Обнаружение выбросов и влиятельных наблюдений в регрессионном анализе (на примере линейной регрессии)

Черновик

Нохрин Д.Ю. (*forum.disser.ru, nokh*)

I. Исходные данные

Каждому значению количественного показателя x соответствует значение количественного показателя y. Величина x контролируется экспериментатором или ошибка x пренебрежимо мала по сравнению с ошибкой y. Показатель y имеет нормальное распределение. Требуется обнаружить возможные выбросы и определить влияющие наблюдения в предположении линейной зависимости y от x вида: $y=a+bx$.

Обозначим через n – количество пар x-y (объём выборки), а через k – количество параметров в модели, включая свободный член. Для нашего примера n=10, k=2.

Таблица 1. Исходные данные

i	1	2	3	4	5	6	7	8	9	10
x_i	1	2	3	3	4	5	6	7	8	12
y_i	4	4	4	8	5	5	6	6	7	10

II. Описательная статистика, уравнение регрессии, оценка статистической значимости регрессионной модели и её параметров

Большинство компьютерных программ выдают все перечисленные в заголовке показатели, поэтому приведу только график регрессии и результаты дисперсионного анализа. Точка $O(\overline{x};\overline{y})$ – центр системы (в множественной регрессии центр системы называют центроидом).

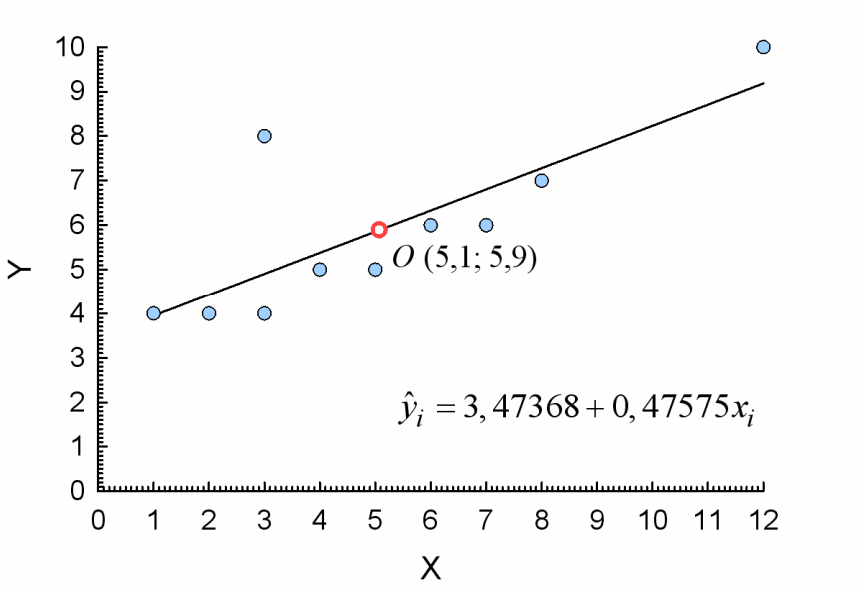

Таблица 2. Результаты дисперсионного анализа линейной регрессии

Источник изменчивости	Сумма квадратов *SS*	Степени свободы *v*	Средний квадрат *MS*	*F*-критерий	*P*
Линейная регрессия	21,93199174	1	21,93199174	13,529906	0,00623439
Отклонения от линейной регрессии (ошибка)	12,96800826	8	**1,621001032**		
Общая	34,9	9			

II. Подготовительные расчёты

Статистические пакеты, содержащие продвинутые модули регрессионного анализа, включают в себя анализ остатков и влияющих наблюдений. Однако набор этих показателей в разных пакетах отличается. Более того, под одинаковыми названиями могут фигурировать различные статистические процедуры, что отражает существующую терминологическую противоречивость нескольких наиболее авторитетных литературных источников, которыми руководствуются составители программ. Такие случаи ниже специально отмечены. При подготовке раздела компиляция формул из разных источников была проведена с изменением по необходимости буквенных обозначений для обеспечения внутреннего единства материала. Все представленные ниже ручные расчёты совпадают с результатами, выдаваемыми пакетами Statistica (v.8.0, StatSotf Inc.) и SPSS (PASW Statistics, v.18.0), с точностью до ошибок округления в последнем знаке.

(1). Ожидаемое (Predicted) значение

– значение, ожидаемое в случае отсутствия изменчивости зависимой переменной *y*. Рассчитываются по теоретической регрессионной зависимости *y* от независимых предикторов (в нашем случае – от *x*).

$\hat{y}_i = a + bx_i$ В нашем случае: $\hat{y}_i = 3,47368 + 0,47575x_i$ и, например,

$$\hat{y}_1 = 3,47368 + 0,47575 \cdot 1 = 3,94943 .$$

Таблица 3. Ожидаемые значения

i	1	2	3	4	5	6	7	8	9	10
$\hat{y}_i$	3,94943	4,42518	4,90093	4,90093	5,37668	5,85243	6,32817	6,80392	7,27967	9,18266

(2). Остаток (Residual)

– разность между наблюдаемым и ожидаемым значением *y* при данном *x*. Сумма всех остатков равна нулю. Анализ распределения остатков позволяет обнаружить отклонение исходных данных от нормального распределения, а также предположить вид нормализующего преобразования. Поскольку в нашем случае переменная *y* по условию имела нормальное распределение, распределение остатков также будет нормальным.

$r_i = y_i - \hat{y}_i$ Например: $r_1 = 4 - 3,94943 = 0,05057$.

Таблица 4. Остатки

i	1	2	3	4	5	6	7	8	9	10
r_i	0,05057	-0,42518	-0,90093	3,09907	-0,37668	-0,85243	-0,32817	-0,80392	-0,27967	0,81734

(3). Удалённый остаток (Deleted residual)

Если в наборе данных присутствует один или несколько выбросов, то в зависимости от своего положения они могут сильно искажать регрессионную зависимость и т.о. препятствовать своему обнаружению. Идея расчета удалённых остатков состоит в том, чтобы получить уравнение регрессии без влияния на него потенциального выброса. Для этого *i*-тое наблюдение удаляется – записывается как (*i*), – по оставшимся (*n*-*i*) наблюдениям строится уравнение регрессии, и рассчитывается ожидаемое значение для *i*-того наблюдения $\hat{y}_{i(i)}$. Разность между наблюдаемым значением и ожидаемым в отсутствие данного наблюдения значением и является удалённым остатком. По абсолютному значению удалённые остатки всегда больше обычных остатков.

$r_{(i)} = y_i - \hat{y}_{i(i)}$ Например, если удалить из набора данных наблюдение № 4, то уравнение регрессии изменится на: $\hat{y}_{i(4)} = 2,71014 + 0,55435x_i$. Согласно ему, ожидаемое значение для наблюдения № 4 составит $\hat{y}_{4(4)} = 2,71014 + 0,55435 \cdot 3 = 4,37319$, а удалённый остаток $r_{(4)} = 8 - 4,37319 = 3,62681$.

Таблица 5. Удалённые остатки

i	1	2	3	4	5	6	7	8	9	10
$r_{(i)}$	0,06960	-0,53093	-1,05435	3,62681	-0,42442	-0,94725	-0,36806	-0,93182	-0,34391	1,99999

(4). Дисперсия ошибки регрессии

– мера разброса значений относительно линии регрессии. В результатах дисперсионного анализа регрессии она представляет собой средний квадрат ошибки ($MS_{\text{ошибки}}$) и может быть взята оттуда (выделена жирным шрифтом в Табл. 2) или рассчитана по формуле:

$$s_e^2 = \frac{\sum r_i^2}{n-k} \qquad s_e^2 = \frac{12,96800826}{10-2} = 1,621001033.$$

(5). Стандартное отклонение ошибки регрессии – квадратный корень из дисперсии ошибки:

$$s_e = \sqrt{s_e^2} \qquad s_e = \sqrt{1,621001033} = 1,273185388.$$

(6). Показатель воздействия наблюдения или «разбалансировка» (Leverage) и (6-а) его центрированный вариант (Centered leverage)

(В литературе показатель влияния обозначатся обычно $\boldsymbol{h_{ii}}$ – этот символ происходит из матричной формы записи, где h_{ii} является диагональным элементом матрицы проекции на пространство регрессоров (hat matrix). Поскольку в данном тексте, по возможности, не употребляются матричные обозначения – далее он приводится как h_i.)

Наклон регрессионной зависимости, изображённой на рисунке, может быть представлен в виде суммы наклонов n частных регрессий, проходящих через каждое i-тое наблюдение и центр системы $O(\bar{x};\bar{y})$, взятых с определёнными весами h_i. Эти веса называются показателями воздействия, поскольку именно их величина определяет итоговый наклон регрессии. Чем дальше от центра системы по оси x находится наблюдение, тем сильнее его вклад в общее уравнение регрессии и тем, следовательно, больше h_i. Таким образом, h_i определяется только абсциссой точки и максимальные значения показателя влияния будут иметь крайние по оси x наблюдения. Величина показателя воздействия находится между 0 и 1, а сумма всех h_i равна числу параметров модели, включая свободный член (т.е. для линейной регрессии $\sum h_i = 2$). Большими считаются значения $h_i > \frac{2k}{n}$.

$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{(n-1)s_x^2}$, где s_x^2 - дисперсия предикторов x_i.

Например: $h_{10} = \frac{1}{10} + \frac{(12-5,1)^2}{(10-1)\cdot 10,76667} = 0,59133$. $\frac{2k}{n} = \frac{2\cdot 2}{10} = 0,4$, и таким образом, в нашем наборе имеется только одно значение с высоким показателем влияния – № 10. Как видно из рисунка, оно наиболее удалено от центра O.

В некоторых компьютерных программах приводится центрированный показатель воздействия. Его величина изменяется от 0 до 1-1/n.

$h_i^* = \frac{(x_i - \bar{x})^2}{(n-1)s_x^2} = h_i - \frac{1}{n}$ Например, $h_{10}^* = 0,59133 - \frac{1}{10} = 0,49133$.

Таблица 6. Показатели воздействия и центрированные показатели воздействия

i	1	2	3	4	5	6	7	8	9	10
h_i	0,27348	0,19917	0,14551	0,14551	0,11249	0,10010	0,10836	0,13725	0,18679	0,59133
h_i^*	0,17348	0,09917	0,04551	0,04551	0,01249	0,00010	0,00836	0,03725	0,08679	0,49133

(7). Стандартная ошибка остатка (Standard error of residual)

Учитывает как общий разброс наблюдений относительно регрессионной зависимости, так и удалённость конкретного наблюдения от центра системы. В отличие от стандартного отклонения остатков, которое одинаково для всех данных, стандартная ошибка остатка индивидуальна для каждого наблюдения и зависит от его показателя воздействия: чем он больше, тем меньше стандартная ошибка.

$m_i = s_e\sqrt{1-h_i}$ Например: $m_1 = 1,273185388\sqrt{1-0,27348} = 1,08522$.

Таблица 7. Стандартные ошибки остатков

i	1	2	3	4	5	6	7	8	9	10
m_i	1,08522	1,13936	1,17691	1,17691	1,19944	1,20778	1,20223	1,18259	1,14813	0,81391

III. Обнаружение выбросов (Outliers)

Выбросами считаются атипичные наблюдения, лежащее в стороне от регрессионной зависимости для большинства других наблюдений. Появление выбросов связано с влиянием на признак редких и/или обычно не учитываемых факторов, а также ошибками на стадии измерения признака. Говоря статистически, выброс не принадлежит данной генеральной совокупности, а потому должен быть исключён из анализа.

(8). Стандартизованный остаток (Standardized residual)

– отношение остатка к его стандартному отклонению. В некоторых литературных источниках и компьютерных программах стандартизованным остатком называют стьюдентизированный остаток (Studentized residual, см. ниже), в связи с чем, при первом использовании программы, необходимо с помощью ручного расчёта или с привлечением референтного набора данных выяснить что авторы понимают под этим термином.

Стандартизованные остатки распределены асимптотически нормально. Однако они имеют разную дисперсию и т.о. ещё не являются приведёнными к единому масштабу, поскольку не учитывают удалённость наблюдения от центра системы. Поэтому величины rs_i можно использовать лишь в качестве самых ориентировочных указателей на возможные выбросы.

$rs_i = \frac{r_i}{s_e}$ Например, $rs_4 = \frac{3,09907}{1,273185388} = 2,43411$

Обычно большими остатками считают rs_i >2. Shiffer (1988) (цит. по Gray, William, 1994) показал, что верхняя граница стандартизованного остатка не превышает $\sqrt{\frac{(n-k)(n-1)}{n}}$. Тем не менее для обнаружения выбросов не рекомендуется использовать и это выражение и следует рассчитывать внешние стьюдентизированные остатки (см. ниже).

Таблица 8. Стандартизованные остатки

i	1	2	3	4	5	6	7	8	9	10
rs_i	0,03972	-0,33395	-0,70762	2,43411	-0,29585	-0,66952	-0,25776	-0,63143	-0,21966	0,64196

(9). Внутренний стьюдентизированный остаток (Internally studentized residual)

– отношение остатка к его стандартной ошибке. Называется внутренним т.к. его расчёт проводится по полным данным внутри всей имеющейся системы наблюдений. В различных литературных источниках и компьютерных программах может называться просто «стьюдентизированным остатком (Studentized residual)» или «стандартизованным остатком (Standardized residual)», в связи с чем при первом использовании программы необходимо с помощью ручного расчёта или с привлечением референтного набора данных выяснить что авторы понимают под этим термином.

Внутренний стьюдентизированный остаток изменяется от 0 до $\sqrt{n-k-1}$ и имеет неизвестное распределение близкое к *t*-распределению Стьюдента с числом степеней свободы $v=n-k$. Отличие от *t*-распределения связано с тем, что при внутренней стьюдентизации числитель и знаменатель в формуле расчёта rt_i не являются независимыми.

$$rt_i = \frac{r_i}{m_i}$$

Например, $rt_4 = \frac{3,09907}{1,17691} = 2,63322$.

Для не слишком малых выборок и небольшого числа параметров регрессии *k* критические значения *t*-распределения для 5%-ного уровня значимости близки к 2. На основании этого наблюдения со стьюдентизированными остатками 2 и более являются подозрительными на выброс. Однако для обнаружения выбросов не рекомендуется использовать и этот показатель и следует предпочесть внешний стьюдентизированный остаток (см. ниже).

Таблица 9. Внутренние стьюдентизированные остатки

i	1	2	3	4	5	6	7	8	9	10
rt_i	0,04660	-0,37318	-0,76550	**2,63322**	-0,31404	-0,70578	-0,27297	-0,67980	-0,24359	1,00421

(10). Внешний стьюдентизированный остаток (Externally studentized residual) или стьюдентизированный удалённый остаток (Studentized deleted residual)

– отношение остатка к его стандартной ошибке, вычисленной по регрессионной зависимости с удалённым *i*-тым наблюдением. Удаление *i*-того наблюдения позволяет исключить его влияние как потенциального выброса на форму регрессионной зависимости. Внешним стьюдентизированным остатком он называется потому, что при его расчёте привлекается внешняя оценка дисперсии ошибки регрессии, свободная от влияния потенциального выброса. Вообще говоря, именно такая внешняя стьюдентизация и является собственно стьюдентизацией. Поэтому в некоторых литературных источниках и компьютерных программах его могут называть просто стьюдентизированным остатком (Studentized residual, см. вше), в связи с чем, при первом использовании программы, необходимо с помощью ручного расчёта или с привлечением референтного набора данных выяснить что авторы понимают под этим термином.

Внешний или удалённый стьюдентизированный остаток имеет *t*-распределение Стьюдента с числом степеней свободы $v=n-k$. Для не слишком малых выборок и небольшого числа параметров регрессии *k* критические значения *t*-распределения для 5%-ного уровня значимости близки к 2. На основании этого наблюдения с удалёнными стьюдентизированными остатками 2 и более обычно относятся к выбросам, хотя для более определённого вывода желательно рассчитать непосредственно достигнутый уровень значимости.

$$rt_{(i)} = \frac{r_i}{m_{(i)}}$$

Например, как было показано в (3), если удалить из набора данных наблюдение № 4, то уравнение регрессии изменится на: $\hat{y}_{i(4)} = 2,71014 + 0,55435x_i$. Дисперсия ошибки также изменится и составит 0,2468944, а стандартное отклонение ошибки – 0,496884694. Тогда, согласно (7), величина стандартной ошибки остатка с удалённым наблюдением № 4 составит:

$m_{(4)} = 0,496884694\sqrt{1-0,14551} = 0,459313395$, а $rt_{(4)} = \frac{3,09907}{0,459313395} = 6,74719$.

Таким образом, использование удалённых стьюдентизированных остатков позволило сделать вывод о наблюдении № 4 как выбросе: $t_{[8]}$= 6,74719; *P*=0,000145. Его следует исключить из анализа.

Таблица 10. Стьюдентизированные удалённые остатки

i	1	2	3	4	5	6	7	8	9	10
$rt_{(i)}$	0,04359	-0,35215	-0,74382	**6,74719**	-0,29559	-0,68176	-0,25654	-0,65510	-0,22870	1,00481
P	ns	ns	ns	**0,000145**	ns	ns	ns	ns	ns	ns

IV. Расчёт статистик влияния (Influence Statistics)

Статистики влияния позволяют выразить степень влияния наблюдения на форму и/или положение регрессионной зависимости. Поскольку влиятельные наблюдения в большей мере определяют регрессионную зависимость, даже небольшие ошибки в их оценках приводят к существенному искажению формы и/или положения регрессии, что отрицательно сказывается на её прогностических свойствах. Все меры влияния можно подразделить на общие и специфические. Общие меры показывают как *i*-тое наблюдение влияет на положение всей регрессионной зависимости. К ним относят: расстояние Кука, ковариационное отношение и *DFFITS*. Также к общим мерам влияния относят обычно и расстояние Махаланобиса. Специфические меры влияния, такие как *DFBETAS*, показвают влияние *i*-того наблюдения на отдельные параметры регрессионной модели. Расчёт всех мер влияния, за исключением расстояния Махаланобиса, использует технику исключения наблюдения из анализа.

(11). Расстояние Махаланобиса (Mahalanobis distance)

– наиболее распространённая мера удаления наблюдения от центра системы в многопеременном анализе. Поскольку она не учитывает зависимый или независимый характер переменной и рассматривает их в облаке рассеяния равнозначно, данная мера не является «заточенной» на решение задач регрессионного анализа. Тем не менее, она удобна тем, что (1) очень близка к показателю воздействия h_i и легко выводится из него и n и (2) может рассматриваться как ещё один тест на выбросы.

Расстояние Махаланобиса представляет собой расстояние от точки до центра системы, отнесенное к диаметру облака в этом направлении. Единичное расстояние MD_i соответствует 1 стандартному отклонению от центра системы в направлении рассматриваемой точки.

$$MD_i = (n-1)\left(h_i - \frac{1}{n}\right)$$

Например, $MD_{10} = (10-1)\left(0,59133 - \frac{1}{10}\right) = 4,42197$.

Критическое значение *MD* для для нужного уровня значимости α рассчитывается согласно Penny (1996) как: $MD_{[\alpha]} = \frac{k\left(\mathrm{n}-1\right)^2 F_{[\alpha;\nu_1=k;\nu_2=n-k-1]}}{n\left(n-k-1+k\cdot F_{[\alpha;\nu_1=k;\nu_2=n-k-1]}\right)}$.

Для нашего примера: k=2, n=10, $F_{[\alpha=0,05;\ 2;\ 7]}$= 4,737414, $F_{[\alpha=0,10;\ 2;\ 7]}$= 3,257442

$$MD_{[\alpha=0,05]} = \frac{2(10-1)^2 \cdot 4,737414}{10(10-2-1+2\cdot 4,737414)} = 4,65839,$$

$$MD_{[\alpha=0,10]} = \frac{2(10-1)^2 \cdot 3,257442}{10(10-2-1+2\cdot 3,257442)} = 3,90463.$$

Таким образом, по величине расстояния Махаланобиса ни одно значение на 5%-ном уровне значимости не является аномально удалённым от центра системы, т.к. любое $MD_i < MD_{[\alpha=0,05]}$. Тем не менее наблюдение № 10 может рассматриваться как подозрительное: 0,05<*P*<0,10.

Таблица 11. Расстояния Махаланобиса

i	1	2	3	4	5	6	7	8	9	10
MD_i	1,56130	0,89257	0,40960	0,40960	0,11234	0,00093	0,07523	0,33529	0,78111	4,42197

(12). Расстояние Кука (Cook's distance)

– общая мера влияния наблюдения. Как видно из представленных ниже формул, расстояние Кука может быть получено разными способами. Во-первых, если рассчитать разность между ожидаемыми значениями исходной регрессии и ожидаемыми значениями регрессии, построенной

с удалённым *i*-тым наблюдением, возвести её в квадрат и суммировать по всем *n* значениям – получим числитель формулы 1, а знаменателями будут дисперсия ошибки регрессии и число параметров *k*. Во-вторых, можно использовать только квадрат разности между ожидаемым для *i*-того наблюдения и ожидаемым для него согласно построенной без него регрессии, и полученную величину разделить на дисперсию ошибки регрессии, число параметров *k* и показатель влияния данного наблюдения. Третья формула показывает, что расстояние Кука является обобщённой мерой, учитывающей как внутренний стьюдентизированный остаток (9), так и показатель влияния наблюдения. Поскольку все необходимые для расчётов по ней значения уже были получены ранее, данная формула представляется наиболее удобной.

$$CD_i = \frac{\sum\left(\hat{y} - \hat{y}_{(i)}\right)^2}{ks_e^2} = \frac{(\hat{y}_i - \hat{y}_{(i)})^2}{ks_e^2 h_i} = \frac{rt_i^2}{k} \cdot \left(\frac{h_i}{1-h_i}\right).$$

Например, $CD_{10} = \frac{1{,}00421^2}{2} \cdot \left(\frac{0{,}59133}{1-0{,}59133}\right) = 0{,}72959$.

Критическое значение для расстояния Кука находится как медиана *F*-распределения с числом степеней свободы $\nu_1 = k$; $\nu_2 = n - k$, где *k* – количество параметров в модели.

Для наших данных, критическое значение статистики Кука составляет $F_{[\alpha=0{,}5;\nu_1=2;\nu_2=8]} = 0{,}75683$.

Таким образом, ни одно значение не может считаться значимо выделяющимся свои влиянием на регрессионную зависимость, хотя и для этой меры наиболее близким к критическому значению является наблюдение № 10.

Таблица 12. Расстояния Кука

i	1	2	3	4	5	6	7	8	9	10
CD_i	0,00041	0,01732	0,04989	0,59038	0,00625	0,02771	0,00453	0,03676	0,00681	0,72959

(13). Ковариационное отношение (CovRatio)

– общая мера влияния наблюдения. Представляет собой отношение детерминанта ковариационной матрицы с удалённым *i*-тым наблюдением к детерминанту ковариационной матрицы для всего набора данных. В двухпеременном случае расчёт удобно проводить по формуле:

$$CR_i = \left(\frac{s_{e(i)}^2}{s_e^2}\right)^2 \cdot \frac{1}{1-h_i}$$

Например, при удалении наблюдения № 10 дисперсия ошибки регрессии $s_{e(10)}^2 = 1{,}6190476$. Тогда $CR_{10} = \left(\frac{1{,}6190476}{1{,}621001032}\right)^2 \cdot \frac{1}{1-0{,}59133} = 2{,}44108$.

Влиятельными считаются наблюдения, значимо отличающиеся от 1. Для такой проверки из CR_i вычитают единицу. Если $|CR_i - 1| > \frac{3k}{n}$, то наблюдение считается существенно влияющим на регрессионную зависимость. В нашем случае критическим значением для $|CR_i - 1|$ является 6/10=0,6 и, как видно из таблицы 13, все наблюдения кроме № 4 следует рассматривать как влиятельные: удаление любого из них исказит регрессионную зависимость и расширит доверительный интервал для предсказания.

Таблица 13. Ковариационные отношения

i	1	2	3	4	5	6	7	8	9	10
CR_i	**1,79680**	**1,57468**	**1,31282**	0,02715	**1,43560**	**1,27630**	**1,43769**	**1,34406**	**1,58240**	**2,44108**

(14). Меры *DFFIT* и (14-а) стандартизованное *DFFIT* (*DFFITS*)

DFFIT – разность между ожидаемым значением для данного наблюдения и удалённым ожидаемым значением для него.

$DFFIT_i = \hat{y}_i - \hat{y}_{(i)}$ Например, при исключении наблюдения № 10 уравнение регрессии изменится на $\hat{y}_{i(10)} = 4 + \frac{1}{3} x_i$. Ожидаемое значение для наблюдения № 10 согласно ему будет $\hat{y}_{10(10)} = 4 + \frac{1}{3} 12 = 8$, тогда как для полной регрессии в пункте (1) оно составляло 9,18266. Тогда $DFFIT_{10} = 9{,}18266 - 8 = 1{,}18266$.

Расчёт стандартизованного значения *DFFIT* возможно по двум формулам. Из первой становится понятным, почему оно называется стандартизованным, а из второй видно, что как и дистанция Кука *DFFITS* объединяет стьюдентизированный остаток и показатель влияния наблюдения. Однако если в дистанции Кука используется остаток с внутренней стьюдендизаций, то в остатке для *DFFITS* она внешняя. По этой причине *DFFITS* имеет преимущество перед расстоянием Кука.

$DFFITS_i = \frac{DFFIT_i}{s_{e(i)}\sqrt{h_i}} = rt_{(i)}\sqrt{\frac{h_i}{1-h_i}}$ Например, как уже было рассчитано в (13), при удалении наблюдения №10, $s^2_{e(10)} = 1{,}6190476$, тогда $s_{e(10)} = \sqrt{1{,}6190476} = 1{,}272418013$, а $DFFITS_{10} = \frac{1{,}18266}{1{,}272418013\sqrt{0{,}59133}} = 1{,}20869$. Или, используя формулу 2 и опираясь на расчёты пункта (10), имеем:

$$DFFITS_{10} = 1{,}00481\sqrt{\frac{0{,}59133}{1-0{,}59133}} = 1{,}20868.$$

Для малых и средних объёмов выборок влиятельными наблюдениями признаются значения более 1, а для больших выборок – более $2\cdot\sqrt{\frac{k}{n}}$. Т.о. в нашей малой выборке влиятельными наблюдениями являются №№ 4 и 10.

Таблица 13. Меры *DFFIT* и *DFFITS*

i	1	2	3	4	5	6	7	8	9	10
$DFFIT_i$	0,01903	-0,10575	-0,15342	0,52774	-0,04774	-0,09482	-0,03988	-0,12790	-0,06424	1,18266
$DFFITS_i$	0,02675	-0,17562	-0,30694	**2,78431**	-0,10523	-0,22738	-0,08943	-0,26129	-0,10961	**1,20869**

(14). Меры *DFBETA* и (14-а) стандартизованное *DFBETA* (*DFBETAS*)

- специфические меры влияния, оценивающие степень изменения отдельных параметров регрессионной модели при исключении из i-того наблюдения.

$DFBETA_{ij} = \hat{B}_j - \hat{B}_{j(i)}$, где $\hat{B}$ - ожидаемое значение для j-того параметра регрессии.

В случае линейной регрессии $\hat{y}_i = a + bx_i$, $\hat{B}_0 = a$ (свободный член), $\hat{B}_1 = b$ (коэффициент регрессии).

Для нашего примера имеем: $\hat{y}_i = 3,47368 + 0,47575x_i$. Если теперь удалить из набора данных *i*-тое наблюдение, уравнение регрессии изменится. Так, при удалении наблюдения № 4 оно изменится на $\hat{y}_{i(4)} = 2,71014 + 0,55435x_i$. Тогда: $DFBETA_{4a}$ = 3,47368 - 2,71014 = 0,76354, $DFBETA_{4b}$ = 0,47575 - 0,55435 = -0,07860.

Стандартизация достигается делением $DFBETA_i$ на соответствующую стандартную ошибку:

$$DFBETAS_{ij} = \frac{DFBETA_{ij}}{s_{e(i)}\sqrt{C_{jj}}}$$, где $s_{e(i)}$ – стандартное отклонение ошибки регрессии при удалении *i*-того наблюдения, а C_{jj} – диагональнй элемент матрицы $(X'X)^{-1}$.

В случае линейной регресии:

$$DFBETAS_{ia} = \frac{DFBETA_{ia}}{s_{e(i)}\sqrt{\frac{\sum x_i^2}{n\left(\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}\right)}}};$$

$$DFBETAS_{ib} = \frac{DFBETA_{ib}}{s_{e(i)}\sqrt{\frac{1}{\sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}}}}$$

Например, для наблюдения № 4 имеем:

$$DFBETAS_{4a} = \frac{0,76354}{0,496884694\sqrt{\frac{357}{10\left(357 - \frac{(51)^2}{10}\right)}}} = 2,53165;$$

$$DFBETAS_{4b} = \frac{-0,07860}{0,496884694\sqrt{\frac{1}{357 - \frac{(51)^2}{10}}}} = -1,55714.$$

Как и в случае *DFFITS* для малых и средних объёмов выборок влиятельными наблюдениями признаются *DFBETAS* со значениями более 1, а для больших выборок – более $2\cdot\sqrt{\frac{k}{n}}$. Т.о. в нашей малой выборке наибольшее влияние на вертикальный сдвиг линии регрессии оказывает наблюдение № 4, а на наклон линии регрессии – наблюдения №№ 4 и 10.

Таблица 14. Меры *DFFIT* и *DFFITS*

i	1	2	3	4	5	6	7	8	9	10
Для свободного члена регрессии (Intercept)										
$DFBETA_i$	0,02197	-0,13972	-0,22197	0,76354	-0,06701	-0,09971	-0,01937	0	0,01810	-0,52632
$DFBETAS_i$	0,02661	-0,17061	-0,27909	**2,53165**	-0,08162	-0,12464	-0,02356	0	0,02199	-0,68147
Для коэффициента регрессии (X)										
$DFBETA_i$	-0,00295	0,01685	0,22850	-0,07860	0,00482	0,00098	-0,00342	-0,01827	-0,01029	0,14241
$DFBETAS_i$	-0,02130	0,12393	0,17166	**-1,55714**	0,03506	0,00730	-0,02484	-0,13613	-0,07472	**1,10176**

Использованные источники

McDonald B. A Teaching Note on Cook's Distance – A Guideline // Res. Lett. Inf. Math. Sci. 2002. № 3. P. 127-128. (URL: http://www.massey.ac.nz/~wwiims/research/letters/)

Muller K.E., Mok M.C. The ditribution of Cook' D statitics // Commun. Statit. – Theory Meth. 1977. V. 26, № 3. P. 525-546.

Penny K.I. Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance // Applied statistics. 1996. V. 45, № 1. P. 73-81.

Rawlings J.O., Pentula S.G., Dickey D.A.Applied regression analysis: A research tool.- 2nd ed. Springer, 659 p.

Cohen J., Cohen P., West S.G., Aiken L.S. Applied multiple regression/correlation analysis for the behavioral sciences. Mahwah, New Jersey, London: Lawrence Erlbaum Associates, 2003. 703 p.

А также:

http://www.stat.rutgers.edu/~saral/pdf/563/Leverages-influential.pdf

http://www.weibull.com/DOEWeb/experiment_design_and_analysis_reference.htm

http://www.aiaccess.net/English/Glossaries/Shop/bookstore.htm#GlosTut

www.davidson.edu/academic/economics/martin/jse.pdf

http://www.emis.de/journals/HOA/IJMMS/Volume13_4/806.pdf

http://www.jstor.org/pss/2684258

Wikipedia

Приложение.
Референтный набор данных для обнаружения выбросов и влиятельных наблюдений в регрессионном анализе.

i	x_i	y_i	Ожида-емое $\hat{y}_i$	Показатель влияния		Остаток r_i	Станд. ошибка остатка m_i	Стандарти-зованный остаток rs_i	Удалённый остаток $r_{(i)}$	Стьюденизированнй остаток	
				h_i	Центри-рованный h_i^*					Внутренний rt_i	Внешний (удалённый) $rt_{(i)}$
1	1	4	3,94943	0,27348	0,17348	0,05057	1,08522	0,03972	0,0696	0,0466	0,04359
2	2	4	4,42518	0,19917	0,09917	-0,42518	1,13936	-0,33395	-0,53093	-0,37318	-0,35215
3	3	4	4,90093	0,14551	0,04551	-0,90093	1,17691	-0,70762	-1,05435	-0,7655	-0,74382
4	3	8	4,90093	0,14551	0,04551	3,09907	1,17691	2,43411	3,62681	2,63322	6,74719
5	4	5	5,37668	0,11249	0,01249	-0,37668	1,19944	-0,29585	-0,42442	-0,31404	-0,29559
6	5	5	5,85243	0,10010	0,00010	-0,85243	1,20778	-0,66952	-0,94725	-0,70578	-0,68176
7	6	6	6,32817	0,10836	0,00836	-0,32817	1,20223	-0,25776	-0,36806	-0,27297	-0,25654
8	7	6	6,80392	0,13725	0,03725	-0,80392	1,18259	-0,63143	-0,93182	-0,6798	-0,6551
9	8	7	7,27967	0,18679	0,08679	-0,27967	1,14813	-0,21966	-0,34391	-0,24359	-0,2287
10	12	10	9,18266	0,59133	0,49133	0,81734	0,81391	0,64196	1,99999	1,00421	1,00481

i	Расстояние		Ковариаци-онное отношение CR_i	$DFFIT_i$	$DFFITS_i$	$DFBETA_i$		$DFBETAS_i$	
	Махала-нобиса MD_i	Кука CD_i				Свободный член a	Коэффи-циент регрессии b	Свободный член a	Коэффи-циент регресии b
1	1,5613	0,00041	1,7968	0,01903	0,02675	0,02197	0,02661	-0,00295	-0,0213
2	0,89257	0,01732	1,57468	-0,10575	-0,17562	-0,13972	-0,17061	0,01685	0,12393
3	0,4096	0,04989	1,31282	-0,15342	-0,30694	-0,22197	-0,27909	0,2285	0,17166
4	0,4096	0,59038	0,02715	0,52774	2,78431	0,76354	2,53165	-0,0786	-1,55714
5	0,11234	0,00625	1,4356	-0,04774	-0,10523	-0,06701	-0,08162	0,00482	0,03506
6	0,00093	0,02771	1,2763	-0,09482	-0,22738	-0,09971	-0,12464	0,00098	0,0073
7	0,07523	0,00453	1,43769	-0,03988	-0,08943	-0,01937	-0,02356	-0,00342	-0,02484
8	0,33529	0,03676	1,34406	-0,1279	-0,26129	0	0	-0,01827	-0,13613
9	0,78111	0,00681	1,5824	-0,06424	-0,10961	0,0181	0,02199	-0,01029	-0,07472
10	4,42197	0,72959	2,44108	1,18266	1,20869	-0,52632	-0,68147	0,14241	1,10176