 50 = -4{,}95$$

Функция регрессии Y на X:

$$\hat{y}_i = 37 + 1{,}33\ x_i,$$

и X на Y:

$$\hat{x}_i = -4{,}95 + 0{,}294\ y_i.$$

Функция регрессии молочности свиноматок на их многоплодие:

$$\hat{y}_i = 37 + 1{,}33\, x_i.$$

Расчет прогностических оценок ($\hat{y}_i$), остатков ($\hat{e}_i$), суммы квадратов остатков (SS_e) и суммы квадратов прогностических оценок ($SS_{\hat{y}}$):

x_i	y_i	$\hat{y}_i$	$\hat{e}_i = (y_i - \hat{y}_i)$	$\hat{e}_i^2$	$(\hat{y}_i - \bar{y})$	$(\hat{y}_i - \bar{y})^2$
10	53	50,3	+2,7	7,29	+0,3	0,09
7	49	46,3	+2,7	7,29	-3,7	13,69
9	54	49,0	+5,0	25,00	-1,0	1,00
6	36	45,0	-9,0	81,00	-5,0	25,00
14	55	55,7	-0,7	0,49	+5,7	32,49
8	52	47,7	+4,3	18,49	-2,3	5,29
11	50	51,7	-1,7	2,89	+1,7	2,89
13	51	54,3	-3,3	10,89	4,3	18,49
Σ 78	400	400	0,0	SS_e =153,34	0,0	$SS_{\hat{y}}$ =98,94

По фактическим (y_i) и прогностическим оценкам ($\hat{y}_i$) проведены фактическая и теоретическая линии регрессии (рис. 27).

Рис. 27. Фактическая и теоретическая линии регрессии молочности свиноматок на многоплодие

Варианса остатков: $\hat{\sigma}_e^2 = \frac{SS_e}{n-(m+1)} = \frac{153{,}34}{8-(1+1)} = 25{,}56$.

Варианса регрессии: $\hat{\sigma}_b^2 = \frac{\hat{\sigma}_e^2}{SS_x} = \frac{25{,}56}{55{,}5} = 0{,}46$.

Проверка нулевой гипотезы для оценки коэффициента регрессии молочности свиноматок (Y) на многоплодие (X):

Ошибка: $m_{\hat{b}} = \pm\sqrt{\hat{\sigma}_{\hat{b}}^2} = \pm\sqrt{0{,}46} = \pm 0{,}678$;

Регрессия с ошибкой: $\hat{b} \pm m_{\hat{b}} = 1{,}33 \pm 0{,}678$ кг/гол.

Фактический t-критерий: $t_{\hat{b}} = \frac{\hat{b}_{yx}}{m_{\hat{b}}} = \frac{+1{,}33}{0{,}678} = 1{,}96$.

При α=10% и числе степеней свободы df=8-1-1=6 по табл. А.8 Приложения А находим $t_{0{,}10;6}$=1,94. Так как $t_{\hat{b}} > t_{0{,}10;6}$, то нулевая гипотеза может быть отклонена. Истинное значение коэффициента регрессии с доверительной вероятностью 90% находится в диапозоне оценок

$$(\hat{b} - t_{0{,}10;6}\, m_{\hat{b}}) < b < (\hat{b} + t_{0{,}10;6}\, m_{\hat{b}}), \quad \text{т.е.}$$

от (1,33-1,94×0,678) до (1,33+1,94×0,678) или

от +0,001 до 2,659 кг/гол.

Если принять α=5%, то $t_{0{,}05;6}$=2,45 и нулевая гипотеза не отклоняется, т.к. $t_{\hat{b}} < t_{0{,}05}$. Из этого следует, что выбор уровня значимости (α) оказывает решающее влияние на принятие или отклонение нулевой гипотезы. Поэтому для исключения субъективности в обсуждении результатов эксперимента, необходимо значение α задавать заранее, до получения результатов.

15.6. Коэффициент детерминации

Общую сумму квадратов *зависимой* переменной (SS_y) характеризующую разброс наблюдаемых значений переменной Y около среднего, можно разложить на компоненты:

$$SS_y = \sum(y_i - \bar{y})^2 =$$

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

$$= SS_e + SS_{\hat{y}}.$$

SS_e - это часть общей суммы квадратов SS_y, которая не объясняется функцией регрессии. Поэтому её называют остаточной. Она характеризует ту часть рассеяния переменной Y, которая возникает из-за случайностей и изменчивости прочих неучтенных факторов.

$SS_{\hat{y}}$ - это сумма квадратов значений регрессии $\hat{y}_i$, т.к. $\bar{y} = \bar{\hat{y}}_i$. Эта сумма квадратов представляет ту часть рассеяния переменной Y, которая, в основном, обусловлена влиянием независимых переменных X_k $(k = 1, \ldots, m)$. В связи с этим $SS_{\hat{y}}$ называют *суммой квадратов, которая обусловлена регрессией*.

Альтернативные формулы расчета $SS_{\hat{y}}$:

$$SS_{\hat{y}} = \sum(\hat{y}_i - \bar{y}_i)^2 =$$

$$= \frac{\left[\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}\right]^2}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}} =$$

$$= \hat{b}\left[\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}\right] =$$

$$= \hat{b}^2\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right] =$$

$$= \hat{b}^2\left[\sum x_i^2 - n\bar{x}^2\right].$$

Так как регрессия равна

$$\hat{b}_{yx} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x^2},$$

то ковариансу и вариансу можно определить отдельно, методом дисперсионного анализа, и затем использовать эти оценки для расчета коэффициента регрессии (табл. 33). Модель, определяющая связь между X и Y, может быть основой для дисперсионного анализа переменной Y.

33. Таблица дисперсионного анализа

Источник	df	Сумма квадратов (SS)	Расчет
Общая	n-1	$SS_y = \sum(y_i - \bar{y})^2$	$\sum y_i^2 - \frac{(\sum y_i)^2}{n}$
Регрессия	1	$SS_{\hat{y}} = \sum(\hat{y}_i - \bar{y})^2$	$\hat{b}^2\left[\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right]$
Ошибка	n-2	$SS_e = \sum(y_i - \hat{y}_i)^2$	$SS_y - SS_{\hat{y}}$

Так, для зависимости молочности свиноматок от многоплодия (пример 15.5) таблица дисперсионного анализа:

Источник		df	Сумма квадратов,SS
Общая,	SS_y	7	252,0
Регрессия,	$SS_{\hat{y}}$	1	98,9
Остаток,	SS_e	6	153,3

Критерий соответствия регрессии опытным данным заложен в требовании метода наименьших квадратов:

$$SS_e = \sum(y_i - \hat{y}_i)^2 = \sum \hat{e}_i^2 \to \min .$$

Однако этот критерий имеет недостаток: при нижней границе равной нулю, верхняя граница не определена. Необходим показатель, который бы отражал, в какой мере функция регрессии определяется независимыми переменными. В качестве такого показателя интенсивности связи, или оценки доли влияния переменной X на Y, используют *коэффициент детерминации*:

$$R_{yx}^2 = \frac{SS_{\hat{y}}}{SS_y} = \frac{98,9}{252} = 0,39 \qquad \text{(пример 15.5).}$$

Коэффициент детерминации показывает, какая часть общего рассеяния значений Y обусловлена изменчивостью переменной X. Чем больше R_{yx}^2, тем лучше выбранная функция регрессии соответствует эмпирическим данным. И наоборот, чем меньше эмпирические значения Y отклоняются от прямой регрессии, тем лучше определена функция регрессии, тем больше R_{yx}^2. Субиндекс при R_{yx}^2 указывает на переменные, связь между

которыми изучается. При этом вначале стоит обозначение зависимой переменной, а затем - независимой, объясняющей переменной.

Коэффициент детерминации - величина безразмерная, не зависит от изменения единиц измерения Y и X и не реагирует на их преобразование. R^2_{yx} всегда находится в пределах от 0 до 1 (или от 0 до 100%). Если $R^2_{yx}=1$, то $y_i = \hat{y}_i$ и $\sigma^2_e=0$. В этом случае говорят о строгом линейном соотношении (линейной функции) между переменными X и Y.

Если $R^2_{yx}=0$, то $\hat{y}_i = \bar{y}$ и $\sigma^2_e=\sigma^2_y$. В этом случае линия регрессии параллельна оси абсцисс. Никакой численной линейной зависимости переменной Y от X в статистическом ее понимании нет. При этом коэффициент регрессии незначительно отличается от нуля. Таким образом, чем больше R^2_{yx} приближается к 1 (или к 100%), тем лучше определена регрессия.

Коэффициент детерминации можно выразить через коэффициент регресии:

$$R^2_{yx} = \hat{b}^2_{yx}\,\frac{\hat{\sigma}^2_x}{\hat{\sigma}^2_y} =$$

$$= \frac{\hat{\sigma}_{xy}}{\hat{\sigma}^2_x} \times \frac{\hat{\sigma}_{xy}}{\hat{\sigma}^2_x} \times \frac{\hat{\sigma}^2_x}{\hat{\sigma}^2_y} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}^2_x} \times \frac{\hat{\sigma}_{xy}}{\hat{\sigma}^2_y} =$$

$$= \hat{b}_{yx} \times \hat{b}_{xy} =$$

$$= 1{,}33 \times 0{,}294 = 0{,}39.$$

Из последнего выражения следует, что $R^2_{yx} = R^2_{xy} = R^2$.

Мерой неопределенности или неточности регрессионного анализа служит соотношение:

$$E = \frac{SS_e}{SS_y} = 1 - R^2.$$

Для **примера 15.5** R^2=0,39. Это указывает на то, что только 39% общей изменчивости молочности свиноматок было обусловлено вариацией многоплодия. Коэффициент неопределенности $E = 0{,}61$. Значит 61% общей изменчивости нельзя было объяснить зависимостью молочности от

многоплодия. Другими словами, бóльшая часть общей изменчивости молочности была вызвана неучтенными в регрессионном анализе факторами и различными случайными причинами.

15.7. Линейная множественная регрессия

Любой признак или явление детерминируется (определяется), как правило, множеством одновременно и совместно действующих причин. Поэтому одной из задач регрессионного анализа является исследование зависимости одной переменной Y от нескольких объясняющих или независимых переменных X_1, X_2, ... X_m в условиях конкретного места и конкретного времени. Эта задача решается с помощью *множественного* (мультифакторного) регрессионного анализа.

При наличии линейных соотношений между переменными, общее выражение уравнения множественной регрессии имеет вид:

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_m x_m + e,$$

где $b_1, b_2, ..., b_m$ - коэффициенты регрессии.

Функция линейной множественной регрессии есть:

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2 + ... + \hat{b}_m x_m .$$

$\hat{y}_i$ (i=1, …, n) - расчетные значения регрессии. Они указывают средние значения переменной Y в точке i при фиксированных значениях x_{ik} (k=0, …, m) - в предположении, что только эти m переменных являются причиной изменения переменной Y.

Коэффициенты b_k (k=0, …, m) - параметры регрессии. Константа регрессии b_0 выполняет в уравнении регрессии функцию выравнивания. Она определяет точку пересечения гиперповерхности регрессии с осью ординат.

Значения $\hat{b}_1, ..., \hat{b}_m$ есть оценки коэффициентов регрессии. Субиндекс при коэффициенте соответствует субиндексу независимой переменной. Так, $\hat{b}_1$ указывает среднюю величину изменения Y при изменении X_1 на одну единицу (при условии, что другие переменные остаются без изменения); $\hat{b}_2$ показывает, на сколько единиц в среднем изменится Y, если бы переменная

X_2 изменилась на единицу (при условии, что переменные X_k $(k \neq 2)$ оставались бы без изменения) и т.д. В то время как функция регрессии охватывает *совокупное одновременное* влияние независимых переменных, коэффициент регрессии $\hat{b}_k$ (k=1, 2, …, m) указывает соответствующие *усредненные частные* влияния переменной X_k в предположении, что остальные независимые переменные сохраняются на постоянном уровне.

Таким образом, с точки зрения статистической методологии нет различия между множественной и частной регрессией. Следовательно, при изучении регрессии нет необходимости различать частную и множественную регрессию. Поэтому в литературе параметры b_k (k=1, 2, …, m) называют как коэффициентами множественной, так и частной регрессии. Следует отметить, что множественная регрессия хотя и охватывает одновременное действие m независимых переменных, коэффициент регрессии $\hat{b}_k$ *исключает влияние остальных переменных-факторов* (при простой линейной регрессии влияние прочих неучтенных факторов частично отражается в коэффициенте регрессии).

Задача множественного регрессионного анализа состоит в оценке параметров регрессии по результатам выборочных наблюдений над переменными, включенными в анализ. Если для константы b_0 ввести фиктивную переменную $x_{i0} \equiv 1$, для всех i=1, 2, …n, то линейную модель множественной регрессии можно представить в виде

$$y = b_0 x_0 + b_1 x_1 + b_2 x_2 + ... + b_m x_m + e$$

или в матричной форме

$$y = Xb + e.$$

Для оценки неизвестных параметров вектора b, как и в случае с простой линейной регрессией, используют метод наименьших квадратов. Нормальные уравнения, которые удовлетворяют требованию о том, что сумма квадратов отклонений эмпирических значений от расчетных значений регрессии должна быть минимальна, имеют вид

$$X'X\hat{b} = X'y.$$

Если матрица $X'X$ обратима, то можно получить в качестве решения системы нормальных уравнений вектор-столбец искомых параметров регрессии:

$$\hat{b} = (X'X)^{-1} X'y .$$

Матрица $X'X$ и вектор $X'y$ имеют следующую структуру:

$$X'X = \begin{bmatrix} n & \sum x_{i1} & \cdots & \sum x_{im} \\ \sum x_{i1} & \sum x_{i1}^2 & \cdots & \sum x_{i1} x_{im} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_{im} & \sum x_{im} x_{i1} & \cdots & \sum x_{im}^2 \end{bmatrix};$$

$$X'y = \begin{bmatrix} \sum y_i \\ \sum x_{i1} y_i \\ \vdots \\ \sum x_{im} y_i \end{bmatrix}.$$

Пример 15.6. Пусть требуется исследовать зависимость переменной Y от трех независимых переменных X_1, X_2 и X_3 (n=14). Их числовые значения сведены в векторе-столбце **y** и в матрице **X**:

$$y = \begin{bmatrix} 20 \\ 24 \\ 28 \\ 30 \\ 31 \\ 33 \\ 34 \\ 37 \\ 38 \\ 40 \\ 41 \\ 43 \\ 45 \\ 48 \end{bmatrix}; \qquad X = \begin{matrix} x_0 & x_1 & x_2 & x_3 \end{matrix} \begin{bmatrix} 1 & 32 & 33 & 127 \\ 1 & 30 & 31 & 120 \\ 1 & 36 & 41 & 116 \\ 1 & 40 & 39 & 117 \\ 1 & 41 & 46 & 106 \\ 1 & 47 & 43 & 128 \\ 1 & 56 & 34 & 109 \\ 1 & 54 & 38 & 114 \\ 1 & 60 & 42 & 115 \\ 1 & 55 & 35 & 121 \\ 1 & 61 & 39 & 110 \\ 1 & 67 & 44 & 111 \\ 1 & 69 & 40 & 108 \\ 1 & 76 & 41 & 113 \end{bmatrix};$$

$$X' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 32 & 30 & 36 & 40 & 41 & 47 & 56 & 54 & 60 & 55 & 61 & 67 & 69 & 76 \\ 33 & 31 & 41 & 39 & 46 & 43 & 34 & 38 & 42 & 35 & 39 & 44 & 40 & 41 \\ 127 & 120 & 116 & 117 & 106 & 128 & 109 & 114 & 115 & 121 & 110 & 111 & 108 & 113 \end{bmatrix}.$$

Для $X'X$ и $X'y$ получим:

$$X'X = \begin{bmatrix} 14 & 724 & 546 & 1615 \\ 724 & 40134 & 28533 & 82884 \\ 546 & 28533 & 21544 & 62840 \\ 1615 & 82884 & 62840 & 186891 \end{bmatrix}; \quad X'y = \begin{bmatrix} 492 \\ 26907 \\ 19384 \\ 56389 \end{bmatrix}.$$

После обращения матрицы $X'X$ и умножения ее на вектор $X'y$ получим вектор оценок параметров регрессии:

$$\hat{b} = \begin{bmatrix} -52{,}88929 & -0{,}06869 & -0{,}26929 & -0{,}33603 \\ -0{,}06869 & 0{,}00052 & -0{,}00034 & 0{,}00048 \\ -0{,}26929 & -0{,}00034 & 0{,}00489 & 0{,}00083 \\ -0{,}33603 & 0{,}00048 & 0{,}00083 & 0{,}00242 \end{bmatrix} \begin{bmatrix} 492 \\ 26907 \\ 19384 \\ 56389 \end{bmatrix} = \begin{bmatrix} 5{,}05729 \\ 0{,}52123 \\ 0{,}15092 \\ -0{,}02389 \end{bmatrix}.$$

Таким образом, функция регрессии для рассматриваемой выборки данных имеет вид

$$\hat{y} = 5{,}05729 + 0{,}52123\,x_1 + 0{,}15092\,x_2 - 0{,}02389\,x_3$$

или

$$\hat{y} = 5{,}06 + 0{,}52\,x_1 + 0{,}15\,x_2 - 0{,}02\,x_3.$$

15.8. Стандартизация

В случае множественной регрессии рекомендуют преобразовывать переменные. Наиболее простым способом преобразования зависимой (Y) и независимых (X_k) переменных является *стандартизация* (нормирование):

$$y'_i = \frac{y_i - \bar{y}}{\hat{\sigma}_y}; \quad x'_{k_i} = \frac{x_{k_i} - \bar{x}_k}{\hat{\sigma}_{x_k}} \qquad (k = 1, ..., m),$$

где $\hat{\sigma}_y$, $\hat{\sigma}_{x_k}$ - стандартные отклонения переменных Y и X_k.

Все переменные и соотношения между ними выражают в стандартизованном масштабе. В этом масштабе за начало отсчета для каждой переменной принимают среднее значение, а за единицу измерения - величину стандартного отклонения. В стандартизованном масштабе упрощаются соотношения между переменными. При стандартизации фиктивную переменную (x_0), а вместе с ней и константу регрессии ($\hat{b}_0$) исключают, что способствует облегчению расчетов. Уравнение множественной линейной регрессии в стандартизованном масштабе:

$$\hat{y}' = \hat{b}'_1 x'_1 + \hat{b}'_2 x'_2 + ... + \hat{b}'_m x'_m$$

где $\hat{y}'$, x'_k (k=1, …, m) - стандартизованные переменные, а $\hat{b}'_k$ (k=1, …, m) - стандартизованные коэффициенты регрессии.

Оценки стандартизованных коэффициентов множественной регрессии также находят методом наименьших квадратов. Формулы аналогичны формулам обычных коэффициентов

регрессии (выраженных в натуральном масштабе), но с учетом того, что отсутствуют x_0 и b_0, и происходит замена переменных y_i на y'_i, а x_{ik} на x'_{ik}.

Стандартизованные коэффициенты регрессии $\hat{b}'_k$ можно вычислить также по натуральным коэффициентам регрессии $\hat{b}_k$. Между ними имеют место следующие соотношения:

$$\hat{b}'_k = \frac{\hat{\sigma}_{x_k}}{\hat{\sigma}_y}\hat{b}_k \quad \text{или} \quad \hat{b}_k = \frac{\hat{\sigma}_y}{\hat{\sigma}_{x_k}}\hat{b}'_k.$$

Стандартизованные переменные y'_i и x'_{ik}, а также стандартизованные коэффициенты регрессии $\hat{b}'_k$ *безразмерны*. Благодаря этому становится возможным сравнение *силы влияния* независимых переменных на зависимую переменную. Использовать для этой цели коэффициенты регрессии в натуральном масштабе ($\hat{b}_k$) нельзя из-за различной размерности переменных и коэффициентов регрессии, а также из-за различных средних значений переменных. Несмотря на небольшое по величине значение «натурального» коэффициента регрессии, соответствующая переменная может оказывать значительное влияние. Это объясняется различным рассеянием (вариацией) значений переменных x_k.

При стандартизации переменные выражают в единицах стандартных отклонений. В результате стандартные отклонения преобразованных переменных становятся равными единице. Стандартизованные коэффициенты множественной регрессии характеризуют скорость изменения среднего значения зависимой переменной по каждой из независимых переменных (при постоянных значениях остальных переменных, включенных в анализ).

Пример 15.7. Продолжим рассмотрение примера 15.6. Допустим, что известны значения коэффициентов регрессии и стандартных отклонений независимых переменных в натуральном масштабе:

$$\hat{b}_1 = 0{,}52123\,\text{ед.Y/ед.X}_1, \qquad \hat{\sigma}_{x_1} = 14{,}3925\,\text{ед.X}_1,$$
$$\hat{b}_2 = 0{,}15092\,\text{ед.Y/ед.X}_2, \qquad \hat{\sigma}_{x_2} = 4{,}3853\,\text{ед.X}_2,$$
$$\hat{b}_3 = -0{,}02389\,\text{ед.Y/ед.X}_3, \qquad \hat{\sigma}_{x_3} = 6{,}7323\,\text{ед.X}_3,$$
$$\hat{\sigma}_y = 8{,}0752\,\text{ед.Y}.$$

Тогда стандартизованные коэффициенты регрессии составят:

$$\hat{b}'_1 = 0{,}52123\,\frac{14{,}3925}{8{,}0752} = 0{,}92899,$$

$$\hat{b}'_2 = 0{,}15092\,\frac{4{,}3853}{8{,}0752} = 0{,}08196,$$

$$\hat{b}'_3 = -0{,}02389\,\frac{6{,}7323}{8{,}0752} = -0{,}01992.$$

Уравнение множественной регрессии в стандартизованном масштабе примет вид:

$$\hat{y}' = 0{,}92899\,x'_1 + 0{,}08196\,x'_2 - 0{,}01992\,x'_3.$$

В отличие от обычных коэффициентов регрессии, выраженных в натуральном масштабе, стандартизованные коэффициенты можно непосредственно сравнивать друг с другом. Стандартизованные коэффициенты множественной регрессии показывают, на какую часть стандартного отклонения изменилось бы среднее значение зависимой переменной, если бы значение соответствующей независимой переменной увеличилось на стандартное отклонение, а прочие переменные остались без изменения. Благодаря тому, что все переменные выражены в сравнимых единицах измерения, стандартизованные коэффициенты регрессии показывают относительную *силу влияния* каждой независимой переменной на изменение зависимой переменной. Так, в рассматриваемом примере значения стандартизованных коффициентов регрессии свидетельствуют о несущественном влиянии на зависимую переменную переменных-факторов X_2 и X_3. Наибольшее влияние на Y оказывает переменная-фактор X_1. При фиксированном значении X_2 и X_3, повышение X_1 на величину стандартного отклонения приводит в среднем к увеличению Y на 0,929 единиц стандартного отклонения. Аналогично интерпретируют стандартизованные коэффициенты регрессии $\hat{b}'_2$ и $\hat{b}'_3$.

15.9. Связь коэффициентов корреляции и регрессии

В регрессионном анализе существуют два коэффициента регрессии. Коэффициент же корреляции является общим мерилом сопряженной вариации двух признаков. Он более искусственен,

чем регрессия. При регрессии один признак выступает в качестве независимой переменной, другой - в качестве зависимой и наоборот. Эти зависимости имеют конкретный смысл.

Определим регрессию Y по X как:

$$b_{yx} = \frac{\sigma_{xy}}{\sigma_x^2}.$$

Умножим обе части уравнения на σ_x / σ_y:

$$\frac{\sigma_x}{\sigma_y} b_{yx} = \frac{\sigma_{xy}}{\sigma_x^2} \frac{\sigma_x}{\sigma_y}.$$

После сокращения правой часть на σ_x получим:

$$\frac{\sigma_x}{\sigma_y} b_{yx} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \quad \text{или} \quad \frac{\sigma_x}{\sigma_y} b_{yx} = r.$$

Тогда:

$$b_{yx} = r \frac{\sigma_y}{\sigma_x} \quad \text{соответственно} \quad b_{xy} = r \frac{\sigma_x}{\sigma_y}.$$

Следовательно, если коэффициент корреляции известен, то с помощью стандартных отклонений можно определить требуемый коэффициент регрессии.

Далее, если перемножить два коэффициента регрессии, то получим

$$b_{yx} \times b_{xy} = r^2.$$

Из этого отношения следует, что

$$r = \pm\sqrt{b_{yx}\, b_{xy}} = \pm\sqrt{1{,}33 \times 0{,}294} = \pm 0{,}63$$

$$= \pm\sqrt{\frac{\sigma_{xy}}{\sigma_x^2} \frac{\sigma_{xy}}{\sigma_y^2}} =$$

$$= \pm\frac{\sigma_{xy}}{\sigma_x \sigma_y} = \pm\frac{10{,}57}{2{,}8 \times 6} = \pm 0{,}63,$$

т.е. математически коэффициент корреляции представляет собой среднюю геометрическую из двух коэффициентов регрессии.

Следует отметить, что для коэффициента регрессии Y по X имеют место соотношения:

$$R^2 = b_{yx}^2 \frac{\sigma_x^2}{\sigma_y^2} =$$

$$= \left(r \frac{\sigma_y}{\sigma_x} \right)^2 \frac{\sigma_x^2}{\sigma_y^2} =$$

$$= r^2.$$

Таким образом, квадрат коэффициента корреляции равен коэффициенту детерминации.

Как в формуле для коэффициента корреляции, так и в формулах коэффициентов регрессии центральное место занимает *ковариансa*. Она, в сущности, настоящее мерило сопряженной вариации анализируемых переменных. Поэтому ковариансa является связующим звеном в корреляционном и регрессионном анализах.

15.10. Нелинейная регрессия

Между явлениями и процессами не всегда существует линейная зависимость. Часто имеющиеся соотношения нельзя выразить линейными функциями из-за возникающих при этом неоправданно больших ошибок. В таких случаях для описания зависимостей используют нелинейную регрессию (и корреляцию).

В зависимости от характера связи различают: (1) положительную равноускоренно и равнозамедленно возрастающую регрессию; (2) отрицательную равноускоренно и равнозамедленно убывающую регрессию и (3) их комбинированные формы.

Для выбора и обоснования типа кривой регрессии нет универсального метода. Односторонняя стохастическая зависимость между явлениями может быть описана, например, с помощью полиномиальной регрессии:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x + \hat{b}_2 x^2 + \hat{b}_3 x^3 + ...,$$

либо с помощью гиперболической регрессии:

$$\hat{y}^* = \hat{b}_0^* + \hat{b}_1^* \frac{1}{x}.$$

Применяют также степенную, показательную, логарифмическую и тригонометрическую функции. Подбор

функции регрессии производят с применением теории по той конкретной проблеме, на базе которой возникает задача измерения связи между явлениями. Чаще всего используют семейства кривых, уравнения которых выражают многочленами целых положительных степеней (полиномами). Полином первой степени (прямая линия) не имеет изгибов. С помощью полинома второй степени можно передать одну точку поворота функции. Полином третьей степени отражает две точки поворота функции.

О характере зависимости между явлениями часто судят по внешнему виду эмпирического графика регрессии. Однако при малом числе наблюдений этот путь приводит к неудовлетворительным результатам, так как резкие зигзаги эмпирической линии регрессии затрудняют выявление закономерности. В каждом случае следует проверять возможность применения линейной регрессии хотя бы на ограниченном участке изменения переменных.

Различают два класса нелинейных регрессий. К первому классу относят регрессии, нелинейные относительно включенных в анализ независимых переменных (X_k), но линейные по неизвестным, подлежащим оценке, параметрам регрессий ($\hat{b}_k$, k=1, 2, ... p). Образующие этот класс регрессии называют *квазилинейными*. Для них возможно непосредственное применение LS-метода. Используют те же самые критерии значимости, аналогично строят доверительные интервалы.

Второй класс регрессий характеризуется нелинейностью по оцениваемым параметрам. Основной недостаток - нельзя использовать LS-метод. Для решения системы нелинейных уравнений привлекают или итерационные методы, или прибегают к аппроксимации параметров искомой зависимости. Широко используют также линейное преобразование функции регрессии, которое позволяет применять к преобразованным параметрам статистические критерии линейной регрессии.

Пример 15.8. Известно, что рост организмов (или рост популяций организмов) во многих случаях происходит таким образом, что прибавка в весе растущего организма во всякий момент времени пропорциональна уже достигнутому весу. Иллюстрацией могут быть следующие данные по

изменению сухого веса куриных эмбрионов от 6- до 16-дневного возраста и логарифмы этого веса:

Возраст, дней, X	Сухой вес, г, W	Логарифм веса, У
6	0,029	-1,538
7	0,052	-1,284
8	0,079	-1,102
9	0,125	-0,903
10	0,181	-0,742
11	0,261	-0,583
12	0,425	-0,372
13	0,738	-0,132
14	1,130	+0,053
15	1,882	+0,275
16	2,812	+0,445

Если представить данные второго столбца на графике (рис. 28), то можно видеть, что возрастание веса происходит значительно быстрее, чем возраста.

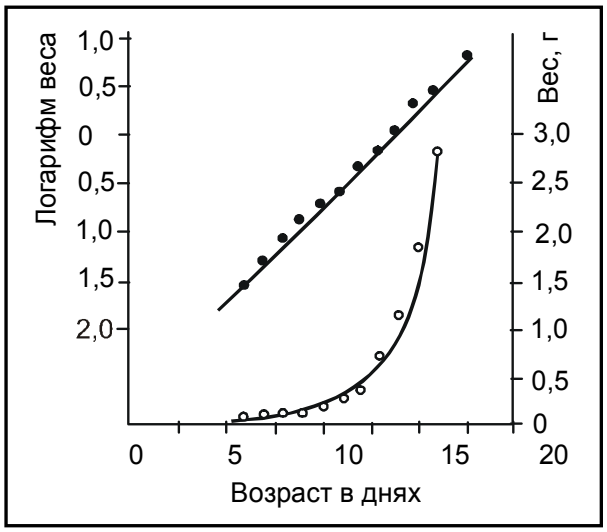


Рис. 28. Изменение сухого веса куриных эмбрионов с возрастом (нижняя кривая линия) и изменение логарифмов веса эмбрионов с возрастом (верхняя прямая линия). На абсциссе логарифмы веса вверх от нуля положительны, вниз - отрицательны

Такая кривая называется *экспоненциальной*; уравнение регрессии для нее выражается формулой:

$$W = A \times B^{x},$$

где A и B - определенные константы.

Однако логарифмируя это уравнение, его можно привести к форме уравнения прямой:

$$\log W = \log A + (\log B)^x .$$

Здесь $\log W$ соответствует *y*, $\log A$ соответствует *a* и $\log B$ соответствует *b* в обычном линейном уравнении:

$$y = a + bx + e .$$

Третий столбец таблицы - есть логарифмы веса эмбрионов с возрастом, изображение которых на рис. 28 дает прямую линию.

Оценки параметров регрессии определяют обычной процедурой. Функция регрессии будет иметь вид

$$\hat{y} = -2{,}689 + 0{,}1959x .$$

15.11. Последовательность регрессионного анализа

В обобщенном виде процедура регрессионного анализа включает следующие этапы.

1. *Формулировка проблемы*. Подразумевает конкретизацию биозоотехнических явлений и процессов, зависимость между которыми подлежит оценке.
2. *Идентификация переменных*. На основе профессионально-теоретических соображений и биологического смысла определяют разумное число переменных, производят их классификацию на зависимые и независимые.
3. *Сбор данных*. Исходя из цели и задач, устанавливают принцип отбора данных и объем выборки. Если для каких-либо явлений не может быть обеспечен необходимый объем данных, то следует вернуться к первому этапу.
4. *Спецификация уравнения регрессии* (параметризация модели). На данном этапе:
   - формулируют гипотезы о форме связи (линейная или нелинейная, простая или множественная) и
   - проверяют предпосылки.

   Большей частью вид уравнения регрессии в процессе исследования определяют поэтапно путем исключения переменных, не оказывающих существенного влияния на зависимую переменную, и включения в анализ новых факторов-причин с проверкой их значимости.
5. *Оценка функции регрессии*. Определяют численные значения параметров уравнения регрессии.

6. *Оценка точности регрессионного анализа*. Вычисляют статистические показатели, характеризующие точность регрессионного анализа.
7. *Биозоотехническая интерпретация результатов*. Результаты сравнивают с гипотезами, оценивают их правдоподобие с биологической и/или зоотехнической точек зрения.
8. *Прогноз неизвестных значений зависимой переменной*. Полученную функцию регрессии используют для прогностического анализа.

Если определена функция регрессии и она биологически обоснована, то прогностические (теоретические) оценки обладают достаточной надежностью. По своей сути они являются средними значениями, которые следует ожидать с бóльшей вероятностью. В силу многофакторности биологических явлений и многогранности их выражений отдельные фактические (эмпирические) значения рассеиваются вокруг средних значений. Поэтому естественно, что фактические значения зависимой переменной не будут совпадать с расчетными, т.е. с прогнозом. С этим необходимо считаться. Степень рассеяния наблюдений вокруг теоретической линии регрессии характеризует надежность получаемых по уравнению регрессии прогностических оценок.

Точность прогноза определяется не только точностью полученных оценок параметров регрессии, но и тем, насколько надежно оценены будущие значения независимых переменных на основе дополнительной информации. Источником такой дополнительной информации могут быть профессионально-теоретические соображения в соответствии с зоотехнической, племенной, экономической и даже социальной политикой хозяйства, региона, государства. Поэтому процесс построения статистической модели должен сопровождаться корректировкой оценок параметров регрессии и статистических характеристик в соответствии с ожидаемыми изменениями обстоятельств их формирования.

Прогнозирование результатов по регрессии лучше поддается содержательной интерпретации, чем простая экстраполяция тенденций, т.к. можно полнее учитывать природу исследуемого явления. Благодаря этому регрессионный анализ широко применяют при решении задач перспективного планирования.

Дополнение

Предпосылки регрессионного анализа

Основные предпосылки регрессионного анализа:

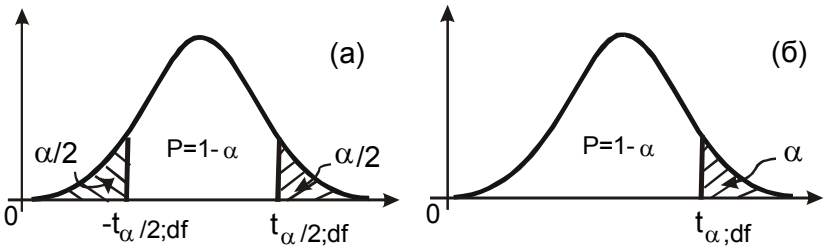
1. Возмущение (остаток) e_i (или зависимая переменная y_i) есть величина случайная, а объясняющая переменная x_i - величина неслучайная.
2. Математическое ожидание e_i равно нулю:
$$E(e_i)=0 \quad \text{или} \quad E(y_i)=b_0+b_1x_i.$$
3. Варианса возмущения e_i (или y_i) постоянна для любого i (неучтенные факторы оказывают одинаковое влияние):
$$Var(e_i)=\sigma_e^2 \ldots\ldots \text{или} \quad Var(y_i)=\sigma^2.$$
4. Возмущения e_i и e_j (или y_i и y_j) не коррелированы:
$$E(e_i,e_j)=0 \quad \text{при } i\neq 0.$$
5. Возмущение e_i (или y_i) есть нормально распределенная случайная величина.
6. Число наблюдений должно превышать число параметров, иначе невозможна их оценка.
7. Объясняющие переменные X не должны коррелировать с возмущающей переменной *e*, т.е.
$$E(x_{ik},e_i)=0 \quad \text{при } (k=1, \ldots, m).$$
8. Переменные x_k объясняют переменную *y*, но *y* не объясняет x_k, т.е. предполагается односторонняя зависимость *y* от x_k и отсутствие взаимосвязи.

Для получения уравнения регрессии достаточно первых четырех предпосылок. Требование выполнения пятой предпосылки необходимо для оценки точности уравнения регрессии и ее параметров.

Приложение

А.8. Критические значения t-распределения Стьюдента

(здесь и далее P - доверительная вероятность)



df	Уровень значимости (ошибка, α)				
	Двусторонняя критическая область (а)				
	0,100	**0,050**	**0,020**	**0,010**	**0,001**
	Односторонняя критическая область (б)				
	0,050	**0,025**	**0,010**	**0,005**	**0,0005**
1	6,314	12,706	31,821	63,657	637
2	2,920	4,303	6,965	9,925	31,598
3	2,353	3,182	4,541	5,841	12,941
4	2,132	2,776	3,747	4,604	8,610
5	2,015	2,571	3,365	4,032	6,859
6	1,943	2,447	3,143	3,707	5,959
7	1,895	2,365	2,998	3,499	5,405
8	1,860	2,306	2,896	3,355	5,041
9	1,833	2,262	2,821	3,250	4,781
10	1,812	2,228	2,764	3,169	4,587
11	1,796	2,201	2,718	3,106	4,437
12	1,782	2,179	2,681	3,055	4,318
13	1,771	2,160	2,650	3,012	4,221
14	1,761	2,145	2,624	2,977	4,140
15	1,753	2,131	2,602	2,947	4,073
20	1,725	2,086	2,528	2,845	3,850
25	1,708	2,060	2,485	2,787	3,725
30	1,697	2,042	2,457	2,750	3,646
35	1,690	2,030	2,432	2,724	3,591
40	1,684	2,021	2,408	2,704	3,551
50	1,676	2,008	2,384	2,678	3,496
100	1,661	1,982	2,360	2,625	3,390
∞	1,645	1,960	2,326	2,576	3,291

Примечание. В последней строке даны значения нормированной случайной величины $t = u \sim N(0;1)$.