Этапы корреляционно-регрессионного анализа данных

Полученные в результате каких-либо исследований данные почти всегда представлены в виде таблиц. Числовые данные, содержащиеся в таблицах, обычно имеют между собой явные (известные) или неявные (скрытые) связи.

Явно связаны показатели, которые получены методами прямого счета, т. е. вычислены по заранее известным формулам. Например, проценты выполнения плана, уровни, удельные веса, отклонения в сумме, отклонения в процентах, темпы роста, темпы прироста, индексы и т. д.

Связи же второго типа заранее неизвестны. Однако люди должны уметь объяснять и предсказывать (прогнозировать) сложные явления, поведение объектов или систем для того, чтобы управлять ими. Поэтому специалисты с помощью наблюдений стремятся выявить скрытые зависимости и выразить их в виде формул, т. е. математически смоделировать явления или процессы. Одну из таких возможностей предоставляет корреляционно-регрессионный анализ.

Пользуясь методами корреляционно-регрессионного анализа, аналитики измеряют тесноту связей показателей с помощью коэффициента корреляции. При этом обнаруживаются связи, различные по силе (сильные, слабые, умеренные и др.) и различные по направлению (прямые, обратные). Если связи окажутся существенными, то целесообразно будет найти их математическое выражение в виде регрессионной модели и оценить статистическую значимость модели. В экономике значимое уравнение используется, как правило, для прогнозирования изучаемого явления или показателя.

Корреляционно-регрессионный анализ связей между переменными показывает, как один набор переменных (X) может влиять на другой набор (У). Таким образом, регрессионные вычисления и подбор хороших уравнений - это ценный, универсальный исследовательский инструмент в самых разнообразных отраслях деловой и научной деятельности (маркетинг, торговля, медицина, техника и т. д.).

Рассмотрим этапы корреляционно-регрессионного анализа данных.

Нулевой этап - это сбор данных. Как в строительстве нулевой цикл обеспечивает фундамент будущему зданию, так в корреляционно-

регрессионном анализе решающую роль играет качество данных. Сбор данных создает фундамент прогнозам. Поэтому имеется ряд требований и правил, которые следует соблюдать при сборе данных.

Данные должны быть наблюдаемыми, т. е. полученными в результате замера, а не расчета. Наблюдения следует спланировать. Чем больше неодинаковых (не повторяющихся) данных, и чем они однороднее, тем лучше получится уравнение, если связи существенны. Подозрительные данные могут быть вызваны ошибками наблюдений и экспериментов.

Первый этап - корреляционный анализ. Его цель - определить характер связи (прямая, обратная) и силу связи (связь отсутствует, связь слабая, умеренная, заметная, сильная, весьма сильная, полная связь). Корреляционный анализ создает информацию о характере и степени выраженности связи (коэффициент корреляции), которая используется для отбора существенных факторов, а также для планирования эффективной последовательности расчета параметров регрессионных уравнений. При одном факторе вычисляют коэффициент корреляции, а при наличии нескольких факторов строят корреляционную матрицу, из которой выясняют два вида связей: связи зависимой переменной с независимыми, связи между самими независимыми.

Для определения тесноты связи результативного признака ***y*** с определяющими его ***x1***, ***x2***, ***x3*** и влияния определяющих признаков друг на друга вычисляются парные коэффициенты корреляции r_{yx1}, r_{yx2}, r_{yx3}, r_{x1x2}, r_{x1x3}, r_{x2x3}.

$$r_{yx_1}=\frac{\frac{1}{m}\sum_{j=1}^{m}\left(\left(y_j-\overline{y}\right)\left(x_{1j}-\overline{x}_1\right)\right)}{\sigma_y\sigma_{x1}}$$

, где $\overline{x_1},\overline{y}$- выборочное среднее значение переменных x_1, y по всем группам значений переменных, $\sigma_y=\sqrt{\frac{\sum_{j=1}^{m}(y_j-\overline{y})^2}{m}}$, $\sigma_{x1}=\sqrt{\frac{\sum_{j=1}^{m}(x_{1_j}-\overline{x_1})^2}{m}}$, где m – количество групп исследуемых изделий, размерность выборки. Аналогично рассчитываются значения r_{yx2}, r_{yx3} по

приведённым выше формулам, только переменная x_1 заменяется на x_2 или на x_3 соответственно. Аналогично рассчитываются коэффициенты корреляции определяющих признаков друг на друга:

$$r_{x1x2}=\frac{\frac{1}{m}\sum_{j=1}^{m}((x_{1_j}-\overline{x_1})(x_{2j}-\overline{x}_2))}{\sigma_{x1}\sigma_{x2}},$$

где $\overline{x_1},\overline{x_2}$- выборочное среднее значение переменных x_1, x_2 по всем группам значений переменных,

$\sigma_{x2}=\sqrt{\frac{\sum_{j=1}^{m}(x_{2j}-\overline{x_2})^2}{m}}$, $\sigma_{x1}=\sqrt{\frac{\sum_{j=1}^{m}(x_{1_j}-\overline{x_1})^2}{m}}$, где m – количество групп исследуемых изделий, размерность выборки.

В библиотеке функций табличного процессора MS Excel есть статистическая функция КОРРЕЛ() для вычисления парных коэффициентов корреляции.

Второй этап - расчет параметров и построение регрессионных моделей. Здесь стремятся отыскать наиболее точную меру выявленной связи, для того чтобы можно было прогнозировать, предсказывать значения зависимой величины Y, если будут известны значения независимых величин X_1, X_2, …, X_n. Эту меру можно выразить математической моделью линейной множественной регрессионной зависимости: $y = a_0 + b_1x_1 + b_2x_2 + \ldots + b_n x_n$.

Для нахождения коэффициентов $a_0,a_1,a_2,..a_n$ используют метод наименьших квадратов. Смысл метода в том, что строят функционал:

$$S(a_0,a_1\ldots a_n)=\sum_{i=1}^{m}¿(Q(x_{1i},x_{2i},x_{3i},a_0,a_1,\ldots,a_n)-y_i)^2¿,$$ затем он

минимизируется $S(a_0,a_1,\ldots,a_n)\to\min$.

Для решения поставленной задачи составляется система уравнений:

$$\begin{cases} \frac{\partial S}{\partial a_0}=0 \\ \frac{\partial S}{\partial a_1}=0 \\ \cdots\cdots \\ \frac{\partial S}{\partial a_n}=0 \end{cases}$$, что эквивалентно $$\begin{cases} 2\sum_{i=1}^{m}(Q(x_{1i},x_{2i},x_{3i},a_0,a_1,\cdots a_n)-y_i)\frac{\partial Q}{\partial a_0}=0 \\ 2\sum_{i=1}^{m}(Q(x_{1i},x_{2i},x_{3i},a_0,a_1,\cdots a_n)-y_i)\frac{\partial Q}{\partial a_1}=0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots(*) \\ 2\sum_{i=1}^{m}(Q(x_{1i},x_{2i},x_{3i},a_0,a_1,\cdots a_n)-y_i)\frac{\partial Q}{\partial a_n}=0 \end{cases}$$

Для $S(a_0,a_1,\ldots a_n)=\sum_{i=1}^{m}(Q(x_{1i},x_{2i},x_{3i},a_0,a_1,\ldots a_n)-y_i)^2$ $S(a_0,a_1,\ldots a_n)=\sum_{i=1}^{m}(a_0+a_1x_{1i}+a_2x_{2i}+a_3x_{3i}-y_i)^2$

система уравнений (¿) имеет вид:

$$\begin{cases} \frac{\partial S}{\partial a_0}=2\sum_{i=1}^{m}(a_0+a_1x_{1i}+a_2x_{2i}+a_3x_{3i}-y_i)=0 \\ \frac{\partial S}{\partial a_1}=2\sum_{i=1}^{m}(a_0+a_1x_{1i}+a_2x_2+a_3x_{3i}-y_i)\times x_{1i}=0 \\ \frac{\partial S}{\partial a_2}=2\sum_{i=1}^{m}(a_0+a_1x_{1i}+a_2x_{2i}+a_3x_{3i}-y_i)\times x_{2i}=0 \\ \frac{\partial S}{\partial a_3}=2\sum_{i=1}^{m}(a_0+a_1x_{1i}+a_2x_{2i}+a_3x_{3i}-y_i)\times x_{3i}=0 \end{cases}$$

После выполнения несложных эквивалентных преобразований получаем систему уравнений:

$$\begin{cases} ma_0+a_1\sum_{i=1}^{m}x_{1i}+a_2\sum_{i=1}^{m}x_{2i}+a_3\sum_{i=1}^{m}x_{3i}=\sum_{i=1}^{m}y_i \\ a_0\sum_{i=1}^{m}x_{1i}+a_1\sum_{i=1}^{m}x_{1i}^2+a_2\sum_{i=1}^{m}x_{1i}x_{2i}+a_3\sum_{i=1}^{m}x_{1i}x_{3i}=\sum_{i=1}^{m}x_{1i}y_i \\ a_0\sum_{i=1}^{m}x_{2i}+a_1\sum_{i=1}^{m}x_{1i}x_{2i}+a_2\sum_{i=1}^{m}x_{2i}^2+a_3\sum_{i=1}^{m}x_{2i}x_{3i}=\sum_{i=1}^{m}x_{2i}y_i \\ a_0\sum_{i=1}^{m}x_{3i}+a_1\sum_{i=1}^{m}x_{1i}x_{3i}+a_2\sum_{i=1}^{m}x_{2i}x_{3i}+a_3\sum_{i=1}^{m}x_{3i}^2=\sum_{i=1}^{m}x_{3i}y_i \end{cases}$$

Решением данной системы являются коэффициенты уравнения регрессии a_0, a_1, a_2, a_3. Их значения можно получить, например, методом Крамера.

Осуществление второго этапа сильно зависит от выводов, которые получены при анализе корреляционной матрицы. Можно значительно ускорить

проведение регрессионного анализа и снизить затраты на исследование, если принять правильную стратегию поиска наилучшего уравнения. Для этого необходимо знать основные и наиболее эффективные методы поиска наилучшего уравнения.

После получения каждого варианта уравнения обязательной процедурой является оценка его статистической значимости, поскольку главная цель - получить уравнение наивысшей значимости, поэтому второй этап корреляционно-регрессионного анализа неразрывно связан с третьим. Однако в связи с тем, что расчеты выполняет ЭВМ, а решение на основе оценки значимости уравнения принимает исследователь (принять или отбросить уравнение), условно можно выделить третий этап этой человеко-машинной технологии как интеллектуальный немашинный этап, для которого почти все данные по оценке значимости уравнения подготавливает ЭВМ.

На третьем этапе выясняют статистическую значимость, т. е. пригодность постулируемой модели для использования ее в целях предсказания значений отклика. Задачей третьего этапа построения регрессионных моделей является вычисление коэффициента детерминации или коэффициента множественной корреляции, на основании которого можно сделать заключение о значимости построенной регрессионной модели и возможности её дальнейшего использования для объяснения и анализа процессов, проявляющихся, в данной задаче. Коэффициент множественной корреляции определяется по формуле: $R_{xy} = \sqrt{\left(K_x^{-1}\overrightarrow{r_{yx}}, \overrightarrow{r_{yx}}\right)}$, где K_x^{-1} – обратная матрица корреляции, а матрица корреляции для трёх факторных признаков x_1, x_2 имеет вид: $K_x = \begin{pmatrix} r_{x1,x1} & r_{x1,x2} & r_{x1,x3} \\ r_{x2,x1} & r_{x2,x2} & r_{x2,x3} \\ r_{x3,x1} & r_{x3,x2} & r_{x3,x3} \end{pmatrix}$. Причём корреляционная матрица любой размерности будет симметричной, так как $r_{xixj} = r_{xjxi}$. Для трёх факторных признаков вектор $\overrightarrow{r_{yx}} = \left(r_{yx1}, r_{yx2}, r_{yx3}\right)$ - вектор коэффициентов парной регрессии результативного признака Y и факторных признаков X_1, X_2, X_3.

На этом этапе исключительно важную роль играют коэффициент детерминации и F-критерий значимости регрессии. R_{xy}^2 = D = R_{xy} Squared - коэффициент детерминации - это квадрат множественного коэффициента корреляции между наблюдаемым значением Y и его теоретическим значением, вычисленным на основе модели с определенным набором факторов. Коэффициент детерминации измеряет действительность модели. Он может принимать значения от 0 до 1. Эта величина особенно полезна для сравнения ряда различных моделей и выбора наилучшей модели.

На четвертом этапе корреляционно-регрессионного исследования, если полученная модель статистически значима, ее применяют для прогнозирования (предсказания), управления или объяснения.

Влияние отдельных факторов в многофакторных моделях может быть охарактеризовано с помощью частных коэффициентов эластичности, которые в случае линейной трёхфакторной модели рассчитываются по формулам:

$Э_{yx1} = \frac{a_1 \overline{x_1}}{\overline{y}}$; $Э_{yx2} = \frac{a_2 \overline{x_2}}{\overline{y}}$; $Э_{yx3} = \frac{a_3 \overline{x_3}}{\overline{y}}$. Частные коэффициенты эластичности показывают на сколько процентов изменится результативный признак, если значение одного из факторных признаков изменится на 1 %, а значения других признаков останутся неизменными.

Если же обнаружена незначимость, то модель отвергают, предполагая, что истинной окажется какая-то другая форма связи, которую надо поискать. Например, с самого начала работы (как бы по умолчанию) строилась и проверялась линейная регрессионная модель. Незначимость ее служит основанием для того, чтобы отвергнуть только линейную форму модели. Возможно, что более подходящей будет нелинейная форма модели.

Для выполнения всех перечисленных выше этапов корреляционно - регрессионного анализа данных можно средствами MS Excel построить имитационную модель. Алгоритм разработки имитационной модели можно представить следующим образом.

1) Оформление главного листа приложения. Размещения на нём названия приложения и кнопок для вызова необходимых модулей. На главном листе

содержатся следующие кнопки «Вспомогательные расчёты», «Расчёт коэффициентов корреляции», «Расчёт коэффициентов модели», «Оценка значимости уравнения», «Построение графика».

2) Разработка макроса «Oforml» для оформления базовой таблицы данных для проведения корреляционно – регрессионного анализа. Опция «Сервис», закладка «Макрос», закладка «Начать запись». Осуществляется оформление границ таблицы, ввод наименований столбцов (зависимой и независимых переменных), наименований строк таблицы. Здесь же производятся вспомогательные расчёты, связанные с перемножением столбцов исходных данных, нахождением их сумм. Выполнение этих расчётов требуется для нахождения значений коэффициентов будущей системы нормальных уравнений. После завершения всех указанных действий нажимается кнопка «Остановить запись». Макрос считается созданным, его можно вызывать по нажатию кнопки «Вспомогательные расчёты» (см. рис. 1).

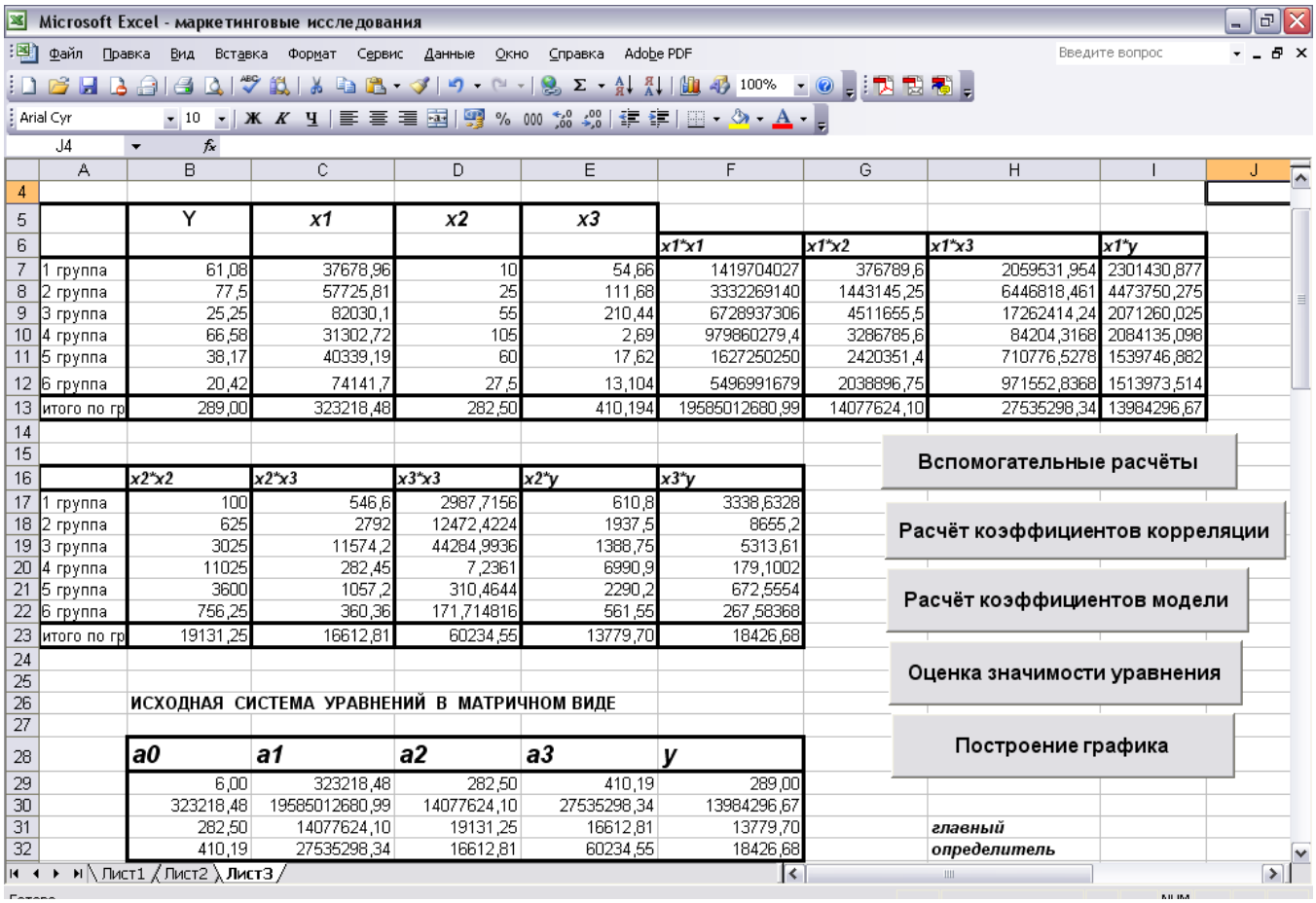


Рис. 1. Главный лист приложения на VBA for Excel для проведения корреляционно - регрессионного анализа данных

3) Запись макроса “Rasch_kor” для расчёта коэффициентов корреляции. Опция

«Сервис», закладка «Макрос», закладка «Начать запись». С помощью функции КОРЕЛЛ() табличного процессора MS Excel'2000 по формулам вида: $r_{yx_1} = \frac{\frac{1}{m}\sum_{j=1}^{m}((y_j - \bar{y})(x_{1j} - \bar{x}_1))}{\sigma_y \sigma_{x1}}$ (вместо y и x_1 могут быть использованы переменные $x_i \neq x_j$) рассчитываются парные коэффициенты факторных признаков и результирующего и каждого из факторных признаков. После окончания ввода формул для расчёта парных коэффициентов корреляции нажимается кнопка «Остановить запись». Макрос считается записанным и назначается кнопке «Расчёт коэффициентов корреляции», размещённой на главном листе приложения.

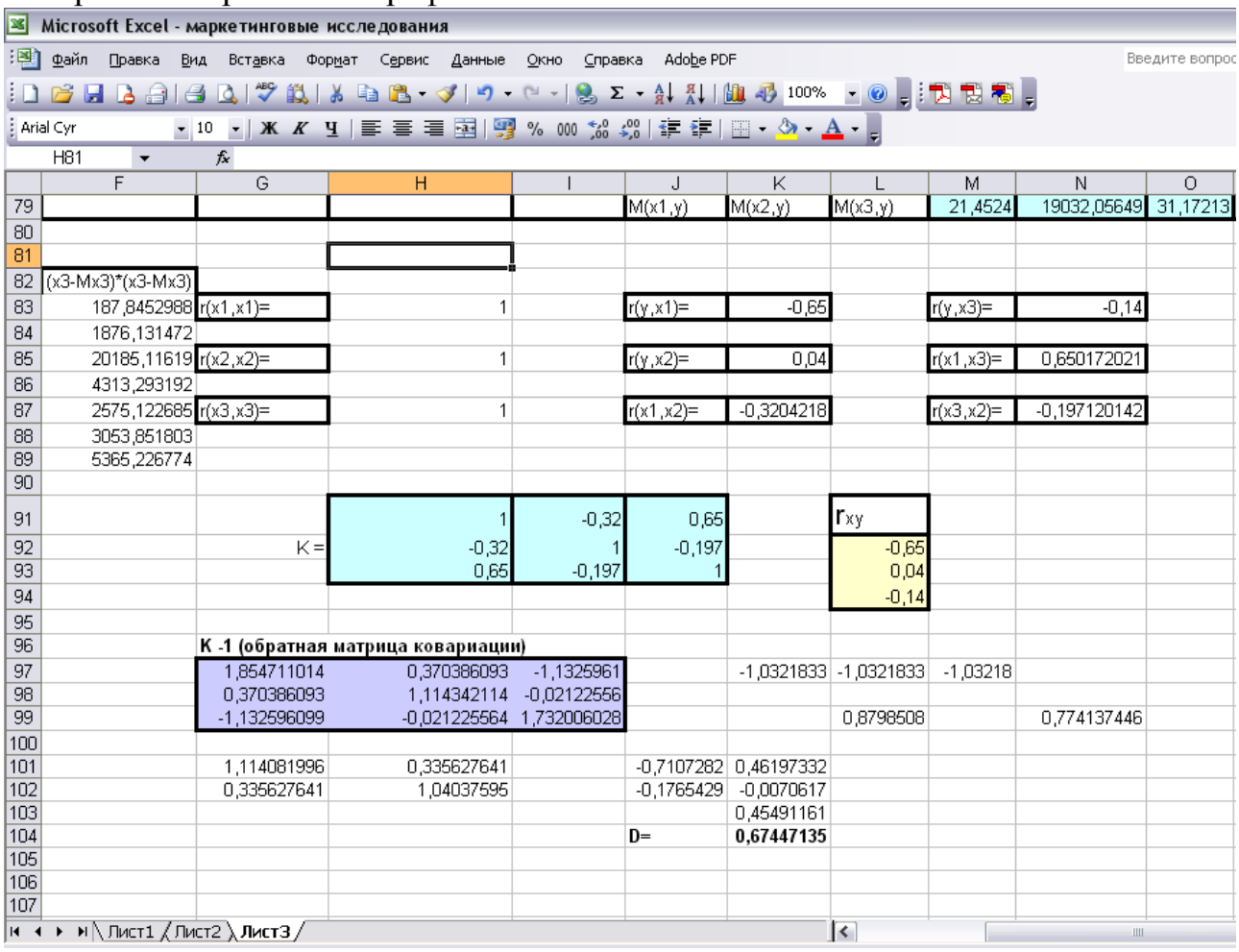
4) Запись макроса «Rasch_koef», который осуществляет расчёт коэффициентов регрессионного уравнения линейного вида. Запись макроса производится из опции «Сервис», закладки «Макрос», закладки «Начать запись». Далее с применением стандартной функции MS Excel'2000 МОПРЕД() осуществляется расчёт главного и остальных определителей, соответствующих системе линейных уравнений, полученной на основе метода наименьших квадратов, из которой будут определяться коэффициенты уравнения регрессии с помощью метода Крамера. После выполнения расчётов определителей системы линейных уравнений выполняется расчёт коэффициентов регрессионного уравнения по формуле: $a_i = \frac{\Delta_i}{\Delta}$, *где* Δ_i - определитель, соответствующий i – й переменной, Δ - главный определитель системы. По заверешении всех указанных действий необходимо остановить запись макроса нажатием кнопки «Остановить запись». Созданный макрос назначается кнопке «Расчёт коэффициентов модели», размещённой на главном листе приложения.

5) Запись макроса «Koef_Det», который позволяет вычислить множественный коэффициент корреляции и коэффициент детерминации по формулам: $R_{xy} = \sqrt{(K_x^{-1} r_{yx}, r_{yx})}$ - коэффициент множественной корреляции, где K_x^{-1} – обратная матрица корреляции; $R_{xy}^2 = D$ – коэффициент детерминации. Обратная матрица корреляции вычисляется с помощью функции MS Excel'2010 МОБР(). Запись макроса осуществляется из опции «Сервис», закладки «Макрос», закладки «Начать запись». После ввода расчётных формул для коэффициентов множественной корреляции и детерминации необходимо остановить запись макроса нажатием кнопки «Остановить запись». Созданный макрос назначается кнопке «Оценка значимости уравнения», размещённой на главном листе

приложения.

6) Запись макроса «Grafik», позволяющего построить график полученного уравнения и исходные значения результирующего в одной системе координат. Запись макроса осуществляется из опции «Сервис», закладки «Макрос», закладки «Начать запись». После этого из опции «Вставка», закладки «Диаграмма» осуществляется выбор типа графика, задание диапазона исходных данных для его построения и непосредственно его построение. По завершении указанных действий остановить запись макроса нажатием кнопки «Остановить запись». Созданный макрос назначается кнопке «Построение графика», размещённой на главном листе приложения.

7) Проверка работоспособности имитационной модели путём нажатия всех кнопок и проверки правильности выполненных модулями VBA расчётов (учитывая, что правильно рассчитанные коэффициенты парной и множественной корреляции должны находиться в диапазоне от –1 до 1, а коэффициент детерминации – в диапазоне от 0 до 1 (см. рис. 2)), визуальный контроль построенного графика.



	F	G	H	I	J	K	L	M	N	O
79					M(x1,y)	M(x2,y)	M(x3,y)	21,4524	19032,05649	31,17213
80										
81										
82	(x3-Mx3)*(x3-Mx3)									
83	187,8452988	r(x1,x1)=	1		r(y,x1)=	-0,65		r(y,x3)=	-0,14	
84	1876,131472									
85	20185,11619	r(x2,x2)=	1		r(y,x2)=	0,04		r(x1,x3)=	0,650172021	
86	4313,293192									
87	2575,122685	r(x3,x3)=	1		r(x1,x2)=	-0,3204218		r(x3,x2)=	-0,197120142	
88	3053,851803									
89	5365,226774									
90										
91			1	-0,32	0,65		rxy			
92		K =	-0,32	1	-0,197		-0,65			
93			0,65	-0,197	1		0,04			
94							-0,14			
95										
96		K -1 (обратная матрица ковариации)								
97		1,854711014	0,370386093	-1,1325961		-1,0321833	-1,0321833	-1,03218		
98		0,370386093	1,114342114	-0,02122556						
99		-1,132596099	-0,021225564	1,732006028			0,8798508		0,774137446	
100										
101		1,114081996	0,335627641		-0,7107282	0,46197332				
102		0,335627641	1,04037595		-0,1765429	-0,0070617				
103						0,45491161				
104					D=	**0,67447135**				
105										
106										
107										

Рис. 2. Результаты расчётов коэффициентов парной корреляции и детерминации