Лекция № 19 (продолжение)

Пример. Построить уравнение регрессии для зависимости количества пропущенных занятий и итоговой оценки, табл. 3 лекция 16.

Решение. Аналогично предыдущему примеру, расчет вспомогательных величин осуществляется в таблице 1, после чего по формуле (*) вычисляются коэффициенты уравнения:

$$a = \frac{511 \cdot 579 - 57 \cdot 3745}{7 \cdot 579 - (57)^2} = 102.43,\ b = \frac{7 \cdot 3745 - 57 \cdot 511}{7 \cdot 579 - (57)^2} = -3.622.$$

Уравнение регрессии для зависимости между числом пропусков и итоговым баллом получено, рис. 1:

$$y' = a + bx = 109.493 - 3.622 \cdot x$$

Таблица 1. Расчет коэффициентов уравнения регресии

Число пропусков x_i	*Суммарный балл* y_i	$x_i y_i$	x_i^2
6	82	492	36
2	86	172	4
15	43	645	225
9	74	666	81
12	58	696	144
5	90	450	25
8	78	624	64
$\sum_{i=1}^{7} x_i = 57$	$\sum_{i=1}^{7} y_i = 511$	$\sum_{i=1}^{7} x_i y_i = 3745$	$\sum_{i=1}^{7} x_i^2 = 579$

Модель линейной регрессии

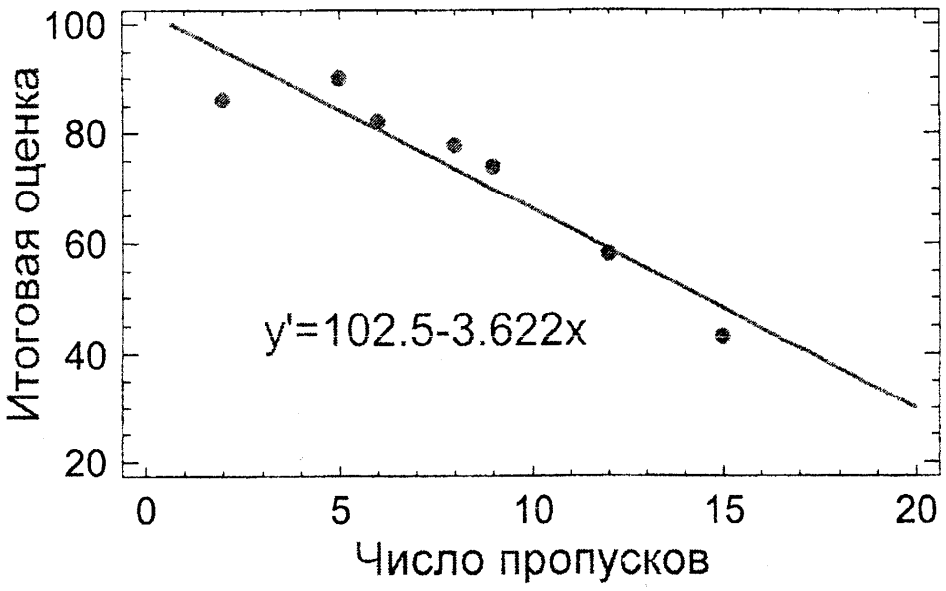

Рис. 1 уравнение линейной регрессии для количества пропусков и итоговой оценки

Одной из основных целей построения уравнения регрессии является прогнозирование. Например, с помощью построенной зависимости давления от возраста найдем, что в возрасте $x = 50$ *лет* расчетное давление составит $y' = a + bx = 81.048 + 0.964 \cdot 50 \approx 130$. Как соотнести полученный результат с реальностью? Интуитивно понятно, что, выбрав произвольного человека 50 лет и измерив его давление, мы не получим значение 130. Ответ на этот и связанные с ним вопросы будет дан в следующих трех пунктах, где будет рассмотрен:

1) Анализ точности модели с помощью коэффициента детерминации.

2) Анализ достоверности (значимости) построенной по выборке модели для генеральной совокупности в целом.

3) Построение интервального прогноза - диапазона значений зависимой переменной, в которой истинное значение попадет с заданной вероятностью.

Анализ точности модели — коэффициент детерминации

При нахождении регрессионного уравнения было отмечено, что его коэффициенты находятся из условия минимизации суммы квадратов остатков. Однако насколько велико это минимальное значение? На рис. 2а и

2б представлены диаграмма рассеяния и регрессионная модель для двух величин x, y. При этом данные на рис. 2а характеризуются коэффициентом корреляции $r = 0.7$, а на рисунке 2б $r = 0.95$. Нетрудно понять, что сумма отклонений значений y от линии регрессии во втором случае будет меньше. Соответственно, и качество модели, и точность прогноза с возрастанием r должно увеличиваться.

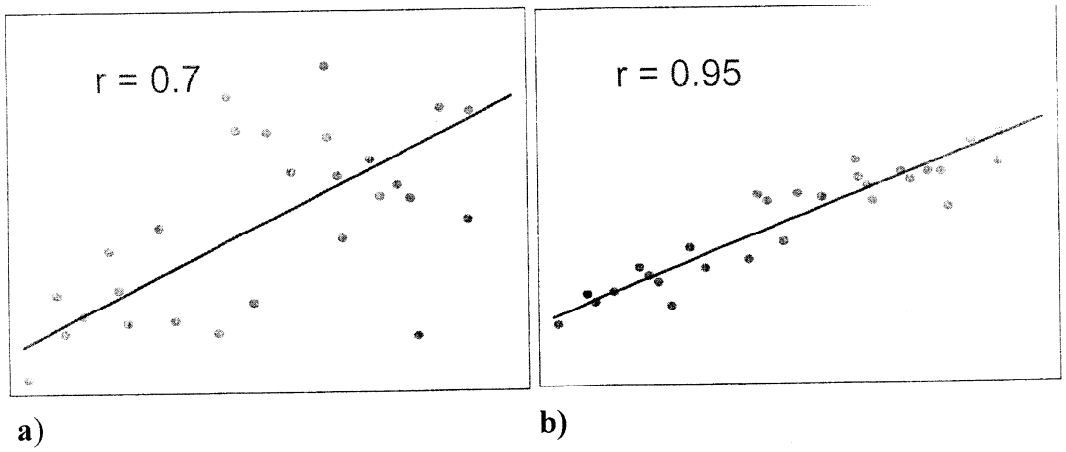

Рис. 2 уравнения регрессии для зависимостей с различными коэффициентами корреляции

Рассмотрим выборочные данные (x_i, y_i), $i = 1,\ldots,n$ $\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, построим уравнение регрессии $y' = a + bx$ и для каждого x_i, рассчитаем значение $y'_i = a + bx_i$, рис. 3.

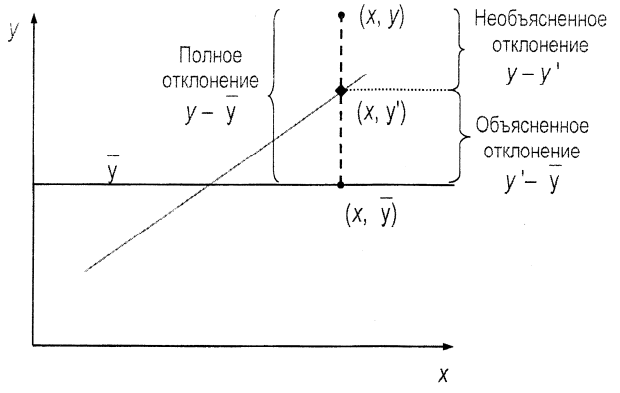

Рис. 3 Разложение отклонения от линии регрессии

Рассмотрим точку i и представим разность $y_i - \overline{y}$ в виде суммы:

$$y_i - \overline{y} = \left(y_i' - \overline{y}\right) + \left(y_i - y_i'\right)$$

Три выражения в скобках имеют специальные названия и широко используются в регрессионном анализе, табл. 2, колонки 1, 2.

Таблица 2. Расчёт величин, характеризующих регрессионную модель

Величина	Название	Величина	Название
$y_i - \overline{y}$	Полное отклонение	$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2$	Дисперсия
$y_i' - \overline{y}$	Объясненное отклонение	$s_{y'}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i' - \overline{y}\right)^2$	Объясненная дисперсия
$y_i - y_i' = e_i$	Необъясненное отклонение, остаток	$s_e^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - y_i'\right)^2$	Необъясненная дисперсия, дисперсия остатков

Возводя обе части в квадрат и выполняя суммирование по i, получаем:

$$\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 = \sum_{i=1}^{n}\left(y_i' - \overline{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - y_i'\right)^2 + 2\sum_{i=1}^{n}\left(y_i' - \overline{y}\right)\left(y_i - y_i'\right)$$

Пользуясь определением y_i' и $\overline{y}$, можно показать, что последняя сумма справа равна нулю, поэтому соотношение принимает вид:

$$\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2 = \sum_{i=1}^{n}\left(y_i' - \overline{y}\right)^2 + \sum_{i=1}^{n}\left(y_i - y_i'\right)^2,$$

то есть сумма квадратов полных отклонений складывается из суммы квадратов объясненных отклонений и суммы квадратов остатков. Разделив каждую из сумм на число единиц выборки, табл. 2, колонки 3,4, можно получить аналогичное соотношение для дисперсий:

Дисперсия = Объясненная дисперсия + Необъясненная дисперсия.

Пример. В табл. 3 приведена выборочная совокупность для величин x, y. Данные характеризуются значимым коэффициентом корреляции $r = 0.919$ и уравнением регрессии $y' = 4.8 + 2.8 \cdot x$. Рассчитать суммы квадратов полных, объясненных и необъясненных отклонений.

Таблица 3. Расчет отклонений для уравнения регрессии

x_i	y_i	$y' = 4.8 + 2.8 \cdot x$	$(y_i - \overline{y})^2$	$(y_i' - \overline{y})^2$	$(y_i - y_i')^2$
1	10	4.8+2.8-1=7.6	$(10-13.2)^2$= 10.24	$(7.6-13.2)^2$=31.36	$(10-7.6)^2$=5.76
2	8	4.8+2.8-2=10.4	$(8-13.2)^2$=27.04	$(10.4-13.2)^2$=7.84	$(8-10.4)^2$=5.76
3	12	4.8+2,8-3=13.2	$(12-13.2)^2$= 1.44	$(13.2-13.2)^2$= 0	$(12-13.2)^2$= 1.44
4	16	4.8+2.8-4=16	$(16-13.2)^2$= 7.84	$(16-13.2)^2$= 7.84	$(16-16)^2$=0
5	20	4.8+2.8-5=18.8	$(20-13.2)^2$=46.24	$(18.8-13.2)^2$=31.36	$(20-18.8)^2$=1.44
Сумма			$\sum_{i=1}^{5}(y_i - \overline{y})^2 = 92.$	$\sum_{i=1}^{5}(y_i' - \overline{y})^2 = 78.4$	$\sum_{i=1}^{5}(y_i - y_i')^2 = 1$

Решение.

1) Вычислить для каждого x_i, прогнозируемые значения y_i', табл. 3, колонка 3.

2) Определить среднее значение $\overline{y} = \frac{1}{5}(10+8+12+16+20) = 13.2$ y = (10 + 8+12 + 16 + 20)/5 = 13.2.

3) Вычислить квадраты полного отклонения $(y_i - \overline{y})^2$ в каждой точке и просуммировать, колонка 4.

4) Вычислить квадраты объясненных отклонений $(y_i' - \overline{y})^2$ в каждой точке и просуммировать, колонка 5.

5) Вычислить квадраты необъясненных отклонений (остатков) $(y_i - y_i')^2$ в

каждой точке и просуммировать, колонка 6.

Легко проверить, что условие $\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(y_i'-\bar{y})^2+\sum_{i=1}^{n}(y_i-y_i')^2$ выполняется:

Представленные определения позволяют выписать выражение для *коэффициента детерминации:*

$$r^2=\frac{\sum_{i=1}^{n}(y_i'-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=\frac{s_{y'}^2}{s_y^2}$$

который равен отношению объясненной дисперсии к полной. Коэффициент детерминации является основной характеристикой регрессионной модели и показывает, какую долю вариации (изменчивости) результативного признака можно объяснить изменением факторного признака. Для рассмотренного выше примера коэффициент детерминации равен

$r^2=\frac{\sum_{i=1}^{n}(y_i'-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=\frac{s_{y'}^2}{s_y^2}=\frac{78.4}{92.8}=0.845$ $\varepsilon^2 = 78.4/92.8 = 0.845.$

Величина $1-r^2$ называется *коэффициентом недетерминации.*

Использование обозначения r коэффициента линейной корреляции в коэффициенте детерминации r^2 не случайно. Пользуясь свойствами корреляции и дисперсии, можно показать, что определенный выше как отношение квадратов отклонений коэффициент r^2 равен коэффициенту корреляции, возведенному в квадрат. В частности, в приведенном примере коэффициент корреляции равен $r=0.919$ и $r^2=0.919^2=0.845$, что совпадает с коэффициентом детерминации, рассчитанным по определению. С учетом этого коэффициент детерминации на практике обычно вычисляется

через коэффициент корреляции. При уменьшении r коэффициент детерминации быстро убывает. Так, при $r = 0.6$ имеем $r^2 = 0.36$, то есть регрессионная модель объясняет только 36% вариации результативного признака.

Одним из практических применений коэффициента детерминации является оценка качества и сравнение между собой различных моделей (линейной и нелинейных) парной регрессии.

Стандартные ошибки. Помимо коэффициента детерминации, качество регрессионной модели характеризуют *стандартные ошибки коэффициентов:*

$$s_{er}(a) = \frac{s_e}{\sqrt{n-2}},\ s_{er}(b) = \frac{s_e}{\sqrt{n-2}s_x},$$

и *стандартная ошибка модели*:

$$s_{er}(y') = \sqrt{\left(1 + \frac{2}{n-2}\right)s_e}$$

где $s_e^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - y_i')^2$ дисперсия остатков, $s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ - дисперсия независимой величины x. Стандартные ошибки коэффициентов используются для проверки значимости коэффициентов модели, а стандартная ошибка модели - для построения интервального прогноза. Все три величины пропорциональны значению s_e - величине “среднего остатка”, но ошибки коэффициентов убывают с увеличением длины выборки, а ошибка модели остается практически неизменной и равной s_e. При использовании большинства программ статистического анализа стандартные ошибки коэффициентов и модели рассчитываются одновременно с расчетом самой модели.

Исследование значимости регрессионной модели

До сих пор при рассмотрении регрессионной модели мы оперировали

набором значений (x_i, y_i) одной единственной выборки. Можно ли и с какой точностью перенести полученные результаты на генеральную совокупность? Этот вопрос в статистике решается с помощью проверки гипотез о значимости коэффициентов и модели в целом. Схема рассуждений во многом является аналогичной проверке гипотез о значимости коэффициента корреляции.

Формулировка гипотез о значимости коэффициентов. До сих мы рассматривали уравнение регрессии, построенное по выборке $y' = a + bx$. Для генеральной совокупности уравнение линейной зависимости между x и y запишем в виде $y = A + Bx + u$, где u - случайный член, соответствующий всем факторам, не учтенным в модели. Гипотезы о коэффициентах регрессионной модели формулируются следующим образом:

Коэффициент	Основная гипотеза	Альтернативная гипотеза
A	$H_0 : A = 0$	$H_1 : A \neq 0$
B	$H_0 : B = 0$	$H_1 : B \neq 0$

Принятие основной гипотезы означает, что выборочный коэффициент незначимый и в генеральной совокупности этот коэффициент предположительно равен нулю. В противоположном случае, когда принимается альтернативная гипотеза, отличие от нуля полагается значимым.

Неформальную интерпретацию гипотез о значимости коэффициентов рассмотрим на примере простейшей функции спроса $y = A + Bx + u$, где y - величина спроса, скажем, на продукты питания, x -доход, u - случайный член. Исходя из вполне разумных теоретических оснований, можно предположить, что спрос на продукты зависит от дохода, но указать при этом конкретное значение, например, для B затруднительно. В этом случае можно в качестве в основной гипотезы взять утверждение, что y *не влияет*

на x, $H_0 : B = 0$. Тогда альтернативная гипотеза $H_1 : B \neq 0$, то есть x *влияет на* y.

Схема проверки гипотез о значимости коэффициентов.

Проверка гипотез о значимости коэффициентов производится путем их сравнения с величиной стандартной ошибки: $t = t(b) = \dfrac{b}{s_{er}(b)}$, $t = t(a) = \dfrac{a}{s_{er}(a)}$ Эти величины подчиняются распределению Стьюдента с числом степеней свободы $k = n - 2$ при выполнении следующих двух условий.

1) Величина остатков для всех наблюдений должна иметь нулевое математическое ожидание и одинаковую дисперсию.

2) Остатки для всех наблюдений должны иметь приближенно нормальное распределение.

Эти утверждения требуют пояснения. В них говориться о распределении остатков в каждом наблюдении, тогда как исследователь имеет в своем распоряжении одну выборку, одну регрессионную модель и одно значение остатка. Термин "распределение остатков" в этом случае следует трактовать как *возможное* поведение остатков *до того,* как сделана выборка. Фактические значения остатков e_i для конкретной выборки будут иногда больше, иногда меньше по величине, иногда положительными, иногда отрицательными, но нет причин ожидать, например, что вероятность появления больших остатков при каком-то одном i будет больше, чем при любом другом. Сложность исследования остатков типична при построении регрессионных моделей в экономике, так как возможность получения повторных выборок y_i для одной и той же совокупности значений x_i чаще всего отсутствует.

Пример. На основании данных наблюдений в США за 25-летний срок (1959-1983 годы) построена зависимость суммарных расходов на питание y

от располагаемых доходов x: $y' = 55.3 + 0.093 \cdot x$, $s_{er}(a) = 2.4$, $s_{er}(b) = 0.003$. На уровне значимости 5% проверить гипотезы о значимости коэффициентов.

1) Гипотезы для обоих коэффициентов формулируются одинаково:

$H_0 : A = 0;\ H_1 : A \neq 0.$

$H_0 : B = 0;\ H_1 : B \neq 0$

2) Критические значения для распределения Стьюдента с 23 степенями свободы равны $t_{\alpha/2} = 2.069$ С, критическая область $|t| > 2.069$.

3) Определить расчетные значения критерия:

$$t = t(b) = \frac{b}{s_{er}(b)} = \frac{0.093}{0.003} = 31,\ t = t(a) = \frac{a}{s_{er}(a)} = \frac{55.3}{2.4} = 23.04.$$

4) Принятие решения. Расчетные значения критерия для обоих коэффициентов превышают критическое значение:

$t(b) = 31 > t_{\alpha/2} = 2.069,\ t(a) = 23.04 > t_{\alpha/2} = 2.069,$

поэтому основные гипотезы отклоняются.

5) Вывод. Оба коэффициента регрессионной модели значимо (не случайно) отличаются от нуля.

Замечание. Зная выборочные значения коэффициентов, величину критерия и его распределение, можно построить для коэффициентов интервальные оценки (доверительные интервалы). Например, с доверительной вероятностью $p_0 = 1 - \alpha$ коэффициент B имеет значение:

$$b - s_{er}(b) \cdot t_{\alpha/2} < B < b + s_{er}(b) \cdot t_{\alpha/2}$$

Подставляя значения из примера, получаем:

$0.093 - 0.003 \cdot 2.069 < B < 0.093 + 0.003 \cdot 2.069$ или

$B \in [0.093 \mp 0.0062]$.