РОССИЙСКИЙ УНИВЕРСИТЕТ ДРУЖБЫ НАРОДОВ

Факультет физико-математических и естественных наук

Кафедра прикладной информатики и теории вероятностей

Практикум по математической статистике

Лабораторная работа №1

Тема: «Множественный регрессионный анализ»

Вариант 10

Выполнил

Студент: Феоктистов Владислав

Группа: НПМбд-01-19б

№ с/б: 1032192939

Преподаватель: Матюшенко Сергей Иванович

МОСКВА

2022 г.

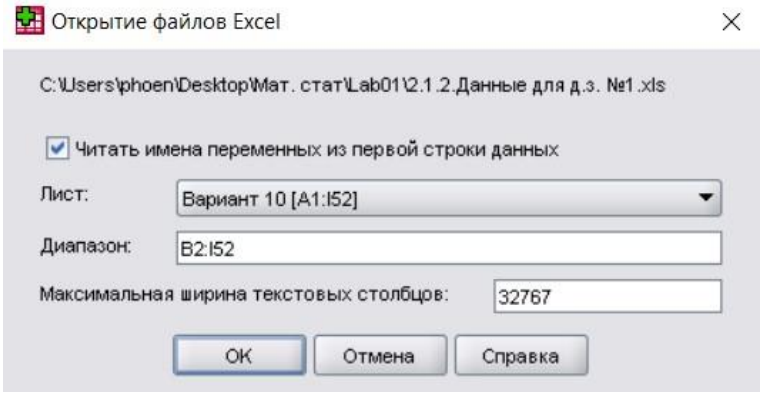
Цель работы: приобрести практические навыки применения множественного регрессионного анализа для решения конкретных задач с использованием статистического пакета SPSS.

Ход работы:

1. Повторил теоретические основы множественного регрессионного анализа.
2. Разобрал пример использования SPSS для построения модели множественной линейной регрессии.
3. Импортировал данные из файла в формате Excel в файл SPSS.

A2

	A	B	C	D	E	F	G	H	I
1	№ п/п	Цена	Общ. площ.	Площ. кухни	Жил. площ.	Район	Этаж	Тип дома	Колич. комнат
2		Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
3	1	520	33	6	19,5	1	4	1	1
4	2	435	28,7	6,1	15,8	1	2	1	1
5	3	800	52	7	32	1	7	1	2
6	4	825	47,4	7,2	28,4	1	7	1	2
7	5	880	53,2	7,6	34	1	9	0	2
8	6	721	51,7	7,3	35,8	1	4	1	2
9	7	630	47,6	5,9	29,7	1	5	1	2
10	8	900	53,1	7	30,4	1	5	0	2
11	9	1030	56,7	11	33	1	2	1	2
12	10	619	43,4	5,8	23,6	1	4	1	2
13	11	670	35	6	23	1	3	1	2
14	12	725	55	6	41	1	1	1	3
15	13	1450	70,3	13,2	39	1	2	0	3
16	14	1130	66,4	7,6	43,1	1	7	0	3
17	15	1030	63	9	37	1	7	1	3
18	16	1070	63,7	9,7	36	1	8	1	3
19	17	790	69	7,5	44	1	2	1	3



	Y	X1	X2	X3	X4	X5	X6	X7
1	520,0	33	6	19,5	1	4	1	1
2	435	29	6	15,8	1	2	1	1
3	800	52	7	32,0	1	7	1	2
4	825	47	7	28,4	1	7	1	2
5	880	53	8	34,0	1	9	0	2
6	721	52	7	35,8	1	4	1	2
7	630	48	6	29,7	1	5	1	2
8	900	53	7	30,4	1	5	0	2
9	1030	57	11	33,0	1	2	1	2
10	619	43	6	23,6	1	4	1	2
11	670	35	6	23,0	1	3	1	2
12	725	55	6	41,0	1	1	1	3
13	1450	70	13	39,0	1	2	0	3
14	1130	66	8	43,1	1	7	0	3
15	1030	63	9	37,0	1	7	1	3
16	1070	64	10	36,0	1	8	1	3
17	790	69	8	44,0	1	2	1	3
18	830	67	8	43,6	1	1	0	3
19	510	37	8	17,8	2	4	0	1

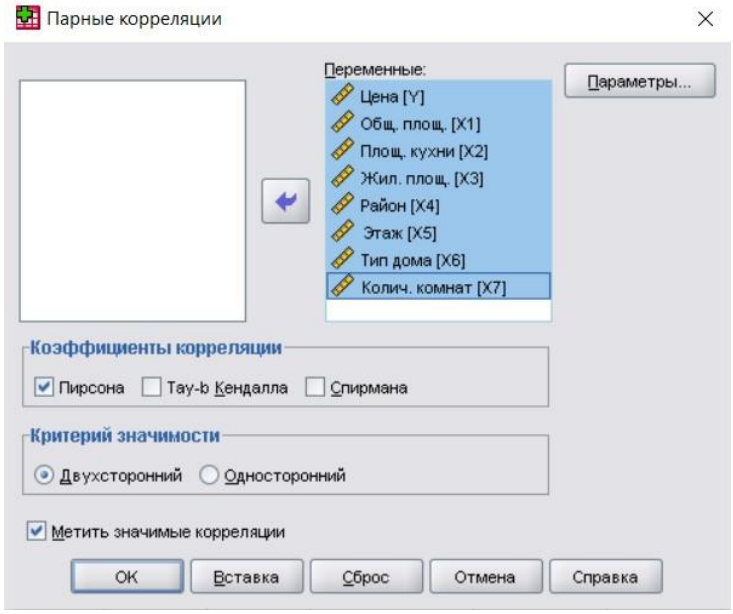
	Имя	Тип	Ширина	Десятич...	Метка	Значения	Пропуски	Столбцы	Выравнивание	Шкала
1	Y	Числовая	11	0		Нет	Нет	11	По право...	Количестве...
2	X1	Числовая	11	0		Нет	Нет	11	По право...	Количестве...
3	X2	Числовая	11	0		Нет	Нет	11	По право...	Количестве...
4	X3	Числовая	11	1		Нет	Нет	11	По право...	Количестве...
5	X4	Числовая	11	0		Нет	Нет	11	По право...	Количестве...
6	X5	Числовая	11	0		Нет	Нет	11	По право...	Количестве...
7	X6	Числовая	11	0		Нет	Нет	11	По право...	Количестве...
8	X7	Числовая	11	0		Нет	Нет	11	По право...	Количестве...

Важно отметить, что перед импортом Excel файла необходимо в начале просмотреть содержимое файла, определить лист и диапазон выборки. Поскольку у меня 10 вариант и в исходных данных первый столбец отвечает за индексы, а первая строка за метки (более подробное описание параметров Y, X1, X2, …, X7), то я выбрал лист с названием «Вариант 10» и установил диапазон выборки B2:I52.

Для удобства дальнейшего работы, параметрам Y, X1, X2, …, X7 можно задать метки.

	Имя	Тип	Ширина	Десятич...	Метка	Значения	Пропуски	Столбцы	Выравнивание	Шкала
1	Y	Числовая	11	0	Цена	Нет	Нет	11	По право...	Количестве...
2	X1	Числовая	11	0	Общ. площ.	Нет	Нет	11	По право...	Количестве...
3	X2	Числовая	11	0	Площ. кухни	Нет	Нет	11	По право...	Количестве...
4	X3	Числовая	11	1	Жил. площ.	Нет	Нет	11	По право...	Количестве...
5	X4	Числовая	11	0	Район	Нет	Нет	11	По право...	Количестве...
6	X5	Числовая	11	0	Этаж	Нет	Нет	11	По право...	Количестве...
7	X6	Числовая	11	0	Тип дома	Нет	Нет	11	По право...	Количестве...
8	X7	Числовая	11	0	Колич. комнат	Нет	Нет	11	По право...	Количестве...

Далее я построил матрицу парной корреляции всех переменных.



	Имя	Тип	Ширина	Десятич...	Метка	Значения	Пропуски	Столбцы	Выравнивание	Шкала
1	Y	Числовая	11	0	Цена	Нет	Нет	11	По право...	Количестве...
2	X1	Числовая	11	0	Общ. площ.	Нет	Нет	11	По право...	Количестве...
3	X2	Числовая	11	0	Площ. кухни	Нет	Нет	11	По право...	Количестве...
4	X3	Числовая	11	1	Жил. площ.	Нет	Нет	11	По право...	Количестве...
5	X4	Числовая	11	0	Район	Нет	Нет	11	По право...	Номинальная
6	X5	Числовая	11	0	Этаж	Нет	Нет	11	По право...	Номинальная
7	X6	Числовая	11	0	Тип дома	Нет	Нет	11	По право...	Номинальная
8	X7	Числовая	11	0	Колич. комнат	Нет	Нет	11	По право...	Номинальная

Корреляции

		Цена	Общ. площ.	Площ. кухни	Жил. площ.	Район	Этаж	Тип дома	Колич. комнат
Цена	Корреляция Пирсона	1	,840**	,589**	,670**	-,346*	,235	-,136	,723**
	Знч.(2-сторон)		,000	,000	,000	,014	,101	,346	,000
	N	50	50	50	50	50	50	50	50
Общ. площ.	Корреляция Пирсона	,840**	1	,444**	,858**	-,244	,168	-,077	,868**
	Знч.(2-сторон)	,000		,001	,000	,088	,243	,593	,000
	N	50	50	50	50	50	50	50	50
Площ. кухни	Корреляция Пирсона	,589**	,444**	1	,233	-,187	,098	-,047	,273
	Знч.(2-сторон)	,000	,001		,103	,195	,500	,745	,055
	N	50	50	50	50	50	50	50	50
Жил. площ.	Корреляция Пирсона	,670**	,858**	,233	1	-,162	,085	-,067	,928**
	Знч.(2-сторон)	,000	,000	,103		,260	,557	,643	,000
	N	50	50	50	50	50	50	50	50
Район	Корреляция Пирсона	-,346*	-,244	-,187	-,162	1	-,155	-,333*	-,215
	Знч.(2-сторон)	,014	,088	,195	,260		,283	,018	,134
	N	50	50	50	50	50	50	50	50
Этаж	Корреляция Пирсона	,235	,168	,098	,085	-,155	1	-,057	,056
	Знч.(2-сторон)	,101	,243	,500	,557	,283		,693	,700
	N	50	50	50	50	50	50	50	50
Тип дома	Корреляция Пирсона	-,136	-,077	-,047	-,067	-,333*	-,057	1	,026
	Знч.(2-сторон)	,346	,593	,745	,643	,018	,693		,856
	N	50	50	50	50	50	50	50	50
Колич. комнат	Корреляция Пирсона	,723**	,868**	,273	,928**	-,215	,056	,026	1
	Знч.(2-сторон)	,000	,000	,055	,000	,134	,700	,856	
	N	50	50	50	50	50	50	50	50

**. Корреляция значима на уровне 0.01 (2-сторон.).

*. Корреляция значима на уровне 0.05 (2-сторон.).

Как можно заметить, SPSS сам установил для параметров X4, X5, X6, X7 шкалу «Номинальная», поскольку эти параметры носят категориальный характер.

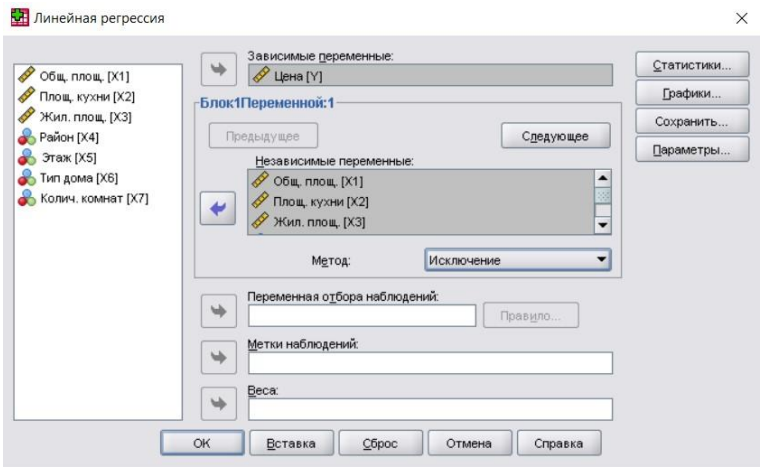
Из таблицы корреляции видно, что зависимая переменная «Цена» сильнее всего связана с независимыми переменными «Общ. площ.» ($r_{Y,X1} = 0{,}840$), «Колич. комнат» ($r_{Y,X7} = 0{,}723$), «Жил. площ» ($r_{Y,X3} = 0{,}670$) и «Площ. кухни» ($r_{Y,X2} = 0{,}589$). Однако, в то же время переменная «Общ. площ.» сильно коррелирует с переменными «Жил. площ» ($r_{Y,X3} = 0{,}858$), «Колич. комнат» ($r_{Y,X7} = 0{,}868$), что

свидетельствует о наличии мультиколлинеарности (это очевидно, поскольку общая площадь напрямую зависит от жилой площади и количества комнат и их площадь, как правило больше, чем площадь кухни, а значит и их корреляция выше). Поэтому, т.к. зависимая переменная «Цена» сильнее зависит от независимой переменной «Общ. площ.», то переменные «Жил. площ» и «Колич. комнат» нужно будет исключить.

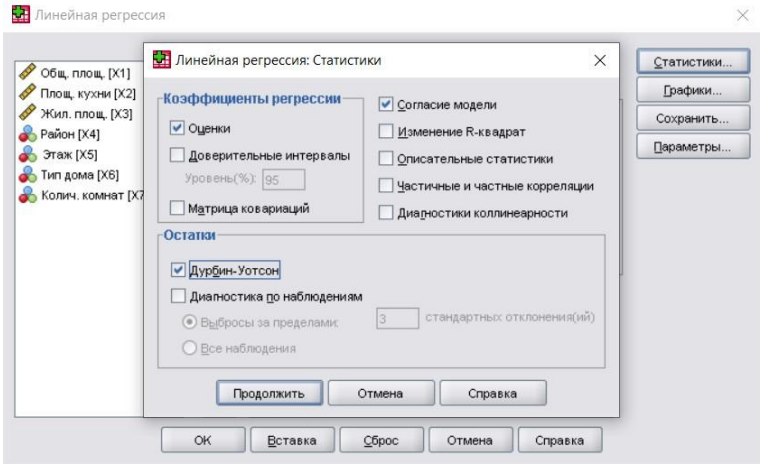
Стоит заметить, что независимая переменная «Этаж» не имеет значимой корреляции на уровне 0,01 и 0,05 ни для одной из других переменных, в том числе и зависимой. Следовательно, эту переменную тоже можно исключить.

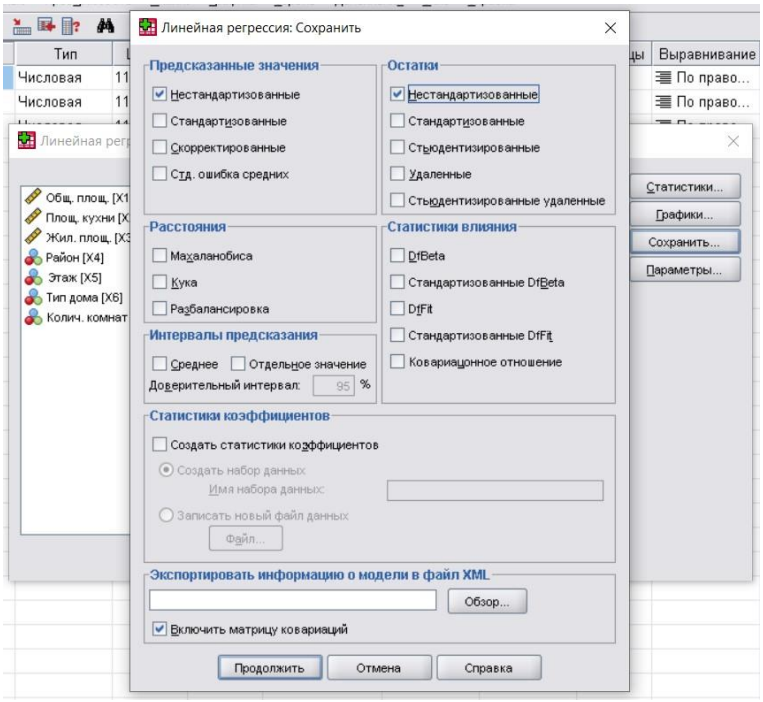
В итоге, воспользовавшись методом исключения, приходим к выводу, что целесообразно построить многофакторное регрессионное уравнение от переменных: «Общ. площ», «Площ. кухни», «Район» и «Тип дома».

Построим линейное уравнение регрессии, используя метод исключения. В качестве зависимой переменной выбираем «Цена», а в качестве независимых – все остальное.



В поле панели «Статистики…» добавляем галочку перед «Дурбин-Уотсон», а в поле панель «Сохранить…» - перед «Нестандартизованные» в разделах «Предсказательные значения» и «Остатки».





В итоге получаем следующие результаты:

Введенные или удаленные переменные[b]

Модель	Включенные переменные	Исключенные переменные	Метод
1	Колич. комнат, Тип дома, Этаж, Площ. кухни, Район, Общ. площ., Жил. площ.[a]	.	Принудительное включение
2	.	Этаж	Исключение (критерий: вероятность F-исключения >= ,100).
3	.	Жил. площ.	Исключение (критерий: вероятность F-исключения >= ,100).
4	.	Колич. комнат	Исключение (критерий: вероятность F-исключения >= ,100).

a. Включены все запрошенные переменные

b. Зависимая переменная: Цена

Из таблицы «Введенные и удаленные переменные» видно, что были исключены переменные «Этаж», «Жил. Площ.» и «Колич. комнат». Это сходится с моими исключениями при анализе таблицы корреляции.

Сводка для модели[e]

Модель	H	R-квадрат	Скорректированный R-квадрат	Стд. ошибка оценки	Дурбин-Уотсон
1	,902[a]	,813	,782	103,487	
2	,898[b]	,807	,780	103,970	
3	,894[c]	,799	,777	104,830	
4	,891[d]	,794	,776	104,979	2,032

a. Предикторы: (конст) Колич. комнат, Тип дома, Этаж, Площ. кухни, Район, Общ. площ., Жил. площ.

b. Предикторы: (конст) Колич. комнат, Тип дома, Площ. кухни, Район, Общ. площ., Жил. площ.

c. Предикторы: (конст) Колич. комнат, Тип дома, Площ. кухни, Район, Общ. площ.

d. Предикторы: (конст) Тип дома, Площ. кухни, Район, Общ. площ.

e. Зависимая переменная: Цена

С помощью коэффициентов детерминации R^2 и множественной корреляции R в таблице «Свода для модели» оценим качество модели.

Коэффициент детерминации $R^2 = 0{,}794$ показывает, что около 79,4% вариаций зависимых переменных учтено в модели и обусловлено влиянием включенных факторов, что достаточно хорошо.

Коэффициент множественной корреляции $R = 0{,}891$ показывает, что зависимая переменная «Цена» достаточно тесно связана со включенными в модель факторами.

Несмотря на то, что коэффициент детерминации в первой модели (до исключения коррелирующих и незначимых переменных) незначительно больше, чем в четвертой, принимаем за верную мы последнюю, поскольку эта разница незначительна, а мультиколлинеарность нежелательна и от нее нужно избавляться.

Дисперсионный анализ[e]

Модель		Сумма квадратов	ст.св.	Средний квадрат	Щ	Знч.
1	Регрессия	1961466,029	7	280209,433	26,164	,000[a]
	Остаток	449801,251	42	10709,554		
	Всего	2411267,280	49			
2	Регрессия	1946449,368	6	324408,228	30,011	,000[b]
	Остаток	464817,912	43	10809,719		
	Всего	2411267,280	49			
3	Регрессия	1927735,226	5	385547,045	35,084	,000[c]
	Остаток	483532,054	44	10989,365		
	Всего	2411267,280	49			
4	Регрессия	1915342,056	4	478835,514	43,449	,000[d]
	Остаток	495925,224	45	11020,561		
	Всего	2411267,280	49			

a. Предикторы: (конст) Колич. комнат, Тип дома, Этаж, Площ. кухни, Район, Общ. площ., Жил. площ.

b. Предикторы: (конст) Колич. комнат, Тип дома, Площ. кухни, Район, Общ. площ., Жил. площ.

c. Предикторы: (конст) Колич. комнат, Тип дома, Площ. кухни, Район, Общ. площ.

d. Предикторы: (конст) Тип дома, Площ. кухни, Район, Общ. площ.

e. Зависимая переменная: Цена

На основании F-критерия Фишера в дисперсионном анализе сделаем проверку значимости уравнения регрессии.

Значение критерия Фишера $F = 43{,}449$. Расчетное значение F-критерия меньше 0,001 (Округленное до тысячных значение равно 0,000. Это не говорит о том, что значение равно 0, а только о том, что оно очень маленькое), следовательно, модель значима.

Коэффициенты[a]

Модель		Нестандартизованные коэффициенты		Стандартизованные коэффициенты	t	Знч.
		B	Стд. Ошибка	Бета		
1	(Константа)	113,365	115,666		,980	,333
	Общ. площ.	9,213	2,684	,564	3,433	,001
	Площ. кухни	34,366	10,748	,252	3,197	,003
	Жил. площ.	-5,610	4,152	-,264	-1,351	,184
	Район	-74,831	34,363	-,164	-2,178	,035
	Этаж	7,392	6,242	,082	1,184	,243
	Тип дома	-69,345	32,665	-,158	-2,123	,040
	Колич. комнат	108,252	58,638	,374	1,846	,072
2	(Константа)	143,209	113,413		1,263	,213
	Общ. площ.	9,801	2,650	,600	3,698	,001
	Площ. кухни	34,011	10,794	,249	3,151	,003
	Жил. площ.	-5,487	4,170	-,258	-1,316	,195
	Район	-80,516	34,185	-,176	-2,355	,023
	Тип дома	-71,503	32,766	-,163	-2,182	,035
	Колич. комнат	98,470	58,324	,340	1,688	,099
3	(Константа)	121,963	113,187		1,078	,287
	Общ. площ.	8,708	2,537	,533	3,432	,001
	Площ. кухни	36,823	10,668	,270	3,452	,001
	Район	-83,168	34,408	-,182	-2,417	,020
	Тип дома	-64,422	32,589	-,147	-1,977	,054
	Колич. комнат	43,944	41,380	,152	1,062	,294
4	(Константа)	113,149	113,042		1,001	,322
	Общ. площ.	11,044	1,266	,676	8,725	,000
	Площ. кухни	34,001	10,346	,249	3,286	,002
	Район	-81,617	34,425	-,178	-2,371	,022
	Тип дома	-57,737	32,020	-,131	-1,803	,078

a. Зависимая переменная: Цена

Исключенные переменные[d]

Модель		Бета включения	t	Знч.	Частная корреляция	Статистики коллинеарности: Толерантность
2	Этаж	,082[a]	1,184	,243	,180	,920
3	Этаж	,080[b]	1,140	,261	,171	,920
	Жил. площ.	-,258[b]	-1,316	,195	-,197	,117
4	Этаж	,065[c]	,931	,357	,139	,949
	Жил. площ.	-,023[c]	-,162	,872	-,024	,236
	Колич. комнат	,152[c]	1,062	,294	,158	,223

a. Предикторы в модели: (конст) Колич. комнат, Тип дома, Площ. кухни, Район, Общ. площ., Жил. площ.

b. Предикторы в модели: (конст) Колич. комнат, Тип дома, Площ. кухни, Район, Общ. площ.

c. Предикторы в модели: (конст) Тип дома, Площ. кухни, Район, Общ. площ.

d. Зависимая переменная: Цена

Статистики остатков[a]

	Минимум	Максимум	Для среднего	Стд. Отклонение	N
Предсказанное значение	348,47	1256,76	735,12	197,708	50
Остаток	-232,599	239,453	,000	100,603	50
Стд. Предсказанное значение	-1,956	2,638	,000	1,000	50
Стд. Остаток	-2,216	2,281	,000	,958	50

a. Зависимая переменная: Цена

На основе данных в таблице «Коэффициенты» составил уравнение регрессии:

$$y = 11{,}044x_1 + 34{,}001x_2 - 81{,}617x_4 - 57{,}737x_6 + 113{,}149$$

При увеличении общей площади на 1 квадратный метр, цена увеличится на 11,044 у.е.; при увеличении площади кухни на 1 квадратный метр, цена увеличится на 34,001 у.е.; при увеличении номера района на 1, цена уменьшится на 81,617 у.е.; при увеличении номера типа дома на 1, цена уменьшается на 57,737 у.е.

Вывод: при выполнении лабораторной работы были приобретены практические навыки применения множественного регрессионного анализа для решения конкретных задач с использованием статистического пакета SPSS.