Лекция 20

Проверка гипотезы о значимости модели с помощью F-критерия.

Даже если между y и x отсутствует связь, по данным выборочного наблюдения может оказаться, что такая зависимость существует. Только по случайному стечению обстоятельств характеризующая «объясняющую способность» модели величина r^2 будет в точности равна 0. Это представляет определенную проблему. Как узнать, действительно ли полученное при оценке регрессии значение r^2 отражает истинную зависимость или оно получено случайно? Для решения вопроса применяется процедура проверки гипотез, основанная на F-критерии, имеющем распределение Фишера:

$$F = \frac{\frac{1}{k}\sum_{i=1}^{n}\left(y_i - y_i'\right)^2}{\frac{1}{n-k-1}\sum_{i=1}^{n}\left(y_i' - \bar{y}\right)^2} = \frac{s_e^2 \big/ k}{s_{y'}^2 \big/ \left(n-k-1\right)}$$

Суммы в числителе и знаменателе соответствуют объясненной и необъясненной дисперсиям, k - число независимых переменных, которое для парной регрессии равно 1. Определенная таким образом величина подчиняется распределению Фишера с числом степеней свободы числителя k и знаменателя $n-k-1$. Значения критерия можно выразить и через коэффициент детерминации r^2:

$$r^2 = \frac{s_e^2}{s_y^2},\ s_y^2 = s_{y'}^2 + s_e^2,\ \frac{s_{y'}^2}{s_y^2} = 1 - r^2.$$

Разделив числитель и знаменатель в выражении для *F,* получим

$$F = \frac{s_e^2 \big/ 1}{s_{y'}^2 \big/ \left(n-2\right)} = \frac{s_e^2 \big/ s_y^2}{\frac{1}{n-2}\cdot\frac{s_{y'}^2}{s_y^2}} = \frac{r^2}{r^2 \big/ n-2}.$$

Свойства распределения Фишера

1) Является семейством кривых, зависящим от числа степеней свободы как числителя, так и знаменателя.

2) Принимает только положительные значения.

3) Является положительно - асимметричным.

4) Имеет математическое ожидание, примерно равное 1.

Несколько характерных кривых плотности распределения Фишера представлены на рис. 1.

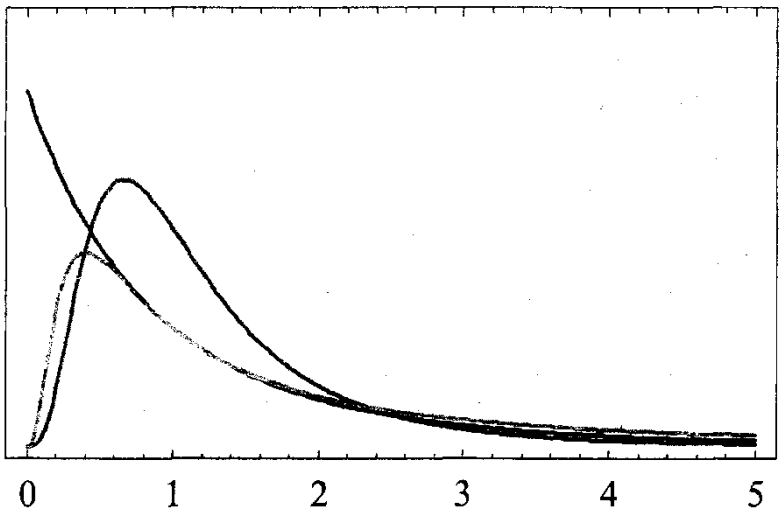

Рис. 1 кривые плотности распределение Фишера

Пример. Выполнить исследование значимости регрессионной модели для зависимости расходов на питание от дохода. Коэффициент детерминации $r^2 = 0.9775$, уровень значимости $\alpha = 5\%$.

1) Сформулировать гипотезы:

H_0: В генеральной совокупности нет зависимости расходов на питание от дохода;

H_1: В генеральной совокупности расходы на питание зависят от дохода.

2) Найти критическое значение для F - распределения со степенями свободы 1 для числителя и 23 для знаменателя для 5% значимости, $F_\alpha = 4.28$ [Гмурман В.Е. Теория вероятности и математическая статистика. – М:. Высшая школа,

2000. 470 с.], критическая область $F > F_{\alpha} = 7.88$.

3) Вычислить значение критерия:

$$F = \frac{r^2}{r^2/n-2} = \frac{0.9775}{0.0025/23} = 999.2$$

4) Сделать вывод. Расчетное значение превосходит критическое 999.2 > 4.28, принимается альтернативная гипотеза.

5) Вывод. В генеральной совокупности расходы на питание зависят от доходов.

Таблица значений F-критерия Фишера при уровне значимости α =0,05

k1 / k2	1	2	3	4	5	6	8	12	24	∞
1	161,4	199,5	215,7	224,5	230,1	233,9	238,8	243,9	249,0	254,32
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
35	4,12	3,26	2,87	2,64	2,48	2,37	2,22	2,04	1,83	1,57
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
45	4,06	3,21	2,81	2,58	2,42	2,31	2,15	1,97	1,76	1,48

Продолжение

	f	2	3	4	5	6	8	f2	24	∞
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	1,95	1,74	1,44
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
70	3,98	3,13	2,74	2,50	2,35	2,23	2,07	1,89	1,67	1,35
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,88	1,65	1,31
90	3,95	3,10	2,71	2,47	2,32	2,20	2,04	1,86	1,64	1,28
100	3,94	3,09	2,70	2,46	2,30	2,19	2,03	1,85	1,63	1,26
125	3,92	3,07	2,68	2,44	2,29	2,17	2,01	1,83	1,60	1,21
150	3,90	3,06	2,66	2,43	2,27	2,16	2,00	1,82	1,59	1,18
200	3,89	3,04	2,65	2,42	2,26	2,14	1,98	1,80	1,57	1,14
300	3,87	3,03	2,64	2,41	2,25	2,13	1,97	1,79	1,55	1,10
400	3,86	3,02	2,63	2,40	2,24	2,12	1,96	1,78	1,54	1,07
500	3,86	3,01	2,62	2,39	2,23	2,11	1,96	1,77	1,54	1,06
1000	3,85	3,00	2,61	2,38	2,22	2,10	1,95	1,76	1,53	1,03
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1,00

Прогнозирование с помощью уравнения регрессии

Построенное регрессионное уравнение позволяет для произвольного x рассчитать значение зависимой переменной $y' = a + bx$ и получить точечный прогноз. Однако очевидно, что вероятность в точности получить предсказанное значение при последующем случайном наблюдении чрезвычайно мала и оно может служить только в качестве общего ориентира. По этой причине чаще используется интервальный прогноз, выполнение которого гарантируется с заранее заданной (доверительной) вероятностью.

Построение интервального прогноза основывается на величине стандартной ошибки модели $S_{er}(y')$. При выполнении условия нормальной распределенности остатков и достаточно большого размера выборки величина $z = \dfrac{y - y'}{S_{er}(y')}$ подчиняется стандартному нормальному распределению. Соответственно, при заданном уровне значимости α (доверительной вероятности $1 - \alpha$) получаем:

$$y' - z_{\frac{\alpha}{2}} \cdot S_{er}(y') < y' < y' + z_{\frac{\alpha}{2}} \cdot S_{er}(y')$$

где $z_{\frac{\alpha}{2}}$ - критическое значение, найденное по таблице стандартного нормального распределения. В частности, для 95% и 99% доверительных вероятностей:

$$y' - 1.96 \cdot S_{er}(y') < y' < y' + 1.96 \cdot S_{er}(y')$$

$$y' - 2.58 \cdot S_{er}(y') < y' < y' + 2.58 \cdot S_{er}(y')$$

Область применимости интервального прогноза ограничена случаем достаточно больших выборок $n > 100$. Помимо ошибки, связанной с наличием

остатков (необъясненных отклонений), неточность прогнозов обусловлена также погрешностью выборочных коэффициентов уравнения a и b рассмотренной ранее. Однако в силу того, что средняя ошибка коэффициентов с увеличением размера выборки убывает как $\approx\sqrt{\frac{1}{n}}$, при большом n вклад коэффициентов в суммарную погрешность становится пренебрежимо малым.

Пример. По выборке из ста единиц построена регрессионная модель зависимости давления крови от возраста $y'=100+0.96\cdot x$ при средней ошибке модели $S_{er}\left(y'\right)=5$. Построить с доверительной вероятностью 95% оценку для давления человека в возрасте 43 года.

Решение.

1) Находим величину точечного прогноза при $x=43$:

$$y'=100+0.96\cdot 43=141.28.$$

2) Для 95% доверительной вероятности интервальный прогноз имеет вид:

$$y'-1.96\cdot S_{er}\left(y'\right)<y'<y'+1.96\cdot S_{er}\left(y'\right)$$

Подставляя рассчитанное y' и $S_{er}\left(y'\right)$ из условий задачи, получаем:

$$141.28-1.96\cdot 5<y'<141.28+1.96\cdot 5$$

$$131.48<y'<151.08$$

С вероятностью 95% можно прогнозировать, что давление человека в возрасте 43 года окажется в интервале [131.48; 151.08].

Нелинейная и множественная регрессия

Парная нелинейная регрессия

Одним из недостатков линейного регрессионного анализа является то, что он применим только к описанию линейных зависимостей. В то же время для моделирования многих физических, химических, социально-экономических процессов необходимо использовать нелинейные соотношения. В общем случае изменение функциональной формы модели сопряжено с выводом новых

формул для коэффициентов, проведением исследований точности и значимости модели. Однако для целого ряда нелинейных моделей удается избежать этой трудоемкой процедуры и распространить на них результаты линейного случая, используя *преобразование переменных.* Рассмотрим его на конкретном примере.

В табл. 1, колонки 1 и 2, приведены данные гипотетического обследования, в котором для 10 семей зафиксированы годовой доход и количество купленных бананов. Диаграмма рассеяния зависимости представлена на рис. 2.

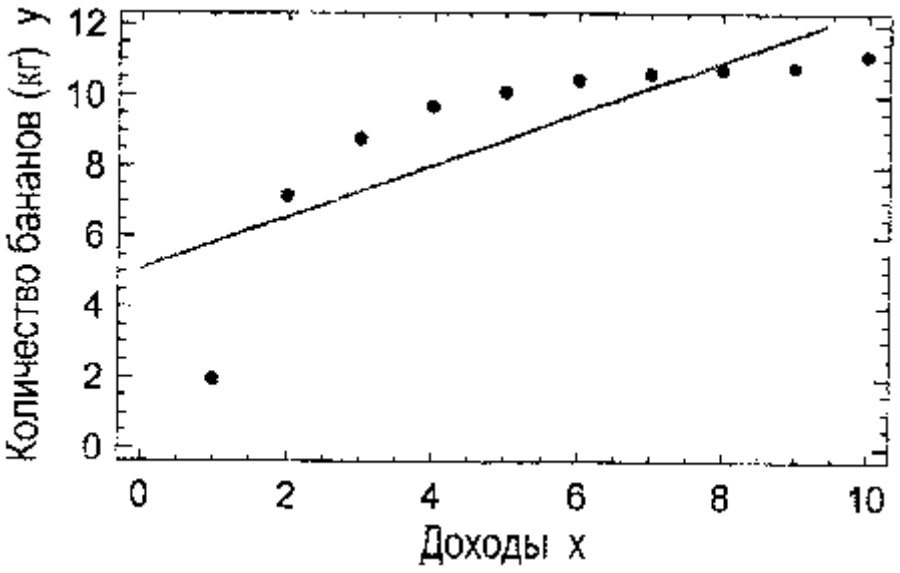

Рис. 2. Модель линейной регрессии зависимости количества бананов от доходов

Таблица 1. *Замена переменной при выводе уравнения нелинейной регрессии*

Годовой доход x_i *(в 1000 у.е.)*	Количество бананов y_i (в кг)	Линейная		Нелинейная регрессия		
		y'_i	e_i	$z_i = \frac{1}{x_i}$	y'_i	e_i
1	1.93	5.82	-3.90	1.000	2.00	-0.07
2	7.13	6.56	0.57	0.500	7.04	0.09

3	8.78	7.29	1.49	0.333	8.72	0.06
4	9.69	8.03	1.67	0.250	9.51	0.18
5	10.09	8.76	1.33	0.200	10.06	0.03
6	10.42	9.50	0.93	0.167	10.40	0,02
7	10.62	10.23	0.39	0.143	10.64	-0.02
8	10.71	10.97	-0.26	0.125	10.82	-0.11
9	10.79	11.70	-0.91	0.111	10.96	-0.17
	11.13	12.43	-1.31	0.100	11.07	0.06

По выборочным данным найдем уравнение линейной регрессии $y' = 5.09 + 0.73 \cdot x$, определим расчетное значение y'_i в каждой точке (колонка 3), величины остатков e_i (колонка 4), стандартные ошибки $S_{er}(a) = 1.23$, $S_{er}(b) = 0.2$ и вычислим коэффициент детерминации $r^2 = 0.64$. Построенная линия регрессии также изображена на рис. 2. Несмотря на то, что тесты на значимость коэффициентов и модели в целом дают положительный результат, соотношение построенной прямой и фактических значений на графике заставляет предположить, что функциональная зависимость определена неправильно. Понять это можно по поведению остатков. Положительные или отрицательны, большие или маленькие остатки должны чередоваться случайным образом. Здесь же, как видно из таблицы, сначала остатки отрицательны, затем они становятся положительными, достигают максимума, а потом снова уменьшаются и становятся отрицательными.

С учетом характерного расположения точек на диаграмме попробуем описать зависимость между величинами с помощью гиперболического закона $y' = a + \frac{b}{x}$. Для этого выполним преобразование $z = \frac{1}{x}$ и рассчитаем значения

$z_i = \frac{1}{x_i}$ (колонка 5). Полученная зависимость $y' = 12.08 - 10.08 \cdot z = 12.08 - \frac{10.08}{x}$ характеризуется коэффициентом детерминации $r^2 = 0.9989$, то есть является намного более точной. Зависимости $y(z)$ и $y(x)$ представлены на рис. 3 а, б.

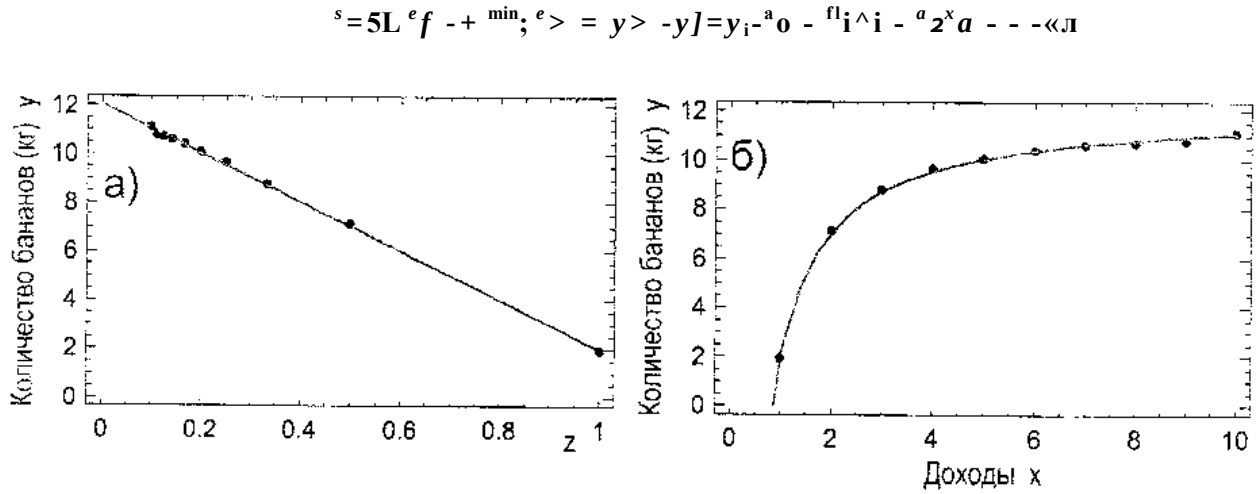

Рис. 3. Модель нелинейной регрессии зависимости количества бананов от доходов

О высоком качестве модели говорит и то, что точки фактических значений почти точно попадают на построенную кривую.

Таким образом, преобразование переменных позволило свести нелинейную регрессионную модель к линейной модели. В таблице 2 представлены наиболее часто используемые нелинейные модели, которые путем преобразования – взятием обратной величины, логарифмированием, возведением в степень и т.п. – могут быть трансформированы в линейную модель.

Таблица 2. *Нелинейные модели парной регрессии и преобразования переменных*

Тип модели	*Связь*	*Преобразования*	*Линейное уравнение*

Экспоненциальная	$y = \exp(a + bx)$	$\ln y = u$	$u = a + b \cdot x$
Обратная по y	$y = \dfrac{1}{a + bx}$	$\dfrac{1}{y} = u$	$u = a + b \cdot x$
Обратная по x	$y = a + \dfrac{b}{x}$	$\dfrac{1}{x} = z$	$y = a + b \cdot z$
Дважды обратная	$y = \dfrac{1}{a + \dfrac{b}{x}}$	$\dfrac{1}{x} = z,\ \dfrac{1}{y} = u$	$u = a + b \cdot z$
Логарифм по x	$y = a + b \cdot \ln x$	$\ln x = z$	$y = a + b \cdot z$
Мультипликативная	$y = a \cdot x^b$	$\ln x = z;$ $\ln y = u\ \ln a = a_1$	$u = a_1 + b \cdot z$
Квадратный корень по x	$y = a + b\sqrt{x}$	$z = \sqrt{x}$	$y = a + b \cdot z$
Квадратный корень по y	$y = \sqrt{a + b \cdot x}$	$u = y^2$	$u = a + b \cdot x$
S-кривая	$y = \exp\left(a + \dfrac{b}{x}\right)$	$\ln y = u;\ \dfrac{1}{x} = z$	$u = a + b \cdot z$

При использовании для регрессионного анализа программ статистической обработки нет необходимости пытаться заранее угадать тип функциональной зависимости между данными. Большинство программ по заданной выборке автоматически рассчитывают коэффициенты представленных зависимостей и их коэффициенты детерминации. После этого в качестве наиболее удачной можно взять модель с наибольшим r^2, при условии, что тесты на значимость коэффициентов и модели в целом дают положительный результат.

В последние годы еще одним стандартным средством парного регрессионного анализа стало *преобразование Бокса-Кокса.* Это

преобразование относится к зависимой переменной y и имеет вид $z=\frac{1}{\lambda}\left(y^{\lambda}-1\right)$, где λ - свободный параметр. Осуществляя перебор значений λ в диапазоне [0;1], ищется то из них, при котором построенная методом наименьших квадратов регрессионная модель $z=a+b\cdot x$ будет обладать наименьшей суммой квадратов остатков $S=\sum e_i^2$.

В заключении краткого обзора парных нелинейных моделей следует упомянуть класс *полиномиальных моделей* $y=a_0+a_1\cdot x+a_2\cdot x^2+\ldots++a_k\cdot x^k$, где k - степень полинома. Определить коэффициенты $a_0,a_1,a_2,\ldots,a_k$ можно, по аналогии с линейной моделью, методом наименьших квадратов. Соответствующая система линейных уравнений задается условиями минимизации функционала $\frac{\partial S}{\partial a_j}=0$. В ряде случаев использование полиномиальной модели позволяет повысить коэффициент детерминации, но обеспечить значимость всех коэффициентов удается только при достаточно большой длине выборки.

Множественная линейная регрессия

Множественный регрессионный анализ является развитием парного применительно к случаям, когда зависимая переменная гипотетически связана с более чем одной независимой. Значительная часть вопросов построения и исследования регрессионной модели в этом случае решается аналогично случаю парной регрессии, однако возникнут и две новые проблемы:

1) Проблема оценки влияния каждой из независимых переменных на зависимую. Сложность задачи связана с явлением *мультиколлинеарности* - взаимовлияния независимых переменных, которая отрицательно сказывается на точности модели.

2) Проблема спецификации модели (отбора переменных). Часто

предполагается, что несколько переменных могут оказывать влияние на зависимую, с другой стороны, некоторые переменные могут не подходить для модели. Необходимо решить, какие из них следует включить в уравнение регрессии, а какие нет.

Строгое математическое рассмотрение множественной регрессии сопряжено со значительными трудностями и выходит за рамки курса. С другой стороны, даже небольшое формальное представление об этом виде исследований позволит проводить его с помощью программ статистической обработки.

Уравнение множественной регрессии. В общем виде модель *множественной линейной регрессии* (относящаяся к генеральной совокупности) имеет вид

$$y = A_0 + A_1 \cdot x_1 + A_2 \cdot x_2 + \ldots + A_k \cdot x_k + u$$

где y - зависимая переменная, $x_1, x_2, \ldots, x_k$ независимые переменные, $A_0, A_1, A_2, \ldots, A_k$ - коэффициенты регрессии, u - остаточный член.

Построение выборочной регрессионной модели $y' = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2 + \ldots + a_k \cdot x_k$ осуществляется по данным выборочного наблюдения y_i, $i = 1, \ldots, n$ и x_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, k$ где x_{ij} - значение j-й переменной в i-м наблюдении. Коэффициенты модели определяются методом наименьших квадратов, из условия минимизации суммы квадратов остатков:

$$S = \sum_{i=1}^{n} e_i^2 \to \min, \; e_i = y_i - y_i' = y_i - a_0 - a_1 \cdot x_{i1} - a_2 \cdot x_{i2} - \ldots - a_k \cdot x_{ik}$$

Коэффициент детерминаций. Объясняющая способность модели характеризуется коэффициентом детерминации r^2, определяемым как отношение объясненной дисперсии к полной $r^2 = \dfrac{S_{y'}^2}{S_y^2} = \dfrac{1 - S_e^2}{S_y^2}$ При добавлении к модели новой переменной этот коэффициент возрастает (не убывает). Однако определить с помощью коэффициента детерминации

объясняющую способность каждой переменной в отдельности не удается из-за взаимной коррелированности - добавление новой переменной изменяет вклад уже включенных. Наряду с "обычным" коэффициентом детерминации применяется также "скорректированный" (adjusted) коэффициент, учитывающий количество переменных в модели:

$$r_1^2 = r^2 - \frac{k}{n-k-1}\left(1-r^2\right)$$

Проверка значимости коэффициентов. Точность определения коэффициентов a_j регрессионной модели характеризуется величиной стандартной ошибки $S_{er}\left(a_j\right)$, пропорциональной "среднему" остатку S_e. Проверка гипотез о значимости коэффициентов проводится с помощью t - критерия $t = \frac{a_j}{S_{er}\left(a_j\right)}$ подчиняющегося распределению Стьюдента с числом степеней свободы $n-k-1$: число наблюдений минус число независимых переменных минус свободный член.

Проверка значимости модели в целом проводится с помощью F-критерия:

$$F = \frac{\frac{r^2}{k}}{\left(1-r^2\right)\Big/ n-k-1}$$

Если значение критерия превышает критическую для данного уровня значимости величину, то нулевая гипотеза $H_0 : A_0 = A_1 = A_2 = \ldots = A_k = 0$ отклоняется. Критерий F используется также для отбора переменных и проверки значимости вклада в модель каждой из переменных.

Новая k-я переменная добавляется в модель, если величина F:

$$F = \frac{\textit{Улучшение качества модели} \,/\, 1}{\textit{Необъясненная сумма квадратов отклонений} \,/\, n-k-1}$$

превышает критическое значение F_a, $F_a = 4.41$ для $\alpha = 5\%$ и $F_a = 8.29$ для $\alpha = 1\%$, где "Улучшение качества модели" равно разности суммы квадратов объясненных отклонений после и до добавления новой переменной.

s=5L ef -+ min; e> = y> -y]=yi-a0 - f1i^i - a2xa - - -«л