Национальный исследовательский университет «Высшая школа экономики»

Факультет: Московский институт электроники и математики

Образовательная программа: Прикладная математика

Отчет по Домашнему заданию No 2 ««Первичная обработка статистических данных»»

по майору

«Прикладной статистический анализ»

Работу выполнил студент 2 курса:

Цыплаков Александр Александрович

Преподаватель:

Кабаева Елена Владимировна

Содержание

Введение

Для анализа были взяты массивы исходных данных под названием «Средняя стоимость квадратного метра жилой площади по субъектам РФ» (Приложение 1), а также для проведения анализа двумерных данных был взят массив данных «Население по субъектам Российский Федерации в 2020 г.» (Приложение 1). Данные получены на конец декабря 2020 г. Данная тема напрямую пересекается со специальностью «Экономика». Целью исследования было проведение анализа и поиск зависимости между средней стоимостью квадратного метра жилой площадью и количеством населения в субъекте РФ. Передо мной как исследователем стояли следующие задачи: описать показатели, выбранные для исследования, проведение корреляционного, регрессионного анализов, также формирование заключений, выводов.

Источником данных, взятых для анализа, является Росстат - федеральный орган исполнительной власти, осуществляющий функции по формированию официальной статистической информации о социальных, экономических, демографических, экологических и других общественных процессах в Российской Федерации.

Независимой переменной (x) является «Население по субъектам Российский Федерации в 2020 г.». Материалы по теме «Население в субъектах Российской Федерации» отражают информацию о количестве населения в субъектах РФ. Исчисляются в миллионах человек. Результативным (зависимым) признаком (y) является «Средняя стоимость квадратного метра жилой площади по субъектам РФ». Информация была подобрана с учетом заданных требований. Таким образом, данные в сходном ряду являются по своему типу пространственными (не временными), моментными, сопоставимыми, с отсутствием разрывов в значениях, абсолютными (являются первичными). После проведения корректировки величина данных составила 30 измерений. С ними были проведены все основные расчеты, для чего использовалась программа MS Excel для Mac версия 16.52.

Корреляционный анализ

Для проведения корреляционного анализа первым делом было построено корреляционное облако (Рис. 1) на основании 30 наблюдений, нанесена линия тренда. Проведя визуальный анализ можно сделать вывод, что между рассматриваемыми наблюдениями присутствует сильная корреляция. Глядя на линию тренда можно предположить, что значение корреляции будет в диапазоне от 0 до 1, ближе к 1.

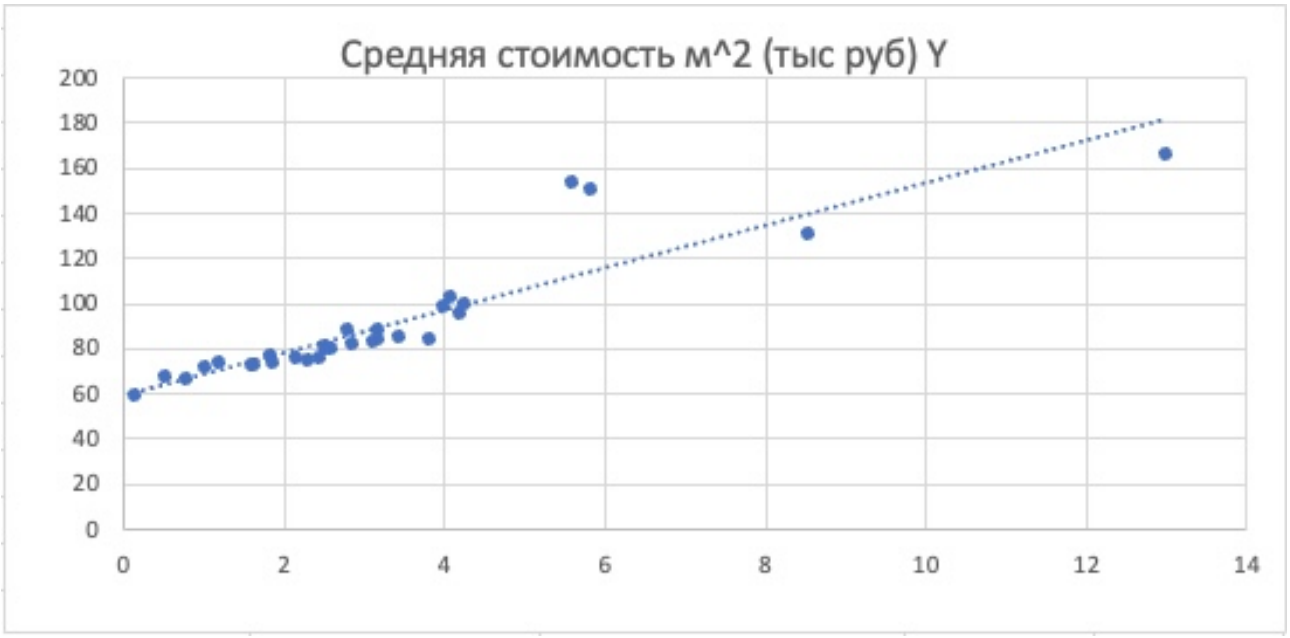

Рис.1

Далее необходимо рассчитать линейный коэффициент корреляции. Сделать это можно двумя способами: вручную и путем использования функции КОРРЕЛ. Для расчета вручную были добавлены 3 столбца: X*Y, X2, Y2 (Приложение 2). А также была добавлена строка «Сумма» для вычисления суммы всех параметров (X, Y, X*Y, X2, Y2). Коэффициент корреляции вычисляется по формуле(1). Таким образом, значение корреляции, вычисленное вручную, равно 0,901. Далее воспользуемся функцией КОРРЕЛ, выделяя необходимые диапазоны, значение корреляции равно 0,901. Вычисления вручную были выполнены верно. Таким образом, прогнозы, выдвинутые на этапе построения корреляционного облака оказались верными. Корреляция действительно есть, однако она не идеально положительная. Значения линейного коэффициента корреляции равно 0,90. Это означает, что «на дистанции» возрастание переменной x (количество активных компаний) приводит к возрастанию переменной y (население). Далее необходимо проверить гипотезы о наличии связи между показателями. Первым шагом является формулировка гипотез. Гипотеза H0: $r = 0$, связь между исследуемыми показателями отсутствует. Гипотеза H1: $r \neq 0$, связь между

показателями есть. Следующим шагом является проверка, ранее выдвинутых гипотез, то есть расчет и сравнение с табличным значением. Сначала рассчитаем t наблюдаемое по формуле . Значение t табл., вычисляется по таблице Стьюдента, t табл. = 2. Соответственно, t наблюдаемое (11,03) > t табличное (2), следовательно H0 отклоняется и принимается H1, значит, что связь между исследуемыми признаками есть.

Формула(1) r = (x (с р)y (с р) – x (с р) * y (с р))/S x * S y, где S x = x (с р)2 – (x (с р))2

Регрессионный анализ

Следующим этапом, для проведения статистического анализа двумерных количественных данных, является регрессионный анализ. Коэффициент регрессии отображает на сколько в среднем величина одной переменной изменяется при увеличении (уменьшении) значения другой переменной. Для начала необходимо построить линейную регрессионная модель по исходным данным. Можно произвести расчеты вручную или же воспользоваться пакетом для анализа данных. Для расчета вручную был создан отдельный лист «Регрессия. Вручную». К таблице исходных данных было создано несколько дополнительных расчетных столбцов: «X2», «XiYi», «$\tilde{Y}$ (расчетный)», «Отклонения Y - $\tilde{Y}$» (Приложение 2). $\tilde{Y}$ = b0 + b1*X - линейное уравнение регрессии. Коэффициент b1 рассчитывается по формуле(2) . Значение b1 вычисляется по формуле(3). Соответственно, линейное уравнение регрессии принимает следующий вид: Y^ = 59,9 + 9,3*X. Далее, после заполнения оставшихся столбцов, необходимо построить диаграмму количества компаний и населения по субъектам Российской Федерации в 2020 г. (Рис.2). Данные строятся на основании «X», «Y», «$\tilde{Y}$ (расчетный)» (Приложение 10). Так же была нанесена линия тренда.

Формула(2): b1=((xy)ср - xср*yср)/Sx^2

Формула(3): b0=yср - b1*x

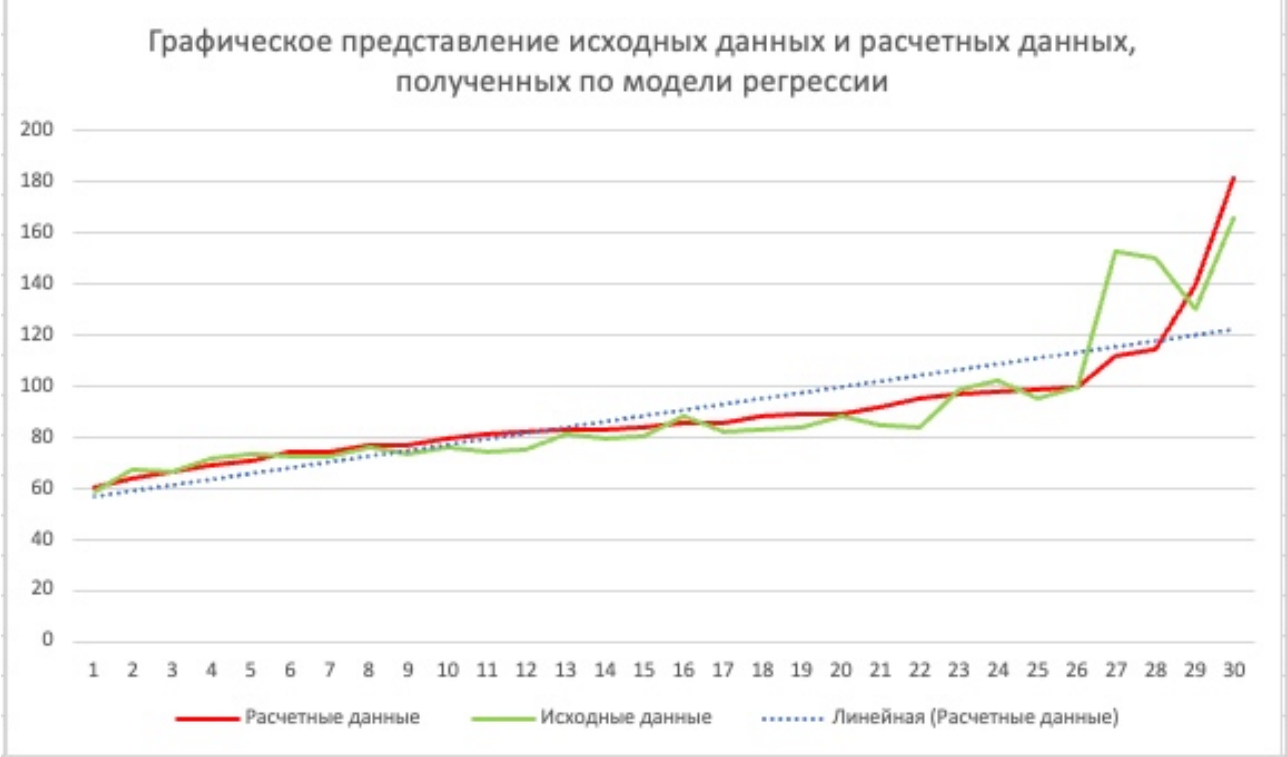

Рис.2

Стоит обратить внимание на значение R - квадрат, которое равно 0,81. Значение "R - квадрат" свидетельствует об устойчивости регрессионной модели. 81% R - квадрат имеет значение чуть выше среднего. Это означает небольшую устойчивость регрессионной модели. Модель проанализированных данных требует уточнения степень устойчивости не самая высокая.

При увеличении численности населения субъекта Российской Федерации на 1 миллион человек, средняя стоимость квадратного метра жилого помещения в среднем увеличивается на 9,3 тысяч рублей

Заключение

В процессе этой работы мы работали с собранными данными и пытались понять, есть ли между ними связь. Что вышло в итоге? Оказалось, что связь действительно есть, в первую очередь нам это показал коэффициент корреляции, который равен 0,90, что является "мигалкой" для аналитика, сигнализирующей о том, что между признаками может быть связь, но помимо коэффициента мы можем рассчитать еще много интересных вещей, которые до конца объяснят зависимость наших значений, поэтому мы идем дальше!

Мы выдвинули две гипотезы: связь есть и связи нет, после чего с помощью t-критерия Стьюдента пришли к тому, что принимаем первую гипотезу (связь есть). На этом закончился корреляционный анализ и мы приступили к регриссионному)

Первым делом мы составили уравнение регрессии: y^ = 59,9 + 9,3*x и затем привели ее интерпретацию! После этого было необходимо вычислить коэффициент детерминации, который показывает точность модели и благодаря нему выяснили, что наша модель дает достаточно точный результат, а в исходных данных 81% вариации зависимой переменной объясняется независимой. Крайними расчетами стали проверки гипотезы о значимости уравнения регрессии и коэффициента уравнения регрессии, мы точно так же выдвигали гипотезы о значимости и не значимости и, сравнивая с 1)Критерием Фишера 2)t-критерией Стьюдента, выяснили, что и уравнение и коэффициенты уравнения значимы!

А вишенкой на торте стал график сравнения полученных из модели данных и исходных!

Выводы:

1) Зависимость между средней стоимостью квадратного метра жилых помещений по субъектам РФ и численностью населения субъекта Российской Федерации есть (r = 0,90(значимый КК))

2) Модель линейной регрессии точная (R^2 = 0,81)

3) Уравнение регрессии значимо (Fрас>>Fтабл)

4) Коэффициент уравнения регресии значим (tрасч>>tтабл)

5) Все полученные значения согласуются друг с другом достаточно сильно!

Приложения

Приложение 1

Таблица с исходными данными

Субъект	Население (млн человек) X	Средняя стоимость м^2 (тыс руб) Y
Магаданская область	0,136	59,035
Республика Карелия	0,533	67,821
Республика Мордовия	0,783	66,345
Архангельская область	1,02	71,452
Иркутская область	1,197	73,2
Ставропольский край	1,604	72,567
Омская область	1,645	72,983
Приморский край	1,845	76,459
Оренбургская область	1,862	73,455
Алтайский край	2,164	75,975
Воронежская область	2,309	74,543
Саратовская область	2,443	75,685
Волгоградская область	2,501	81,136
Пермский край	2,532	79,694
Кемеровская область	2,601	80,191
Новосибирская область	2,797	88,582
Красноярский край	2,857	82,271
Нижегородская область	3,119	83,329
Самарская область	3,172	84,349
Республика Дагестан	3,182	88,596
Челябинская область	3,431	85,207
Тюменская область	3,824	84,123
Республика Татарстан	4,004	98,741
Республика Башкортостан	4,091	102,603
Ростовская область	4,201	95,034
Свердловская область	4,269	99,641
Санкт-Петербург	5,602	152,963
Краснодарский край	5,838	150,641
Московская область	8,525	130,65
Москва	13,01	166,044

Приложение 2

Таблица с расчетными значениями

X^2	Y^2	X*Y	Y^	Y-Y^	(Y-Y^)^2	Y^-Ycp	(Y^-Ycp)^2
0,018496	3485,131	8,02876	60,53826023	-1,50326	2,259791	-29,23857	854,8942
0,284089	4599,688	36,14859	64,28199961	3,539	12,52452	-25,49483	649,9865
0,613089	4401,659	51,94814	66,63951811	-0,294518	0,086741	-23,13732	535,3354
1,0404	5105,388	72,88104	68,87444564	2,577554	6,643786	-20,90239	436,9098
1,432809	5358,24	87,6204	70,54356874	2,656431	7,056627	-19,23326	369,9185
2,572816	5265,969	116,3975	74,38160886	-1,814609	3,292805	-15,39522	237,0129
2,706025	5326,518	120,057	74,76824189	-1,785242	3,187089	-15,00859	225,2578
3,404025	5845,979	141,0669	76,65425669	-0,195257	0,038125	-13,12258	172,202
3,467044	5395,637	136,7732	76,81456795	-3,359568	11,2867	-12,96227	168,0203
4,682896	5772,201	164,4099	79,6624503	-3,68745	13,59729	-10,11438	102,3007
5,331481	5556,659	172,1198	81,02981103	-6,486811	42,07872	-8,747022	76,5104
5,968249	5728,219	184,8985	82,29344094	-6,608441	43,67149	-7,483392	56,00116
6,255001	6583,05	202,9211	82,84038524	-1,704385	2,904929	-6,936448	48,11431
6,411024	6351,134	201,7852	83,13271753	-3,438718	11,82478	-6,644116	44,14427
6,765201	6430,596	208,5768	83,78339264	-3,592393	12,90528	-5,993441	35,92133
7,823209	7846,771	247,7639	85,63168714	2,950313	8,704346	-4,145146	17,18224
8,162449	6768,517	235,0482	86,19749158	-3,926492	15,41734	-3,579342	12,81169
9,728161	6943,722	259,9032	88,66817097	-5,339171	28,50675	-1,108662	1,229132
10,06158	7114,754	267,555	89,16796489	-4,818965	23,22242	-0,608868	0,370721
10,12512	7847,479	281,8807	89,26226563	-0,676266	0,457335	-0,514568	0,26478
11,77176	7260,233	292,3452	91,61035405	-6,403354	41,00294	1,833521	3,361798
14,62298	7076,679	321,6864	95,31637314	-11,19337	125,2916	5,53954	30,6865
16,03202	9749,785	395,359	97,01378646	1,727214	2,983267	7,236953	52,37349
16,73628	10527,38	419,7489	97,83420289	4,768797	22,74143	8,05737	64,9212
17,6484	9031,461	399,2378	98,87151103	-3,837511	14,72649	9,094678	82,71316
18,22436	9928,329	425,3674	99,51275606	0,128244	0,016447	9,735923	94,78819
31,3824	23397,68	856,8987	112,0830447	40,87996	1671,171	22,30621	497,5671
34,08224	22692,71	879,4422	114,3085422	36,33246	1320,047	24,53171	601,8047
72,67563	17069,42	1113,791	139,647151	-8,997151	80,94873	49,87032	2487,049
169,2601	27570,61	2160,232	181,9410329	-15,89703	252,7157	92,1642	8494,24