3.2. Задачи

Задача 3.1

По различным районам (i — номер района) исследуется зависимость урожайности зерновых культур от ряда переменных, характеризующих различные факторы сельскохозяйственного производства (табл. 3.3).

Таблица 3.3

Зависимость урожайности зерновых культур y_i, ц/га от факторов сельскохозяйственного производства: x_{i1} — число тракторов (приведенной мощности) на 100 га; x_{i2} — число зерноуборочных комбайнов на 100 га; x_{i3} — число орудий поверхностной обработки почвы на 100 га; x_{i4} — количество удобрений, расходуемых на гектар (т/га); x_{i5} — количество химических средств защиты растений, расходуемых на гектар (ц/га)

i	y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}
1	9,7	1,59	0,26	2,05	0,32	0,14
2	8,4	0,34	0,28	0,46	0,59	0,66
3	9	2,53	0,31	2,46	0,3	0,31
4	9,9	4,63	0,4	6,44	0,43	0,59
5	9,6	2,16	0,26	2,16	0,39	0,16
6	8,6	2,16	0,3	2,69	0,32	0,17
7	12,5	0,68	0,29	0,73	0,42	0,23
8	7,6	0,35	0,26	0,42	0,21	0,08
9	6,9	0,52	0,24	0,49	0,2	0,08
10	13,5	3,42	0,31	3,02	1,37	0,73
11	9,7	1,78	0,3	3,19	0,73	0,17
12	10,7	2,4	0,32	3,3	0,25	0,14
13	12,1	9,36	0,4	11,51	0,39	0,38
14	9,7	1,72	0,28	2,26	0,82	0,17
15	7	0,59	0,29	0,6	0,13	0,35

Окончание табл. 3.3

i	y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}
16	7,2	0,28	0,26	0,3	0,09	0,15
17	8,2	1,64	0,29	1,44	0,2	0,08
18	8,4	0,09	0,22	0,05	0,43	0,2
19	13,1	0,08	0,25	0,03	0,73	0,2
20	8,7	1,36	0,26	0,17	0,99	0,42

Требуется:

1. Построить выборочное уравнение линейной множественной регрессии для исходных данных:
   - найти вектор коэффициентов b;
   - рассчитать общую сумму квадратов Q, сумму квадратов, объясненную регрессией Q_r, остаточную сумму квадратов Q_e, несмещенные оценки соответствующих дисперсий s^2, s_r^2, s_e^2 и средних квадратических отклонений s, s_r, s_e;
   - найти стандартные отклонения коэффициентов регрессии s_{b_j} и с доверительной вероятностью $\gamma = 0{,}95$ оценить значимость коэффициентов;
   - рассчитать выборочный множественный коэффициент детерминации $\overline{R}^2_{y.X}$ и его скорректированное значение $\overline{R}^2_{adj}$;
   - с доверительной вероятностью $\gamma = 0{,}95$ оценить значимость уравнения регрессии.
2. Проверить полученные результаты с помощью стандартной статистической программы ЛИНЕЙН и инструмента РЕГРЕССИЯ из пакета анализа Microsoft Excel.
3. Провести анализ признаков мультиколлинеарности в следующем порядке:
   - вычислить элементы матрицы выборочных парных коэффициентов корреляции $\mathbf{R}_{XX}$ объясняющих переменных и оценить их значения;

- найти значение определителя матрицы $\mathbf{R}_{XX}$ и с доверительной вероятностью $\gamma = 0{,}95$ проверить гипотезу о наличии мультиколлинеарности на основе статистического χ^2-критерия.

4. С помощью метода пошагового отбора переменных провести устранение мультиколлинеарности, отобрав наиболее информативные предикторы в регрессионной модели. Построить графики зависимостей $\overline{R}^2(l)$, $\overline{R}^{*2}(l)$ и $\overline{R}^2_{\min}(l)$ от номера шага l, полагая их равными нулю при $l = 0$.

5. Построить выборочное уравнение линейной множественной регрессии при учете только отобранных предикторов:

- найти вектор коэффициентов b;
- рассчитать общую сумму квадратов Q, сумму квадратов, объясненную регрессией Q_r, остаточную сумму квадратов Q_e, несмещенные оценки соответствующих дисперсий s^2, s_r^2, s_e^2 и средних квадратических отклонений s, s_r, s_e;
- найти стандартные отклонения коэффициентов регрессии s_{b_j} и с доверительной вероятностью $\gamma = 0{,}95$ оценить значимость коэффициентов;
- рассчитать выборочный множественный коэффициент детерминации $\overline{R}^2_{y.x}$ и его скорректированное значение $\overline{R}^2_{adj}$;
- с доверительной вероятностью $\gamma = 0{,}95$ оценить значимость уравнения регрессии.

6. Проверить полученные результаты с помощью стандартной статистической программы ЛИНЕЙН и инструмента РЕГРЕССИЯ из пакета анализа Microsoft Excel.

Решение

1. Для определения параметров выборочного уравнения линейной регрессии строим расчетную таблицу (табл. 3.4, столбцы 1–7).

Таблица 3.4

Расчетная таблица

i	y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	x_{i5}	$\hat{y}_i$	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})^2$	$(y_i - \hat{y}_i)^2$
1	2	3	4	5	6	7	8	9	10	11
1	9,7	1,59	0,26	2,05	0,32	0,14	8,79	0,03	0,54	0,82
2	8,4	0,34	0,28	0,46	0,59	0,66	8,62	1,27	0,82	0,05
3	9	2,53	0,31	2,46	0,3	0,31	9,02	0,28	0,25	0,00
4	9,9	4,63	0,4	6,44	0,43	0,59	10,60	0,14	1,17	0,50
5	9,6	2,16	0,26	2,16	0,39	0,16	9,06	0,01	0,22	0,30
6	8,6	2,16	0,3	2,69	0,32	0,17	9,39	0,86	0,02	0,63
7	12,5	0,68	0,29	0,73	0,42	0,23	9,30	8,85	0,05	10,22
8	7,6	0,35	0,26	0,42	0,21	0,08	8,30	3,71	1,49	0,50
9	6,9	0,52	0,24	0,49	0,2	0,08	7,96	6,89	2,46	1,11
10	13,5	3,42	0,31	3,02	1,37	0,73	12,63	15,80	9,66	0,75
11	9,7	1,78	0,3	3,19	0,73	0,17	11,28	0,03	3,10	2,51
12	10,7	2,4	0,32	3,3	0,25	0,14	9,54	1,38	0,00	1,34
13	12,1	9,36	0,4	11,51	0,39	0,38	11,57	6,63	4,18	0,28
14	9,7	1,72	0,28	2,26	0,82	0,17	11,27	0,03	3,06	2,48
15	7	0,59	0,29	0,6	0,13	0,35	7,64	6,38	3,55	0,41
16	7,2	0,28	0,26	0,3	0,09	0,15	7,55	5,41	3,90	0,12
17	8,2	1,64	0,29	1,44	0,2	0,08	8,83	1,76	0,48	0,40
18	8,4	0,09	0,22	0,05	0,43	0,2	8,28	1,27	1,56	0,02
19	13,1	0,08	0,25	0,03	0,73	0,2	10,08	12,78	0,31	9,10
20	8,7	1,36	0,26	0,17	0,99	0,42	10,76	0,68	1,54	4,26
Σ	190,5							74,16	38,36	35,80
ср.	9,5									
$\overline{\sigma}^2$								3,90	7,67	2,56
$\overline{\sigma}$								1,98	2,77	1,60

Записываем матрицу $\mathbf{X}$ объясняющих переменных размера 20×6 и находим транспонированную матрицу $\mathbf{X}^T$:

$$\mathbf{X} = \begin{pmatrix} 1 & 1{,}59 & 0{,}26 & 2{,}05 & 0{,}32 & 0{,}14 \\ 1 & 0{,}34 & 0{,}28 & 0{,}46 & 0{,}59 & 0{,}66 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 1{,}36 & 0{,}26 & 0{,}17 & 0{,}99 & 0{,}42 \end{pmatrix};$$

$$\mathbf{X}^T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 1{,}59 & 0{,}34 & \cdots & 1{,}36 \\ 0{,}26 & 0{,}28 & \cdots & 0{,}26 \\ 2{,}05 & 0{,}46 & \cdots & 0{,}17 \\ 0{,}32 & 0{,}59 & \cdots & 0{,}99 \\ 0{,}14 & 0{,}66 & \cdots & 0{,}42 \end{pmatrix}$$

Вычисляем произведение матриц $\mathbf{X}^T\mathbf{X}$

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 20 & 37{,}68 & 5{,}78 & 43{,}77 & 9{,}31 & 5{,}41 \\ 37{,}68 & 156{,}84 & 12{,}47 & 189{,}42 & 18{,}98 & 12{,}87 \\ 5{,}78 & 12{,}47 & 1{,}71 & 14{,}73 & 2{,}70 & 1{,}64 \\ 43{,}77 & 189{,}42 & 14{,}73 & 235{,}12 & 20{,}87 & 14{,}62 \\ 9{,}31 & 18{,}98 & 2{,}70 & 20{,}87 & 6{,}30 & 3{,}20 \\ 5{,}41 & 12{,}87 & 1{,}64 & 14{,}62 & 3{,}20 & 2{,}18 \end{pmatrix}$$

и произведение $\mathbf{X}^T\mathbf{y}$:

$$\mathbf{X}^T\mathbf{y} = \begin{pmatrix} 190{,}50 \\ 393{,}23 \\ 55{,}70 \\ 457{,}89 \\ 95{,}65 \\ 53{,}96 \end{pmatrix}.$$

Рассчитываем обратную матрицу $(\mathbf{X}^T\mathbf{X})^{-1}$:

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 11{,}48 & -0{,}42 & -45{,}16 & 0{,}91 & -1{,}29 & 3{,}77 \\ -0{,}42 & 0{,}34 & 1{,}78 & -0{,}28 & -0{,}08 & -0{,}29 \\ -45{,}16 & 1{,}78 & 180{,}84 & -3{,}76 & 4{,}40 & -15{,}78 \\ 0{,}91 & -0{,}28 & -3{,}76 & 0{,}27 & 0{,}02 & 0{,}39 \\ -1{,}29 & -0{,}08 & 4{,}40 & 0{,}02 & 0{,}93 & -1{,}11 \\ 3{,}77 & -0{,}29 & -15{,}78 & 0{,}39 & -1{,}11 & 3{,}73 \end{pmatrix}.$$

Определяем вектор оценок коэффициентов регрессии $\mathbf{b}$ по формуле (2.8):

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{pmatrix} = \begin{pmatrix} 3{,}515 \\ -0{,}006 \\ 15{,}542 \\ 0{,}110 \\ 4{,}475 \\ -2{,}933 \end{pmatrix}.$$

Подставляя рассчитанные значения b_j в выборочное уравнение регрессии $\hat{y}_i = b_0 + \sum_{j=1}^{5} b_j x_{ij}$, находим значения $\hat{y}_i$, (i = 1, 2, ..., n) и сводим их в 8-й столбец табл. 3.4.

Вычисляем значения $(y_i - \bar{y})^2$, $(\hat{y}_i - \bar{y})^2$, $(y_i - \hat{y}_i)^2$ (табл. 3.4, столбцы 9–11) и по формулам из табл. 2.1, 2.2 рассчитываем соответствующие суммы квадратов, дисперсии на степень свободы и средние квадратические отклонения:

$$Q = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 74{,}16;\ s^2 = \frac{Q}{n-1} = 3{,}90;\ s = 1{,}98;$$

$$Q_r = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 38{,}36;\ s_r^2 = \frac{Q_R}{p} = 7{,}67;\ s_r = 2{,}77;$$

$$Q_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 35{,}796;\ s_e^2 = \frac{Q_e}{n-p-1} = 2{,}56;\ s_e = 1{,}599.$$

Предварительно сформировав вспомогательный вектор из диагональных элементов $[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}$ матрицы $(\mathbf{X}^T\mathbf{X})^{-1}$, по формуле (2.16) находим векторы стандартных отклонений коэффициентов регрессии:

$$\mathbf{s_b} = s_e\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}} = \begin{pmatrix} s_{b_0} \\ s_{b_1} \\ s_{b_2} \\ s_{b_3} \\ s_{b_4} \\ s_{b_5} \end{pmatrix} = \begin{pmatrix} 5{,}42 \\ 0{,}93 \\ 21{,}50 \\ 0{,}83 \\ 1{,}54 \\ 3{,}09 \end{pmatrix}.$$

Записываем выборочное уравнение регрессии в общепринятом виде, указывая в скобках под коэффициентами их стандартные отклонения:

$$\hat{y}_i = \underset{(5{,}42)}{3{,}515} - \underset{(0{,}93)}{0{,}006}x_{i1} + \underset{(21{,}50)}{15{,}542}x_{i2} + \underset{(0{,}83)}{0{,}110}x_{i3} + \underset{(1{,}54)}{4{,}475}x_{i4} - \underset{(3{,}09)}{2{,}932}x_{i5}.$$

Как видно из уравнения, для всех коэффициентов, кроме b_4, стандартные отклонения превышают полученные значения оценок коэффициентов регрессии. Это позволяет предположить, что за исключением четвертого, все коэффициенты регрессии по t-критерию будут не значимыми.

Согласно (2.21) формируем вектор t-статистик $t_{b_j} = \dfrac{b_j}{s_{b_j}}$ критерия значимости коэффициентов регрессии

$$\mathbf{t_b} = \begin{pmatrix} t_{b_0} \\ t_{b_1} \\ t_{b_2} \\ t_{b_3} \\ t_{b_4} \\ t_{b_5} \end{pmatrix} = \begin{pmatrix} 0{,}649 \\ -0{,}007 \\ 0{,}723 \\ 0{,}132 \\ 2{,}899 \\ -0{,}950 \end{pmatrix}.$$

Для заданной доверительной вероятности $\gamma = 0{,}95$ (уровня значимости $\alpha = 0{,}05$) находим значение $t_{\text{кр}}$ с помощью стандартной статистической функции СТЬЮДРАСПОБР $(a;\ n-p-1)$

$t_{\text{кр}} = t_{\text{кр}}(\alpha; k = n-p-1) = 2{,}145,\ (n-p-1 = 20-5-1 = 14).$

Поскольку только $|t_{b_4}| > t_{\text{кр}}$, с доверительным уровнем 95% делаем вывод о том, что коэффициент β_4 значим, а остальные не значимы.

Вычисляем вектор P-значений P_{b_j} для коэффициентов с помощью статистической функции СТЬЮДРАСП $(|t_{b_j}|;\ n-p-1;\ 2)$

$$\mathbf{P_b} = \begin{pmatrix} P_{b_0} \\ P_{b_1} \\ P_{b_2} \\ P_{b_3} \\ P_{b_4} \\ P_{b_5} \end{pmatrix} = \begin{pmatrix} 0{,}527 \\ 0{,}995 \\ 0{,}482 \\ 0{,}897 \\ 0{,}012 \\ 0{,}358 \end{pmatrix}.$$

В силу того, что только $P_{b_4} < \alpha$, полученный с помощью t-критерия вывод подтверждается.

По формулам (2.25), (2.34) рассчитываем величину выборочного множественного коэффициента детерминации и его скорректированного значения

$$\overline{R}^2_{y.X} = \frac{Q_r}{Q} = 1 - \frac{Q_e}{Q} = 0{,}517;$$

$$\overline{R}^2_{adj} = 1 - \frac{Q_e/(n-p-1)}{Q/(n-1)} = 1 - (1-\overline{R}^2_{y.X})\frac{n-1}{n-p-1} = 0{,}345.$$

Величина коэффициента $\overline{R}^2_{y.X}$ показывает, что около 52% вариации зависимой переменной обусловлены влиянием включенных факторов, а остальные 48% — влиянием других неучтенных в модели и случайных факторов.

Согласно (3.35) рассчитываем значение F-статистики

$$F = \frac{s_r^2}{s_e^2} = \frac{Q_r(n-p-1)}{Q_e p} = \frac{\overline{R}^2_{y.X}(n-p-1)}{(1-\overline{R}^2_{y.X})p} = 3{,}001.$$

Критическое значение для доверительного уровня $\gamma = 0{,}95$ (уровня значимости $\alpha = 0{,}05$) находим с помощью стандартной статистической функции FРАСПОБР $(a;\ p;\ n-p-1)$

$$F_{кр} = F_{кр}(\alpha; k_1 = p, k_2 = n-p-1) = 2{,}96.$$

В силу того, что $F > F_{кр}$, с доверительным уровнем 0,95 делаем вывод о том, что уравнение регрессии значимо.

Вычисляем величину P-значения с помощью стандартной статистической функции FРАСП $(F;\ p;\ n-p-1)$

$$P = 0{,}048$$

и, поскольку $P < \alpha$, вывод о значимости уравнения регрессии подтверждается.

2. Проверяем полученные результаты с помощью стандартной статистической программы ЛИНЕЙН и инструмента РЕГРЕССИЯ из пакета анализа Microsoft Excel.

Анализ полученных результатов показывает, что имеются типичные следствия мультиколлинеарности:

- коэффициенты b_1 и b_5, соответственно характеризующие зависимости урожайности зерновых культур от числа тракторов (приведенной мощности) на 100 га и количества химических средств защиты растений, расходуемых на гектар, имеют отрицательные знаки;
- за исключением b_4, все коэффициенты регрессии по t-критерию не значимы, в то время как модель регрессии в целом по F-критерию является значимой.

3. Проводим анализ признаков мультиколлинеарности в следующем порядке.

Вычисляем элементы матрицы выборочных парных коэффициентов корреляции $\mathbf{R}_{XX}$ объясняющих переменных с помощью статистической функции КОРРЕЛ $(\mathbf{x}_j;\ \mathbf{x}_k)$ $(j,\ k = 1,\ 2,\ \ldots,\ p)$

$$\mathbf{R}_{XX} = \begin{pmatrix} 1 & \overline{r}_{12} & \overline{r}_{13} & \overline{r}_{14} & \overline{r}_{15} \\ \overline{r}_{21} & 1 & \overline{r}_{23} & \overline{r}_{24} & \overline{r}_{25} \\ \overline{r}_{31} & \overline{r}_{32} & 1 & \overline{r}_{34} & \overline{r}_{35} \\ \overline{r}_{41} & \overline{r}_{42} & \overline{r}_{43} & 1 & \overline{r}_{45} \\ \overline{r}_{51} & \overline{r}_{52} & \overline{r}_{53} & \overline{r}_{54} & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & 0{,}85 & 0{,}98 & 0{,}11 & 0{,}34 \\ 0{,}85 & 1 & 0{,}88 & 0{,}03 & 0{,}46 \\ 0{,}98 & 0{,}88 & 1 & 0{,}03 & 0{,}28 \\ 0{,}11 & 0{,}03 & 0{,}03 & 1 & 0{,}57 \\ 0{,}34 & 0{,}46 & 0{,}28 & 0{,}57 & 1 \end{pmatrix},$$

а затем проверяем полученные значения, используя инструмент КОРРЕЛЯЦИЯ из пакета анализа.

Находим значение определителя матрицы $\mathbf{R}_{XX}$ с помощью математической функции МОПРЕД ($\mathbf{R}_{XX}$)

$$\det \mathbf{R}_{XX} = 0{,}003.$$

Малое значение определителя свидетельствует о возможном наличии частичной мультиколлинеарности.

Проверяем гипотезу о наличии мультиколлинеарности, для чего согласно (3.3) формируем статистику

$$\chi^2 = -\left[n-1-\frac{1}{6}(2p+5)\right]\ln\det \mathbf{R}_{XX} = 95{,}601$$

и для заданной доверительной вероятности $\gamma = 0{,}95$ (уровня значимости $\alpha = 0{,}05$) находим значение $\chi^2_{\text{кр}}$ с помощью стандартной статистической функции ХИ2ОБР $(a;\ \frac{1}{2}p(p-1))$

$$\chi^2_{\text{кр}} = \chi^2_{\text{кр}}[\alpha; k = \frac{1}{2}p(p-1)] = 18{,}307,$$

$$k = \frac{1}{2}p(p-1) = \frac{5\cdot 4}{2} = 10).$$

Поскольку $\chi^2 > \chi^2_{\text{кр}}$, с доверительным уровнем $\gamma = 0{,}95$ принимаем альтернативную гипотезу H_1 и делаем вывод о том, что в регрессионной модели имеется мультиколлинеарность.

Вычисляем величину P-значения с помощью стандартной статистической функции ХИ2РАСП $(\chi^2; \frac{1}{2}p(p-1))$:

$$P = 4{,}121\cdot 10^{-16}.$$

В силу того, что $P < \alpha$, вывод о наличии мультиколлинеарности подтверждается.

4. С помощью метода пошагового отбора переменных проводим устранение мультиколлинеарности.

а) Определяем исходные данные при учете всех факторных переменных (при $l = p = 5$).

Вычисляем элементы матрицы выборочных парных коэффициентов корреляции $\mathbf{R}_{yX}$ с помощью статистической функции КОРРЕЛ $(\mathbf{y}; \mathbf{x}_j)$,$(j = 1, 2, \ldots, p)$ и проверяем полученные значения, используя инструмент КОРРЕЛЯЦИЯ из пакета анализа

$$\mathbf{R}_{yX} = \begin{pmatrix} 1 & \bar{r}_{01} & \bar{r}_{02} & \bar{r}_{03} & \bar{r}_{04} & \bar{r}_{05} \\ \bar{r}_{10} & 1 & \bar{r}_{12} & \bar{r}_{13} & \bar{r}_{14} & \bar{r}_{15} \\ \bar{r}_{20} & \bar{r}_{21} & 1 & \bar{r}_{23} & \bar{r}_{24} & \bar{r}_{25} \\ \bar{r}_{30} & \bar{r}_{31} & \bar{r}_{32} & 1 & \bar{r}_{34} & \bar{r}_{35} \\ \bar{r}_{40} & \bar{r}_{41} & \bar{r}_{42} & \bar{r}_{43} & 1 & \bar{r}_{45} \\ \bar{r}_{50} & \bar{r}_{51} & \bar{r}_{52} & \bar{r}_{53} & \bar{r}_{54} & 1 \end{pmatrix} =$$

$$= \begin{pmatrix} 1 & 0{,}43 & 0{,}37 & 0{,}40 & \underline{0{,}58} & 0{,}33 \\ 0{,}43 & 1 & 0{,}85 & 0{,}98 & 0{,}11 & 0{,}34 \\ 0{,}37 & 0{,}85 & 1 & 0{,}88 & 0{,}03 & 0{,}46 \\ 0{,}40 & 0{,}98 & 0{,}88 & 1 & 0{,}03 & 0{,}28 \\ \underline{0{,}58} & 0{,}11 & 0{,}03 & 0{,}03 & 1 & 0{,}57 \\ 0{,}33 & 0{,}34 & 0{,}46 & 0{,}28 & 0{,}57 & 1 \end{pmatrix}.$$

С помощью математической функции МОПРЕД ($\mathbf{R}_{yX}$) находим значение $\det \mathbf{R}_{yX} = 0{,}0015$ и, учитывая, что величина $\Delta_{00} = = \det \mathbf{R}_{XX} = 0{,}0030$, вычисляем множественный выборочный коэффициент детерминации 5-го порядка по формуле (3.12)

$$\bar{R}^2_{y\cdot X} = \bar{R}^2(5) = 1 - \frac{\det \mathbf{R}_{yX}}{\Delta_{00}} = 0{,}517.$$

Затем по формулам (3.13), (3.14) определяем скорректированный выборочный коэффициент детерминации

$$\overline{R}^2_{adj} = \overline{R}^{*2}(5) = 1 - [1 - \overline{R}^2(5)]\frac{n-1}{n-5-1} = 0{,}345$$

и вычисляем нижнюю доверительную границу

$$\overline{R}^2_{\min}(5) = \overline{R}^{*2}(5) - 2\sqrt{\frac{2l(n-5-1)}{(n-1)(n^2-1)}}[1 - \overline{R}^2(5)] = 0{,}214.$$

Понятно, что величины $\overline{R}^2_{y.X}$ и $\overline{R}^2_{adj}$ совпадают со значениями коэффициентов, полученных ранее при построении выборочного уравнения регрессии.

б) В качестве оптимального числа l_{opt} предикторов регрессионной модели выбираем то значение l, при котором величина $\overline{R}^2_{\min}(l)$ по мере добавления объясняющих переменных достигает своего максимума, осуществляя этот процесс по шагам.

1-й шаг ($l = 1$):

- среди всех возможных предикторов выбираем объясняющую переменную x_4, которая имеет наибольшее значение выборочного коэффициента парной корреляции $\overline{r}_{04} = 0{,}58$, находящегося в матрице $\mathbf{R}_{yX}$ на 4-й позиции в 0-й строке или столбце (это значение подчеркнуто);
- рассчитываем коэффициент детерминации 1-го порядка

$$\overline{R}^2(1) = \overline{r}^2_{04} = 0{,}333,$$

его скорректированное значение

$$\overline{R}^{*2}(1) = 1 - [1 - \overline{R}^2(1)]\frac{n-1}{n-1-1} = 0{,}296$$

и нижнюю доверительную границу

$$\overline{R}^2_{\min}(1) = \overline{R}^{*2}(1) - 2\sqrt{\frac{2l(n-1-1)}{(n-1)(n^2-1)}}[1 - \overline{R}^2(1)] = 0{,}204.$$

2-й шаг ($l = 2$):

- строим подматрицы $\mathbf{R}_{yx_4x_k}$, которые находятся из матрицы $\mathbf{R}_{yX}$ путем вычеркивания всех строк и столбцов, кроме тех, которые отвечают y, x_4 и x_k ($k = 1, 2, 3, 5$)

$$\mathbf{R}_{yx_4x_1} = \begin{pmatrix} 1 & 0{,}43 & 0{,}58 \\ 0{,}43 & 1 & 0{,}11 \\ 0{,}58 & 0{,}11 & 1 \end{pmatrix}; \quad \mathbf{R}_{yx_4x_2} = \begin{pmatrix} 1 & 0{,}37 & 0{,}58 \\ 0{,}37 & 1 & 0{,}03 \\ 0{,}58 & 0{,}03 & 1 \end{pmatrix};$$

$$\mathbf{R}_{yx_4x_3} = \begin{pmatrix} 1 & 0{,}40 & 0{,}58 \\ 0{,}40 & 1 & 0{,}03 \\ 0{,}58 & 0{,}03 & 1 \end{pmatrix}; \quad \mathbf{R}_{yx_4x_5} = \begin{pmatrix} 1 & 0{,}58 & 0{,}33 \\ 0{,}58 & 1 & 0{,}57 \\ 0{,}33 & 0{,}57 & 1 \end{pmatrix};$$

- с помощью математической функции МОПРЕД (·) находим значения соответствующих определителей и алгебраических дополнений 1-го элемента 1-й строки полученных подматриц

$$\det \mathbf{R}_{yx_4x_1} = 0{,}524; \ \Delta_{11} = 0{,}988;$$

$$\det \mathbf{R}_{yx_4x_2} = 0{,}538; \ \Delta_{11} = 0{,}999;$$

$$\det \mathbf{R}_{yx_4x_3} = 0{,}517; \ \Delta_{11} = 0{,}999;$$

$$\det \mathbf{R}_{yx_4x_5} = 0{,}450; \ \Delta_{11} = 0{,}674;$$

- по формуле $\overline{R}^2_{y/x_4x_k} = 1 - \dfrac{\det \mathbf{R}_{yx_4x_k}}{\Delta_{11}}$ вычисляем соответствующие выборочные множественные коэффициенты детерминации 2-го порядка

$$\overline{R}^2_{y/x_4x_1} = 0{,}470;\ \overline{R}^2_{y/x_4x_2} = 0{,}462;$$

$$\overline{R}^2_{y/x_4x_3} = 0{,}482;\ \overline{R}^2_{y/x_4x_5} = 0{,}333$$

и выбираем объясняющую переменную x_3, имеющую наибольшее значение коэффициента детерминации

$$\overline{R}^2(2) = \overline{R}^2_{y/x_4x_3} = 0{,}482;$$

- рассчитываем соответствующий скорректированный коэффициент детерминации, нижнюю доверительную границу

$$\overline{R}^{*2}(2) = 1 - [1 - \overline{R}^2(2)]\frac{n-1}{n-2-1} = 0{,}421;$$

$$\overline{R}^2_{\min}(2) = \overline{R}^{*2}(2) - 2\sqrt{\frac{2l(n-2-1)}{(n-1)(n^2-1)}}[1-\overline{R}^2(2)] = 0{,}323$$

и, поскольку $\overline{R}^2_{\min}(2) > \overline{R}^2_{\min}(1)$, переходим к следующему шагу.

3-й шаг ($l = 3$):

- строим подматрицы $\mathbf{R}_{yx_4x_3x_k}$, которые находятся из матрицы $\mathbf{R}_{yX}$ путем вычеркивания всех строк и столбцов, кроме тех, которые отвечают y, x_4, x_3 и x_k ($k = 1, 2, 5$):

$$\mathbf{R}_{yx_4x_3x_1} = \begin{pmatrix} 1 & 0{,}43 & 0{,}40 & 0{,}58 \\ 0{,}43 & 1 & 0{,}98 & 0{,}11 \\ 0{,}40 & 0{,}98 & 1 & 0{,}03 \\ 0{,}58 & 0{,}11 & 0{,}03 & 1 \end{pmatrix};$$

$$\mathbf{R}_{yx_4x_3x_2} = \begin{pmatrix} 1 & 0{,}37 & 0{,}40 & 0{,}58 \\ 0{,}37 & 1 & 0{,}88 & 0{,}03 \\ 0{,}40 & 0{,}88 & 1 & 0{,}03 \\ 0{,}58 & 0{,}03 & 0{,}03 & 1 \end{pmatrix};$$

$$\mathbf{R}_{yx_4x_3x_5} = \begin{pmatrix} 1 & 0{,}40 & 0{,}58 & 0{,}33 \\ 0{,}40 & 1 & 0{,}03 & 0{,}28 \\ 0{,}58 & 0{,}03 & 1 & 0{,}57 \\ 0{,}33 & 0{,}28 & 0{,}57 & 1 \end{pmatrix};$$

- с помощью математической функции МОПРЕД (·) находим значения соответствующих определителей и алгебраических дополнений 1-го элемента 1-й строки полученных подматриц

$$\det \mathbf{R}_{yx_4x_3x_1} = 0{,}019;\ \Delta_{11} = 0{,}037;$$
$$\det \mathbf{R}_{yx_4x_3x_2} = 0{,}115;\ \Delta_{11} = 0{,}222;$$
$$\det \mathbf{R}_{yx_4x_3x_5} = 0{,}304;\ \Delta_{11} = 0{,}606;$$

- по формуле $\overline{R}^2_{y/x_4x_3x_k} = 1 - \dfrac{\det \mathbf{R}_{yx_4x_3x_k}}{\Delta_{11}}$ вычисляем соответствующие выборочные множественные коэффициенты детерминации 3-го порядка

$$\overline{R}^2_{y/x_4x_3x_1} = 0{,}485;\ \overline{R}^2_{y/x_4x_3x_2} = 0{,}484;\ \overline{R}^2_{y/x_4x_3x_5} = 0{,}498$$

и выбираем объясняющую переменную x_5, имеющую наибольшее значение коэффициента детерминации

$$\overline{R}^2(2) = \overline{R}^2_{y/x_4x_3x_5} = 0{,}498;$$

- рассчитываем соответствующий скорректированный коэффициент детерминации, нижнюю доверительную границу

$$\overline{R}^{*2}(3) = 1 - [1 - \overline{R}^2(3)]\frac{n-1}{n-3-1} = 0{,}404;$$

$$\overline{R}^2_{\min}(3) = \overline{R}^{*2}(3) - 2\sqrt{\frac{2l(n-3-1)}{(n-1)(n^2-1)}}[1-\overline{R}^2(3)] = 0{,}291$$

и, поскольку $\overline{R}^2_{\min}(3) < \overline{R}^2_{\min}(2)$, ограничиваемся $l_{opt} = 2$ объясняющими переменными x_3 и x_4, отобранными на 2-м шаге;

- строим графики зависимостей $\overline{R}^2(l)$, $\overline{R}^{*2}(l)$ и $\overline{R}^2_{\min}(l)$ от номера шага l (рис. 3.8).

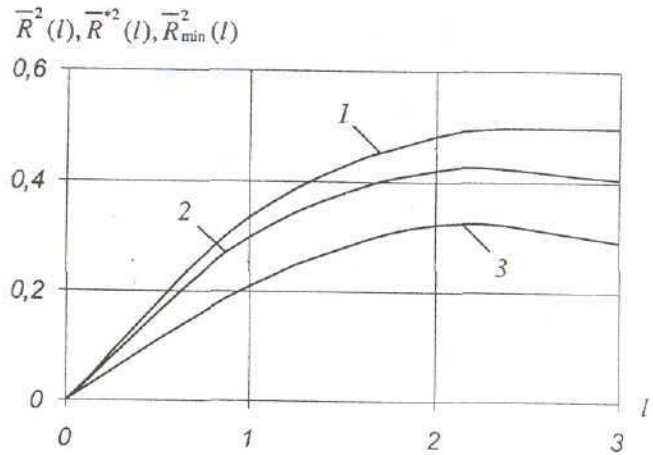

Рис. 3.8. Графики зависимостей выборочных коэффициентов детерминации от номера шага l: $1 - \overline{R}^2(l)$; $2 - \overline{R}^{*2}(l)$; $3 - \overline{R}^2_{\min}(l)$

5. Для определения параметров выборочного уравнения линейной регрессии при учете только $l_{opt} = 2$ отобранных объясняющих переменных x_{i3} и x_{i4} строим расчетную таблицу (табл. 3.5, столбцы 1–4).

Строим матрицу $\mathbf{X}$ объясняющих переменных размером 20×3 и транспонированную матрицу $\mathbf{X}^T$

$$\mathbf{X} = \begin{pmatrix} 1 & 2{,}05 & 0{,}32 \\ 1 & 0{,}46 & 0{,}59 \\ \dots & \dots & \dots \\ 1 & 0{,}17 & 0{,}99 \end{pmatrix}; \quad \mathbf{X}^T = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 2{,}05 & 0{,}46 & \dots & 0{,}17 \\ 0{,}32 & 0{,}59 & \dots & 0{,}99 \end{pmatrix}.$$

Таблица 3.5

Расчетная таблица

i	y_i	x_{i3}	x_{i4}	$\hat{y}_i$	$(y_i - \overline{y})^2$	$(\hat{y}_i - \overline{y})^2$	$(y_i - \hat{y}_i)^2$
1	9,7	2,05	0,32	8,98	0,03	0,30	0,52
2	8,4	0,46	0,59	9,47	1,27	0,00	1,15
3	9	2,46	0,3	9,03	0,28	0,25	0,00
4	9,9	6,44	0,43	10,60	0,14	1,16	0,49
5	9,6	2,16	0,39	9,25	0,01	0,07	0,12
6	8,6	2,69	0,32	9,16	0,86	0,13	0,31
7	12,5	0,73	0,42	8,96	8,85	0,32	12,56
8	7,6	0,42	0,21	8,14	3,71	1,92	0,29
9	6,9	0,49	0,2	8,12	6,89	1,96	1,50
10	13,5	3,02	1,37	12,90	15,80	11,41	0,36
11	9,7	3,19	0,73	10,73	0,03	1,44	1,05
12	10,7	3,3	0,25	9,09	1,38	0,19	2,59
13	12,1	11,51	0,39	11,89	6,63	5,59	0,04
14	9,7	2,26	0,82	10,78	0,03	1,57	1,16
15	7	0,6	0,13	7,91	6,38	2,60	0,83
16	7,2	0,3	0,09	7,69	5,41	3,37	0,24
17	8,2	1,44	0,2	8,39	1,76	1,28	0,04
18	8,4	0,05	0,43	8,80	1,27	0,53	0,16
19	13,1	0,03	0,73	9,84	12,78	0,10	10,66
20	8,7	0,17	0,99	10,78	0,68	1,57	4,32
Σ	190,5				74,16	35,77	38,39
ср.	9,53				74,16		
$\overline{\sigma^2}$					3,90	17,89	2,26
$\overline{\sigma}$					1,98	4,23	1,50

Находим произведение матриц $\mathbf{X}^T\mathbf{X}$ и $\mathbf{X}^T\mathbf{y}$

$$\mathbf{X}^T\mathbf{X} = \begin{pmatrix} 20 & 43{,}77 & 9{,}31 \\ 43{,}77 & 235{,}12 & 20{,}87 \\ 9{,}31 & 20{,}87 & 6{,}30 \end{pmatrix}; \quad \mathbf{X}^T\mathbf{y} = \begin{pmatrix} 190{,}50 \\ 457{,}89 \\ 95{,}65 \end{pmatrix}.$$

Рассчитываем обратную матрицу $(\mathbf{X}^T\mathbf{X})^{-1}$

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 0{,}191 & -0{,}015 & -0{,}233 \\ -0{,}015 & 0{,}007 & 0{,}002 \\ -0{,}233 & -0{,}002 & 0{,}509 \end{pmatrix}.$$

По формуле (2.8) определяем вектор оценок коэффициентов регрессии $\mathbf{b}$

$$\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \begin{pmatrix} b_0 \\ b_3 \\ b_4 \end{pmatrix} = \begin{pmatrix} 7{,}291 \\ 0{,}282 \\ 3{,}475 \end{pmatrix}.$$

Подставляя рассчитанные значения b_j в выборочное уравнение регрессии $\hat{y}_i = b_0 + b_3 x_{i3} + b_4 x_{i4}$, находим значения $\hat{y}_i$, $(i = 1, 2, \ldots, n)$ и сводим их в 5-й столбец табл. 3.5.

Вычисляем значения $(y_i - \bar{y})^2$, $(\hat{y}_i - \bar{y})^2$, $(y_i - \hat{y}_i)^2$ (табл. 3.5, столбцы 6–8) и по формулам из табл. 2.1, 2.2 рассчитываем соответствующие суммы квадратов, дисперсии на степень свободы и средние квадратические отклонения

$$Q = \sum_{i=1}^{n}(y_i - \bar{y})^2 = 74{,}16;\ s^2 = \frac{Q}{n-1} = 3{,}90;\ s = 1{,}98;$$

$$Q_r = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = 35{,}77;\ s_r^2 = \frac{Q_r}{p} = 17{,}89;\ s_r = 4{,}23;$$

$$Q_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 38{,}39;\ s_e^2 = \frac{Q_e}{n-p-1} = 2{,}26;\ s_e = 1{,}50.$$

Предварительно сформировав вспомогательный вектор из диагональных элементов $[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}$ матрицы $(\mathbf{X}^T\mathbf{X})^{-1}$, по формуле (2.16) находим векторы стандартных отклонений коэффициентов регрессии

$$\mathbf{s_b} = s_e\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}} = \begin{pmatrix} s_{b_0} \\ s_{b_3} \\ s_{b_4} \end{pmatrix} = \begin{pmatrix} 0{,}66 \\ 0{,}13 \\ 1{,}07 \end{pmatrix}.$$

Записываем выборочное уравнение регрессии в общепринятом виде, указывая в скобках под коэффициентами их стандартные отклонения

$$\hat{y}_i = \underset{(0{,}66)}{7{,}291} + \underset{(0{,}13)}{0{,}282}\,x_{i3} + \underset{(1{,}07)}{3{,}475}\,x_{i4}.$$

Как видно из уравнения, для всех коэффициентов стандартные отклонения меньше полученных значений оценок коэффициентов регрессии. Это позволяет предположить, что все коэффициенты регрессии по t-критерию будут значимыми.

Согласно (2.21) формируем вектор t-статистик $t_{b_j} = \dfrac{b_j}{s_{b_j}}$ критерия значимости коэффициентов регрессии

$$\mathbf{t_b} = \begin{pmatrix} t_{b_0} \\ t_{b_3} \\ t_{b_4} \end{pmatrix} = \begin{pmatrix} 11{,}10 \\ 2{,}21 \\ 3{,}24 \end{pmatrix}.$$

Для заданной доверительной вероятности $\gamma = 0{,}95$ (уровня значимости $\alpha = 0{,}05$) находим значение $t_{кр}$ с помощью стандартной статистической функции СТЬЮДРАСПОБР $(\alpha;\ n-p-1)$

$$t_{кр} = t_{кр}(\alpha; k = n-p-1) = 2{,}11,\ (n-p-1 = 20-2-1 = 17).$$

Поскольку $|t_{b_j}| > t_{кр}$, с доверительным уровнем 95% делаем вывод о том, что все коэффициенты β_j значимы.

Вычисляем вектор P-значений P_{b_j} для коэффициентов с помощью стандартной функции СТЬЮДРАСП $(|t_{b_j}|;\ n-p-1;\ 2)$

$$\mathbf{P_b} = \begin{pmatrix} P_{b_0} \\ P_{b_3} \\ P_{b_4} \end{pmatrix} = \begin{pmatrix} 3{,}28 \cdot 10^{-9} \\ 0{,}041 \\ 0{,}005 \end{pmatrix}.$$

В силу того, что $P_{b_j} < \alpha$, полученный с помощью t-критерия вывод о значимости коэффициентов β_0, β_3 и β_4 подтверждается.

По формулам (2.25), (2.34) рассчитываем величину выборочного множественного коэффициента детерминации и его скорректированного значения

$$\overline{R}^2_{y.x} = \frac{Q_r}{Q} = 1 - \frac{Q_e}{Q} = 0{,}482;$$

$$\overline{R}^2_{adj} = 1 - \frac{Q_e/(n-p-1)}{Q/(n-1)} = 1 - (1 - \overline{R}^2_{y.x})\frac{n-1}{n-p-1} = 0{,}421.$$

Величина коэффициента $\overline{R}^2_{y.x}$ показывает, что более 48% вариации зависимой переменной обусловлены влиянием включенных факторов, а остальные 52% — влиянием других неучтенных в модели и случайных факторов.

Используя (2.35), рассчитываем значение F-статистики

$$F = \frac{s_r^2}{s_e^2} = \frac{Q_r(n-p-1)}{Q_e p} = \frac{\overline{R}^2_{y.x}(n-p-1)}{(1-\overline{R}^2_{y.x})p} = 7{,}92.$$

Критическое значение для доверительного уровня $\gamma = 0{,}95$ (уровня значимости $\alpha = 0{,}05$) находим с помощью стандартной статистической функции FРАСПОБР (a; p; $n-p-1$)

$$F_{кр} = F_{кр}(\alpha; k_1 = p, k_2 = n-p-1) = 3{,}59.$$

В силу того, что $F > F_{кр}$, с доверительным уровнем 0,95 делаем вывод о том, что уравнение регрессии значимо.

Вычисляем величину $P = 0{,}0037$ с помощью стандартной статистической функции FРАСП (F; p; $n-p-1$) и, поскольку $P < \alpha$, этот вывод подтверждается.

6. Проверяем полученные результаты с помощью стандартной статистической программы ЛИНЕЙН и инструмента РЕГРЕССИЯ из пакета анализа Microsoft Excel.

Задача 3.2

Имеются экономические данные по экономическому развитию некоторой крупной страны за определенный период (i — условный номер года) (табл. 3.6).

Таблица 3.6

Экономические данные: y_i, — общее число занятых в экономике (тыс. чел.); x_{i1} — дефлятор (индекс) цен (%); x_{i2} — валовой национальный продукт (млрд дол.); x_{i3} — общее число безработных (тыс. чел.); x_{i4} — число военнослужащих (тыс. чел.); x_{i5} — неработающее население от 14 лет (тыс. чел.) $\overline{R}^2(l)$; 3 — $\overline{R}^2_{\min}(l)$

i	y_i	y_{i1}	y_{i1}	y_{i1}	y_{i1}	y_{i1}
1	60323	83	234,289	2356	1590	107608
2	61122	88,5	259,426	2325	1456	108632
3	60171	88,2	258,054	3682	1616	109773
4	61187	89,5	284,599	3351	1650	110929
5	63221	96,2	328,975	2099	3099	112075
6	63639	98,1	346,999	1932	3594	113270
7	64989	99	365,385	1870	3547	115094
8	63761	100	363,112	3578	3350	116219
9	66019	101,2	387,469	2904	3048	117388
10	67857	104,6	419,18	2822	2857	118734
11	68169	108,4	442,769	2936	2798	120445
12	66513	110,8	444,546	4681	2637	121950