

Расчетный файл выложен на странице https://www.matburo.ru/ex_ec.php?p1=ecexcel

Эконометрика

Решение задачи на множественную регрессию в Excel

По заданным статистическим данным постройте линейную модель множественной регрессии и исследуйте её.

№ район а	Урожайность зерновых культур, ц/га у	Число тракторов приведённой мощности на 100 га x_1	Число зерноуборочных комбайнов на 100 га x_2	Количество удобрений, расходуемых на 1 га, т/га x_3	Количество химических средств защиты на 1га, ц/га x_4
1	9.70	1.59	0.25	0.32	0.14
2	9.95	0.46	0.26	0.77	0.66
3	8.40	0.28	0.29	0.59	0.46
4	8.78	1.13	0.27	0.55	0.64
5	10.5	0.64	0.24	0.76	0.82
6	11.2	0.59	0.31	0.99	0.89
7	12.0	0.73	0.27	0.64	0.20
8	10.1	1.34	0.27	0.38	0.35
9	9.67	1.06	0.28	0.48	0.42
10	9.55	1.33	0.26	0.62	0.51
11	10.3	0.95	0.23	0.81	0.44
12	11.6	0.86	0.28	0.75	0.37
13	12.4	0.97	0.27	0.46	0.28
14	11.3	1.05	0.26	0.42	0.19
15	12.1	1.38	0.26	0.57	0.30
16	11.7	1.25	0.25	0.76	0.38
17	10.4	0.77	0.28	0.79	0.47
18	9.93	0.94	0.26	0.77	0.40
19	9.48	0.82	0.25	0.86	0.63
20	8.85	0.78	0.26	0.88	0.52

1) Постройте линейную модель множественной регрессии.

2) Запишите стандартизованное уравнение множественной регрессии. На основе стандартизованных коэффициентов регрессии и средних коэффициентов эластичности ранжировать факторы по степени их влияния на результат.

3) Найдите коэффициенты парной, частной и множественной корреляции. Проанализируйте их.

4) Найдите скорректированный коэффициент множественной детерминации.

Сравните его с нескорректированным (общим) коэффициентом детерминации.

5) С помощью F -критерия Фишера оценить статистическую надежность уравнения регрессии и коэффициента детерминации R^2_{yx1x2} .

6) С помощью частных F -критериев Фишера оценить целесообразность включения в уравнение множественной регрессии фактора x_1 после x_2 и фактора x_2 после x_1 .

7) Составьте уравнение линейной парной регрессии, оставив лишь один значащий фактор.

Решение:

1. Для построения линейной модели множественной регрессии используем инструмент Регрессия пакета Анализ данных в MS Excel. Вводим исходные данные в ячейки A1-E21.

	A	B	C	D	E	F	G	H	I	J	K	L
1	Y	X1	X2	X3	X4							
2	9.7	1.59	0.25	0.32	0.14							
3	9.95	0.46	0.26	0.77	0.66							
4	8.4	0.28	0.29	0.59	0.46							
5	8.78	1.13	0.27	0.55	0.64							
6	10.5	0.64	0.24	0.76	0.82							
7	11.2	0.59	0.31	0.99	0.89							
8	12	0.73	0.27	0.64	0.2							
9	10.1	1.34	0.27	0.38	0.35							
10	9.67	1.06	0.28	0.48	0.42							
11	9.55	1.33	0.26	0.62	0.51							
12	10.3	0.95	0.23	0.81	0.44							
13	11.6	0.86	0.28	0.75	0.37							
14	12.4	0.97	0.27	0.46	0.28							
15	11.3	1.05	0.26	0.42	0.19							
16	12.1	1.38	0.26	0.57	0.3							
17	11.7	1.25	0.25	0.76	0.38							
18	10.4	0.77	0.28	0.79	0.47							
19	9.93	0.94	0.26	0.77	0.4							
20	9.48	0.82	0.25	0.86	0.63							
21	8.85	0.78	0.26	0.88	0.52							

Нажимаем ОК и получаем результат:

Регрессионная статистика									
Множественный	0.456134219								
R-квадрат	0.208058426								
Нормированный	-0.003125994								
Стандартная оши	1.175708942								
Наблюдения	20								
Дисперсионный анализ									
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Значимость F</i>				
Регрессия	4	5.447322256	1.361830564	0.985197802	0.445156649				
Остаток	15	20.73437274	1.382291516						
Итого	19	26.181695							
Коэффициенты									
	<i>Коэффициенты</i>	<i>Стандартная ошибка</i>	<i>t-статистика</i>	<i>P-значение</i>	<i>Нижние 95%</i>	<i>Верхние 95%</i>	<i>Нижние 95.0%</i>	<i>Верхние 95.0%</i>	
Y-пересечение	6.6496103	5.331794261	1.247161832	0.231460337	-4.714840102	18.0140607	-4.714840102	18.0140607	
X1	0.604774143	1.097906152	0.550843204	0.589851924	-1.735357415	2.9449057	-1.735357415	2.9449057	
X2	10.67132745	16.46969348	0.647937223	0.526819716	-24.43299308	45.77564798	-24.43299308	45.77564798	
X3	2.843684421	2.183162895	1.302552561	0.212372042	-1.809617118	7.496985959	-1.809617118	7.496985959	
X4	-3.366471014	1.891180849	-1.780089416	0.09532221	-7.397427556	0.664485529	-7.397427556	0.664485529	

Используя значения в столбце «Коэффициенты» получаем уравнение линейной множественной регрессии в естественной форме:

$$\tilde{y} = 6.6496 + 0.6048 \cdot x_1 + 10.6713 \cdot x_2 + 2.8437 \cdot x_3 - 3.3665 \cdot x_4.$$

Полученное уравнение регрессии показывает взаимосвязь между урожайностью зерновых культур, числом тракторов, числом комбайнов, количеством удобрений и количеством химических средств защиты. Из уравнения видно, что с ростом числа тракторов приведенной мощности на 100 га на 1 ед. урожайность зерновых вырастает на 0.6048 ц/га. С ростом числа комбайнов на 1 ед. урожайность зерновых вырастает на 10.6713 ц/га. Увеличение на 1 т/га количества удобрений влечет за собой рост урожайности на 2,8437 ц/га. А вот рост на 1 ц/га количества средств химической защиты снижает урожайность на 3,3665 ц/га.

2. На основе естественных коэффициентов регрессии рассчитаем стандартизированные коэффициенты по формулам:

$$\beta_i = b_i \cdot \frac{\sigma_{x_i}}{\sigma_y}.$$

Расчет будем проводить в Excel. Для определения σ используем функцию СТАНДОТКЛОН().

Получаем уравнение в стандартизированной форме:

$$t_y = 0.171 \cdot t_{x_1} + 0.163 \cdot t_{x_2} + 0.445 \cdot t_{x_3} - 0.571 \cdot t_{x_4}.$$

Так как стандартизированные коэффициенты регрессии можно сравнивать между собой, то можно сказать, что наибольшее влияние на урожайность оказывают количество удобрений и количество химических средств защиты, так как их коэффициенты по модулю примерно одинаковы и существенно превышают значения стандартизированных коэффициентов для факторов 1 и 2.

Вычисляем коэффициенты эластичности по формулам:

$$\bar{\varepsilon}_i = b_i \cdot \frac{\bar{x}_i}{\bar{y}_{xi}}$$

Расчет будем проводить в Excel. Для определения средних значений используем функцию СРЗНАЧ().

Вычисляем:

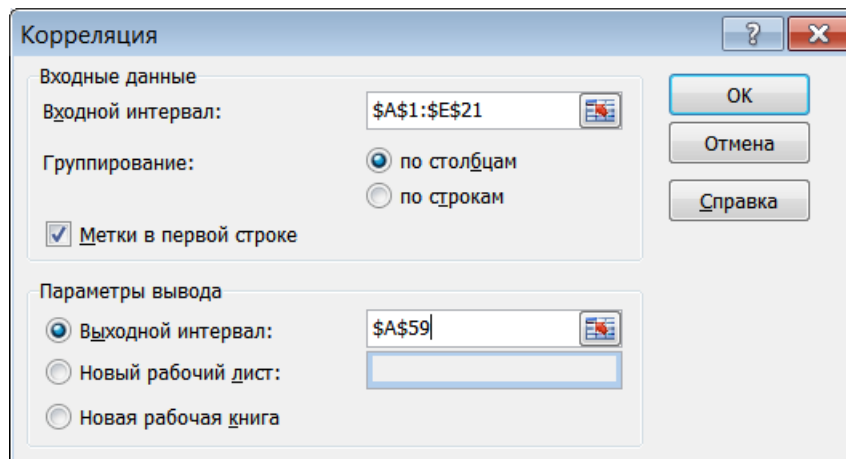
$$\bar{\varepsilon}_1 = 0.055, \bar{\varepsilon}_2 = 0.272, \bar{\varepsilon}_3 = 0.18, \bar{\varepsilon}_4 = -0.147.$$

Частный коэффициент эластичности показывает, насколько процентов в среднем изменяется признак-результат y с увеличением признака-фактора x_j на 1% от своего среднего уровня при фиксированном положении других факторов модели. Все коэффициенты эластичности существенно ниже 1, значит, влияние факторов на результативный признак Y незначительно. В то же время максимальное значение имеет фактор 2, то есть увеличение только числа комбайнов (от своего среднего значения) на 1% увеличивает в среднем урожайность зерновых культур на 0,272 %.

3. Найдем коэффициенты парной корреляции. Расчет можно вести по формуле

$$r_{yx_i} = \frac{\text{cov}(y, x_i)}{\sigma_y \cdot \sigma_{x_i}} = \frac{\overline{y \cdot x_i} - \bar{y} \cdot \bar{x}_i}{\sigma_y \cdot \sigma_{x_i}},$$

однако проще использовать инструмент Корреляция пакета Анализ данных, который сразу строит матрицу парных коэффициентов корреляции.



Результат:

	Y	X1	X2	X3	X4
Y	1.000				
X1	0.157	1.000			
X2	0.027	-0.347	1.000		
X3	-0.030	-0.577	0.050	1.000	
X4	-0.333	-0.524	0.173	0.673	1.000

Коэффициенты парной корреляции указывают между y и факторами имеет достаточно низкие значения (все коэффициенты по модулю ниже 0,4), то есть связь слабая. Заметная связь наблюдается между переменными x_1 и x_3 (коэффициент корреляции -0,577), x_1 и x_4 (коэффициент корреляции -0,524), x_3 и x_4 (коэффициент корреляции 0,673), то есть данные коэффициенты могут быть коллинеарными.

Определяем частные коэффициенты корреляции по формулам:

$$r_{yx_i/x_j} = \frac{r_{yx_i} - r_{yx_j} \cdot r_{xix_j}}{\sqrt{(1 - r_{yx_j}^2) \cdot (1 - r_{xix_j}^2)}}.$$

Получаем:

$$r_{yx_1/x_2} = 0.177, r_{yx_1/x_3} = 0.171, r_{yx_1/x_4} = -0.022.$$

Теснота связи везде низкая.

$$r_{yx_2/x_1} = 0.088, r_{yx_2/x_3} = 0.029, r_{yx_2/x_4} = 0.091.$$

Теснота связи везде низкая.

$$r_{yx_3/x_1} = 0.0745, r_{yx_3/x_2} = -0.032, r_{yx_3/x_4} = 0.278.$$

Теснота связи везде низкая.

$$r_{yx_4/x_1} = -0.298, r_{yx_4/x_2} = -0.343, r_{yx_4/x_3} = -0.423.$$

Теснота связи низкая для первых двух случаев, в последнем – умеренная.

Коэффициент множественной корреляции:

$$R_{yx_1x_2} = \sqrt{\sum \beta_i \cdot r_{yx_i}} = 0.456$$

(поле Множественный R в регрессионной статистике).

Коэффициент множественной корреляции показывает на умеренную связь всего набора факторов с результатом.

4. Нескорректированный коэффициент множественной детерминации $R^2_{yx_1x_2} = 0.208$ (поле R-квадрат), что говорит о том, что 20.8 % вариации результата объясняется вариацией представленных в уравнении факторов.

Скорректированный коэффициент множественной детерминации (поле Нормированный R-квадрат):

$$R_{корр}^2 = 1 - (1 - R^2) \cdot \frac{(n-1)}{(n-m-1)} = 1 - (1 - 0.208) \cdot \frac{(20-1)}{(20-4-1)} \approx -0.00313.$$

Значение скорректированного коэффициента множественной детерминации определяет тесноту связи с учетом степеней свободы общей и остаточной дисперсий. Скорректированный коэффициент детерминации практически равен 1, то есть уравнение регрессии не объясняет вариацию урожайности.

5. Оценим значимость уравнения:

$$F_{набл} = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m} = \frac{0.208}{1-0.208} \cdot \frac{20-4-1}{4} = 0.99.$$

Такое же значение получено в поле «F» дисперсионного анализа.

Табличное значение критерия при уровне значимости $\alpha = 0,05$ и $k_1 = m = 4$, $k_2 = n - m - 1 = 20 - 4 - 1 = 15$:

$$F_{табл} = F(0,05; 4; 15) = 3,06.$$

Так как $F_{табл} > F_{набл}$, то с вероятностью $1 - \alpha = 0,95$ делаем заключение о незначимости уравнения регрессии и коэффициента множественной детерминации (поле «Значимость F» выше 0,05).

6. С помощью частных F -критериев Фишера оценим целесообразность включения в уравнение множественной регрессии факторов x_i после остальных факторов.

$$F_{x1} = \frac{R^2 - R^2(x_2, x_n)}{1 - R^2} \cdot (n - m - 1) = \frac{0.208 - 0.181}{1 - 0.208} \cdot (20 - 4 - 1) = 0.508,$$

$$\text{где } R^2(x_2, x_n) = \beta_2 r_{yx2} + \beta_3 r_{yx3} + \beta_4 r_{yx4} = 0.181.$$

Критическое значение: $F_{кр}(k_1=3; k_2=15) = 3.29$.

$F_{x1} < 3.29$, следовательно, фактор x_1 не целесообразно включать в модель после введения факторов x_j .

$$F_{x2} = \frac{R^2 - R^2(x_3, x_n)}{1 - R^2} \cdot (n - m - 1) = \frac{0.208 - 0.204}{1 - 0.204} \cdot (20 - 4 - 1) = 0.084,$$

$$\text{где } R^2(x_3, x_n) = \beta_1 r_{yx1} + \beta_3 r_{yx3} + \beta_4 r_{yx4} = 0.204.$$

$F_{x_2} < 3.29$, следовательно, фактор x_2 не целесообразно включать в модель после введения факторов x_j .

$$F_{x_3} = \frac{R^2 - R^2(x_3, x_n)}{1 - R^2} \cdot (n - m - 1) = \frac{0.208 - 0.222}{1 - 0.222} \cdot (20 - 4 - 1) = -0.257,$$

где $R^2(x_3, x_n) = \beta_1 r_{yx1} + \beta_2 r_{yx2} + \beta_4 r_{yx4} = 0.222$.

$F_{x_3} < 3.29$, следовательно, фактор x_3 не целесообразно включать в модель после введения факторов x_j .

$$F_{x_4} = \frac{R^2 - R^2(x_4, x_n)}{1 - R^2} \cdot (n - m - 1) = \frac{0.208 - 0.0177}{1 - 0.222} \cdot (20 - 4 - 1) = 3.606,$$

где $R^2(x_4, x_n) = \beta_1 r_{yx1} + \beta_2 r_{yx2} + \beta_3 r_{yx3} = 0.0177$.

$F_{x_4} > 3.29$, следовательно, фактор x_4 целесообразно включать в модель после введения факторов x_j .

7. Найдем уравнение парной регрессии $y = a + bx_4$. Для этого снова используем инструмент Регрессия пакета Анализ данных. Фактор x_4 выбран как наиболее тесно связанный с результирующей переменной.

Регрессионная статистика									
70	Множественный	0.333199774							
71	R-квадрат	0.111022089							
72	Нормированный	0.061634428							
73	Стандартная оши	1.137124749							
74	Наблюдения	20							
75									
Дисперсионный анализ									
77		df	SS	MS	F	Значимость F			
78	Регрессия	1	2.906746484	2.906746484	2.247972178	0.151124302			
79	Остаток	18	23.27494852	1.293052695					
80	Итого	19	26.181695						
81									
	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%	Нижние 95.0%	Верхние 95.0%	
83	Y-пересечение	11.28577006	0.64593229	17.47206362	9.78974E-13	9.928716682	12.64282345	9.928716682	12.64282345
84	X4	-1.963109293	1.309329681	-1.499323907	0.151124302	-4.713908874	0.787690288	-4.713908874	0.787690288

Получаем уравнение линейной регрессии $y = 11.286 - 1.9631x_4$. Коэффициент $b = -1.9631$ показывает, что при увеличении количества химических средств защиты на 1 ц/га величина урожайности снижается в среднем на 1.9631 ц/га. Построенное уравнение регрессии на 11,1 % объясняет вариацию урожайности вариацией количества химических средств защиты (значение R-квадрат). Построенное

Задача по эконометрике с решением в Excel. Выполнена в <https://www.matburo.ru/>

©МатБюро – Решение задач по математике, экономике, программированию

уравнение линейной регрессии не является значимым (значимость $F = 0,1511 > 0,05$).

Список литературы:

1. Елисеева И. И. Эконометрика: учебник/ И.И. Елисеева – М.: Финансы и статистика, 2007.
2. Елисеева И. И. Практикум по эконометрике: учеб. пособие/ И. И. Елисеева – М.: Финансы и статистика, 2008.
3. Кремер Н. Ш. Математика для экономистов : от Арифметики до Эконометрики : учебно-справочное пособие / Н. Ш. Кремер, Б. А. Путко, И. М. Тришин ; ред. Н. Ш. Кремер. - 2-е изд., перераб. и доп. - М. : Юрайт, 2010.