УДК 519.25

ИССЛЕДОВАНИЕ ВЛИЯНИЯ РАЗЛИЧНЫХ ФАКТОРОВ НА ПРОДАЖИ ТАБАЧНОЙ ПРОДУКЦИИ В ГОРОДАХ КЕМЕРОВСКОЙ ОБЛАСТИ

Мысякина А.А., студент гр. СПмоз-221, I курс
Научный руководитель: Ермакова И. А., д.т.н, профессор
Кузбасский государственный технический университет имени Т.Ф. Горбачева
г. Кемерово

Табачная продукция занимает значительную часть мировой экономики. Российский табачный рынок – четвертый в мире по объемам продукции и является одним из крупнейших для производителей табака. Проблема потребления табака является чрезвычайно важной. И за последние годы в стране на государственном уровне было предпринято не мало мер, направленных на снижение спроса на табачные изделия. Так же активно ведется пропаганда здорового образа жизни. В 2022 году курение остается вредной привычкой миллионов россиян, хотя наметилась тенденция к сокращению числа потребителей никотиносодержащей продукции.

В своей работе, я бы хотела остановиться на определении факторов, которые оказывают наибольшее влияние на объемы продаж табачных изделий.

Данная работа будет выполнена в рамках одного из субъектов РФ – Кемеровской области. Исходные данные были собраны по 26 населенным пунктам, таким как: г.Анжеро-Судженск, г.Белово, г.Березовский, пгт.Грамотеино, г.Гурьевск, г.Калтан, г.Кемерово, г.Киселевск, пгт.Краснобродский, г.Ленинск-Кузнецкий, г.Мариинск, г.Междуреченск, г.Мыски, г.Новокузнецк, г.Осинники, г.Полысаево, г.Прокопьевск, пгт.Промышленная, г.Тайга, г.Таштагол, пгт.Тисуль, г.Топки, пгт.Тяжинский, г.Юрга, пгт.Яшкино, пгт.Яя.

Целью данной работы, является выявление факторов, влияющих на объемы продаж табачной продукции в Кемеровской области.

Данное исследование будем проводить с помощью регрессионного анализа данных, который будет заключаться в поиске уравнения регрессии, проверки значимости самого уравнения, и его коэффициентов. Тем самым определим и проанализируем значимость факторов, определяющих объемы продаж.

В качестве исходных показателей для проведения анализа будут использованы статистические данные на 01.11.2022 года, полученные из открытого источника, а именно с сайта https://bde*X*.ru , такие как:

1. численность населения, чел. (X1);
2. количество мужчин (X2);

3. количество женщин (X3);
4. количество постоянно занятого населения, чел. (X4);
5. количество людей с высшим образованием (X5);
6. количество людей со средне профессиональным образованием (X6);
7. количество людей с образованием 11 классов (X7);
8. средняя зарплата, руб. (X8);
9. средняя стоимость пачки сигарет, руб. (X9);

и показатели, актуальные на 01.11.2022 года, предоставленные для исследования одной из международных табачных компаний:

10. количество точек по продаже сигарет, шт. (X10);
11. среднее количество наименований сигарет в точке, шт. (X11);
12. продажи, шт. пачек (Y).

Все вышеперечисленные показатели приведены в таблице 1. Следует отметить, что данные в колонке «Продажи, шт. пачек» это количество проданных пачек сигарет только за последнюю неделю октября 2022 года.

Таблица 1. Исходные данные для анализа

Название городов	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	Y
Анжеро-Судженск	68116	29 549	38 567	40 597	12 738	26 565	11 375	51000	130	215	100	127 845
Белово	71240	30904	40336	42459	13 322	27 784	11 897	51000	130	159	105	113341
Березовский	45598	19 780	25 818	27 176	8 527	17 783	7 615	46100	107	98	98	77 359
Грамотеино	12172	5 280	6 892	7 255	2 276	4 747	2 033	35400	107	34	93	22 919
Гурьевск	22375	9 706	12 669	13 336	4 184	8 726	3 737	45100	107	67	103	49 320
Калтан	20464	8 877	11 587	12 197	3 827	7 981	3 417	45100	107	45	93	36 374
Кемерово	556382	241 359	315 023	331 604	104 043	216 989	92 916	60700	149	1273	99	982 913
Киселевск	86573	37 555	49 018	51 598	16 189	33 763	14 458	51000	130	254	97	150 385
Краснобродский	11419	4 954	6 465	6 806	2 135	4 453	1 907	44600	107	26	91	20 310
Ленинск-Кузнецкий	94398	40 950	53 448	56 261	17 652	36 815	15 764	51000	130	258	95	255 389
Мариинск	37912	16 446	21 466	22 596	7 090	14 786	6 331	45600	107	112	107	78 670
Междуреченск	96299	41 775	54 524	57 394	18 008	37 557	16 082	51000	130	201	101	173 909
Мыски	40787	17 693	23 094	24 309	7 627	15 907	6 811	46100	107	107	109	82 910
Новокузнецк	549403	238 331	311 072	327 444	102 738	214 267	91 750	60700	149	1147	106	903 324
Осинники	41673	18 078	23 595	24 837	7 793	16 252	6 959	46100	107	97	99	75 059
Полысаево	25867	11 221	14 646	15 417	4 837	10 088	4 320	45100	107	73	94	60 707
Прокопьевск	190334	82 567	107 767	113 439	35 592	74 230	31 786	53400	130	450	103	306 838
Промышленная	17345	7 524	9 821	10 338	3 244	6 765	2 897	44600	107	60	107	26 832
Тайга	23108	10 024	13 084	13 772	4 321	9 012	3 859	45100	107	55	95	33 602
Таштагол	22861	9 917	12 944	13 625	4 275	8 916	3 818	45100	107	68	112	51 766
Тисуль	7361	3 193	4 168	4 387	1 377	2 871	1 229	44200	107	27	114	16 138
Топки	27779	12 051	15 728	16 556	5 198	10 834	4 639	45100	107	72	96	39 950
Тяжинский	9687	4 202	5 485	5 773	1 811	3 778	1 618	44200	107	31	116	15 382
Юрга	80840	35 068	45 772	48 181	15 117	31 528	13 500	51000	130	175	87	111 451
Яшкино	13432	5 827	7 605	8 005	2 512	5 238	2 243	44600	107	35	93	20 163
Яя	10305	4 470	5 835	6 142	1 927	4 019	1 721	44600	107	46	88	23 170

Воспользуемся инструментом «Анализ данных/Регрессия» в EXCEL.

На рисунке 1 представлен вывод итогов.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,99797
R-квадрат	0,99594
Нормированный R-квадрат	0,92657
Стандартная ошибка	20181,4
Наблюдения	26

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	11	1498590828259,61	136235529841,78	367,944	0,0000000000000030
Остаток	15	6109312968	407287531,20		
Итого	26	1504700141227,54			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	143720,3	113268,6034	1,268844615	0,2238335	-97706,05576	385146,5707
Переменная X 1	18375,38	11290,00036	1,627580115	0,1244319	-5688,686051	42439,44624
Переменная X 2	0	0	65535	#ЧИСЛО!	0	0
Переменная X 3	-21488,2	16951,16542	-1,267656111	#ЧИСЛО!	-57618,80227	14642,30539
Переменная X 4	-11550,5	16112,40724	-0,71686843	0,4844689	-45893,25917	22792,30701
Переменная X 5	-12537,4	7149,018029	-1,753719486	0,0998818	-27775,14345	2700,399003
Переменная X 6	17955,16	18335,09394	0,979278054	0,3429738	-21125,17253	57035,48276
Переменная X 7	-23846,6	15575,27596	-1,531054914	0,1465709	-57044,51765	9351,312078
Переменная X 8	-1,43983	2,312441718	-0,622646918	0,5428672	-6,368687558	3,48901814
Переменная X 9	-466,779	835,3904383	-0,558755149	0,5845746	-2247,371279	1313,813861
Переменная X 10	674,3828	198,7605439	3,392940895	0,0040155	250,7347067	1098,030848
Переменная X 11	-291,666	660,4789313	-0,441598296	0,6650795	-1699,443888	1116,111148

Рис.1. Вывод итогов при использовании инструмента «Регрессия»

В таблице показаны множественный коэффициент корреляции R и индекс детерминации R-квадрат (R^2) , которые оценивают, на сколько близко находятся наблюдаемые значения *Y* от расчетных [1]. В нашем случае, R^2=0,99594, близок к единице, и можно говорить о влиянии факторов *X* на *Y*.

На следующем этапе исследования необходимо проверить значимость уравнения регрессии и коэффициентов при *X*. Это можно сделать, проанализировав результаты, полученные в таблице «Дисперсионный анализ» (см. рис. 1).

Для проверки значимости уравнения необходимо сравнить показатель «Значимость *F*», который равен 0,0000000000000030, с заданной вероятностью ошибки α=0,05 (по умолчанию в программе уже задан уровень надежности равный 95%, что соответствует уровню значимости (вероятности ошибки) α=0,05). Если «Значимость *F*» меньше заданной вероятности ошибки, то уравнение значимо, если больше, то уравнение не значимо. Поскольку в нашем случае 0,0000000000000030<0,05, можно сделать вывод что уравнение в целом значимо.

Далее следует проверить значимость коэффициентов при факторах *X*. Если коэффициент уравнения значим, то его следует оставить в уравнении

(если не значим - исключить) [1]. Обратим внимание на столбец «*P*-Значение», сравниваем полученные вероятности с заданной вероятностью ошибки α=0,05. Если показатель «*P*-Значение» меньше заданной вероятности ошибки, то коэффициент значим, если больше, то коэффициент не значим, и его не следует включать в уравнение. В нашем случае только одно значение *X*10, равное 0,0040155<0,05, из чего можно сделать вывод, что только один фактор «Количество точек по продаже сигарет, шт.» значим и оказывает влияние на значение Y (продажи, шт. пачек). Но это лишь на первый взгляд, и в случае, когда в качестве входного интервала X мы выбираем все значения из столбцов *X*1 ÷ *X*11.

После этого было бы рациональным рассмотреть парную линейную регрессию с наиболее значимым фактором. Но для начала попробуем разобраться почему согласно таблице итогов (см. рис. 1), остальные факторы не оказывают существенного влияния на объемы продаж табачной продукции в Кемеровской.

Мной были рассчитаны процентные соотношения факторов *X*4 (количества постоянно занятого населения), *X*5 (количества людей с высшим образованием), *X*6 (количества людей со средне профессиональным образованием), *X*7 (количества людей с образованием 11 классов) к фактору *X*1 (количеству проживающих людей в населенных пунктах). Они оказались абсолютно одинаковы во всех рассматриваемых населенных пунктах Кемеровской области. Отношение *X*4 к *X*1 во всех городах составило 60%, *X*5 к *X*1 – 19%, *X*6 к *X*1 – 39%, *X*7 к *X*1 - 17%. На основании полученных результатов, можно сделать вывод, что в данном контексте задачи факторы *X*4, *X*5, *X*6, *X*7 не влияют на объем продаж табачной продукции (*Y*), и поэтому нам следовало бы исключить их из нашего исследования.

Далее были проверены на мультиколлениарность такие факторы как *X*8 (средняя зарплата, руб.) и *X*9 (средняя стоимость пачки сигарет, руб.). Для этого в пакете EXCEL была построена точечная диаграмма рассеивания с «линейной» (поскольку мультколлениарность – это наличие именно линейной зависимости между факторами регрессионной модели) линией тренда и рассчитана величина R^2.

R^2=0,8397, соответственно |r|=0,9164, это значит что между факторами X8 и X9 существует взаимосвязь. Оба этих фактора включать в уравнение регрессии нельзя, нужно оставить только один и необходимо определить какой из этих факторов оказывает наибольшее влияние на *Y*. Для этого нам необходимо найти уравнение множественной линейной регрессии для факторов *X*8, *X*9 и *Y*, найти стандартизованные коэффициенты регрессии и сравнить их по модулю. Воспользовавшись инструментом «Анализ данных/Регрессия» в EXCEL, мы получили:

1. Коэффициент при *X*8 равен 27,66, его «*P*-значение» равно 0,032614196<0,05;
2. Коэффициент при *X*9 равен 4926,16, его «*P*-значение» равно 0,294874396>0,05.

Также для факторов X8 и X9 мною были найдены стандартизованные коэффициенты регрессии

$$\beta_8 = b_8 \frac{S_{x8}}{S_y} = 27{,}66 \frac{5197{,}74}{240568{,}25} = 0{,}5977,$$

$$\beta_9 = b_9 \frac{S_{x9}}{S_y} = 4926{,}16 \frac{13{,}76}{240568{,}25} = 0{,}2817,$$

и выполнено их сравнение по модулю $|\beta_8|>|\beta_9|$.

Из расчетов, выполненных для факторов *X*8 и *X*9, можно говорить о том, что только фактор *X*8 (средняя зарплата, руб.) значим и оказывает наибольшее влияние на *Y*, следовательно, фактор *X*9 также исключаем из нашего исследования.

Для каждого из оставшихся факторов *X*1 (численность населения, чел.), *X*2 (количество мужчин), *X*3 (количество женщин), *X*8 (средняя зарплата, руб.), *X*10 (количество точек по продаже сигарет, шт.), *X*11 (среднее количество наименований сигарет в точке, шт.) по отношению к фактору *Y* (продажи, шт. пачек) были построены диаграммы рассеивания, найдены линии тренда парной регрессии и рассчитаны R^2. Также мною были вычислены коэффициенты корреляции, которые отражены в таблице 2.

Исходя из полученных данных получаем, что между *Y* и *X*11 (среднее количество наименований сигарет в точке, шт.), связи не существует, поскольку коэффициент корреляции больше близок к нулю. Поэтому фактор *X*11 также исключаем из нашего исследования.

Остаются пять факторов, которые имеют взаимосвязь с фактором *Y*. Это факторы *X*1, *X*2, *X*3, *X*8, *X*10.

Таблица 2. Коэффициенты корреляции для рассматриваемых факторов

	Коэффициент корреляции, |r|
Зависимость Y от X1	0,9956
Зависимость Y от X2	0,9957
Зависимость Y от X3	0,9957
Зависимость Y от X8	0,8559
Зависимость Y от X10	0,9967
Зависимость Y от X11	0,0748

В таблице на рис. 1 у факторов *X*2 (количества мужчин), *X*3 (количества женщин) не показана вероятность ошибки. Процентное соотношение этих факторов к фактору *X*1 во всех городах одинаково и составило *X*2 к *X*1 – 43%, *X*3 к *X*1 – 57%. Но, рассмотрев по отдельности факторы *X*2 и *X*3 к *Y*, и использовав инструмент «Анализ данных/Регрессия» в EXCEL, мы видим, что по отдельности данные факторы значимы, причем в равной степени и также оказывают влияние на количество проданных пачек сигарет. Поэтому данные факторы нельзя исключить. Проверив их на мультиколлениарность, мы

увидели, что они тесно связаны между собой, поскольку имеют коэффициент детерминации R^2=1. Найдя стандартизованные коэффициенты регрессии и сравнив их по модулю ($\beta_2 = 2530{,}032, \beta_3 = 2529{,}04$), пришли к выводу, что фактор *X*2 (количество мужчин) оказывает б*о*льшее влияние на *Y* чем *X*3 (количество женщин).

Далее мной была проведена проверка на мультиколлениарность факторов *X*1, *X*8 и *X*10, по результатам которой стало очевидно, что между этими тремя факторами существует взаимосвязь, но наибольшее влияние на продажи сигарет оказывает фактор *X*10.

Воспользуемся инструментом «Анализ данных/Регрессия» в E*X*CEL, мы проанализировали факторы *X*2(количество мужчин) и X10 (количество точек по продаже сигарет) к *Y* (продажи, шт. пачек). Из полученных таблиц следует, что уравнение регрессии значимо, так как «Значимость F»=0,00000000000000000000000000186. Поскольку *P*-Значение переменной *X*10=0,0028, а переменной *X*2=0,1071, то фактор *X*2 нельзя включить в уравнение регрессии. Самую высокую степень взаимосвязи с Y (продажи, шт. пачек), судя по таблице 2, имеет фактор *X*10 (количество точек по продаже сигарет), значение коэффициента корреляции составляет 0,9967. Отсюда следует, что только один фактор *X*10 наиболее значим и оказывает существенное влияние на продажи сигарет в Кемеровской области.

Рассмотрим парную линейную регрессию зависимости фактора *Y* от *X*10. Для этого построим точечную диаграмму рассеивания и линию регрессии (рис. 2). Найдем уравнение линейной парной регрессии, проверим его значимость по критерию Фишера.

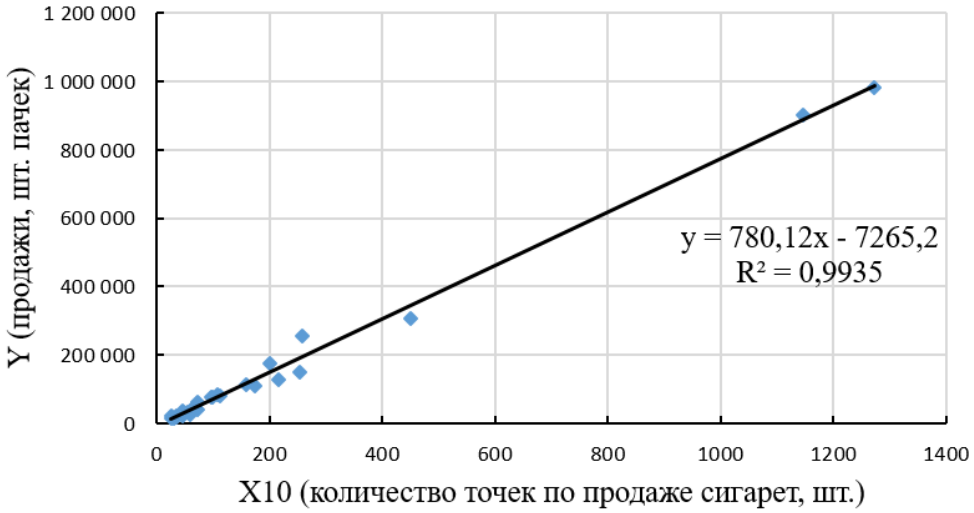

Рис. 2. Диаграмма зависимости Y (продажи, шт. пачек) от X10 (количество точек по продаже сигарет, шт.)

Уравнение парной линейной регрессии имеет вид y=780,12x-7265,2 и величина R^2=0,9935. Проверим значимость уравнения с помощью критерия Фишера.

Для этого необходимо рассчитать наблюдаемое значение критерия:

$$F_{\text{набл}} = \frac{R^2}{1-R^2} \cdot (n - m - 1),$$

где n – число наблюдений (в нашем случае n=26), m – число параметров при X, в парной регрессии m=1 [1].

$$F_{\text{набл}} = \frac{0{,}9935}{1-0{,}9935} \cdot 24 = 3668{,}31.$$

Полученное значение сравниваем с $F_{\text{крит}}$, которое находим по таблице значений F-критерия Фишера при уровне значимости α=0,05 [1]. Для этого необходимо посчитать число степеней свободы k_1 и k_2.

k_1=m=1, k_2=n-m-1=26-1-1=24, $F_{\text{крит}} = 4{,}26$.

Если $F_{\text{набл}} > F_{\text{крит}}$, то связь между X и Y существует, и уравнение регрессии значимо. Если $F_{\text{набл}} < F_{\text{крит}}$, то связи между X и Y не существует, уравнение регрессии не значимо [1].

В нашем случае $F_{\text{набл}} > F_{\text{крит}}$, что говорит о том что связь между X10 (количество точек по продаже сигарет, шт.) и Y (продажи, шт. пачек) существует, и уравнение y=780,12x-7265,2 значимо.

По результатам исследования можно подвести следующий итог. Продажи сигарет в Кузбассе безусловно зависят от численности населения, и в большей мере от количества мужчин, проживающих в городах.

Но всё же из всех проанализированных факторов, фактор X10 (количество точек по продаже сигарет, шт.) имеет самую высокую степень взаимосвязи, является значимым и оказывает наибольшее влияние на значение Y (продажи, шт. пачек). При увеличении в городе количества точек по продаже табачной продукции на 1 единицу, количество проданных пачек сигарет увеличивается на 780 штук. И так как связь между этими величинами прямая, то чем больше в населенном пункте точек по продаже сигарет, тем больше объем продаж табачной продукции. Таким образом, объём продаж в Кемеровской области, в большей степени, напрямую зависит только от количества точек по продаже сигарет.

Однако следует отметить, что на основании проведенного эксперимента нельзя утверждать, что этот же фактор будет являться единственным значимым при аналогичном исследовании объема продаж табачной продукции в другой области России или в Российской Федерации в целом. Значимость могут приобрести другие факторы, которые при данном исследовании не оказали никакого влияния на объем продаж.

Также в ходе проведенного исследования мною был выполнен расчет количества проданных пачек на 1000 человек, проживающих в населенных пунктах Кемеровской области, и рассчитаны удельные веса таких показателей как численность населения, количество точек по продаже сигарет и продажи сигарет за неделю. Удельный вес – это статистический показатель, который показывает соотношение частного к совокупности, значение части от целого. Результаты данных расчетов представлены в таблице 4.

По количеству точек по продаже сигарет места распределились следующим образом:

1. г. Кемерово
2. г. Новокузнецк
3. г. Прокопьевск

Таблица 4. Дополнительные расчеты

Название городов	Удельный вес населения, %	Удельный вес количества точек по продаже сигарет, %	Удельный вес продаж сигарет, %	Количество точек на 1000 человек населения нас. пункта, шт.	Количество проданных пачек на 1000 человек, в неделю
Анжеро-Судженск	3,12	4,15	3,32	3	1 877
Белово	3,26	3,07	2,94	2	1 591
Березовский	2,09	1,89	2,01	2	1 697
Грамотеино	0,56	0,66	0,59	3	1 883
Гурьевск	1,02	1,29	1,28	3	2 204
Калтан	0,94	0,87	0,94	2	1 777
Кемерово	25,48	24,55	25,49	2	1 767
Киселевск	3,96	4,90	3,90	3	1 737
Краснобродский	0,52	0,50	0,53	2	1 779
Ленинск-Кузнецкий	4,32	4,98	6,62	3	2 705
Мариинск	1,74	2,16	2,04	3	2 075
Междуреченск	4,41	3,88	4,51	2	1 806
Мыски	1,87	2,06	2,15	3	2 033
Новокузнецк	25,16	22,12	23,43	2	1 644
Осинники	1,91	1,87	1,95	2	1 801
Полысаево	1,18	1,41	1,57	3	2 347
Прокопьевск	8,72	8,68	7,96	2	1 612
Промышленная	0,79	1,16	0,70	3	1 547
Тайга	1,06	1,06	0,87	2	1 454
Таштагол	1,05	1,31	1,34	3	2 264
Тисуль	0,34	0,50	0,42	4	2 192
Топки	1,27	1,39	1,03	3	1 438
Тяжинский	0,44	0,60	0,40	3	1 588
Юрга	3,70	3,38	2,89	2	1 379
Яшкино	0,62	0,68	0,52	3	1 501
Яя	0,47	0,88	0,60	4	2 248
ИТОГО:	100,00	100,00	100,00		47 946

Если провести расчет количества торговых точек по продаже сигарет на 1000 человек населения, то лидерами являются поселки городского типа Тисуль и Яя, этот показатель равен 4.

Но больше всего пачек сигарет на 1000 человек, было продано в городе Ленинск-Кузнецком, и составило 2705 штук за последнюю неделю октября 2022 года. Таким образом, можно сделать вывод, что город Ленинск-Кузнецкий является самым курящим городом Кемеровской области.

Список литературы:

1. Ермакова, И.А. Прикладная математика [Электронный ресурс] : методические материалы для обучающихся направления подготовки 08.04.01 "Строительство", профили: "Промышленное и гражданское строительство",

"Автомобильные дороги", всех форм обучения / И. А. Ермакова, В. А. Гоголин; Кузбасский государственный технический университет им. Т. Ф. Горбачева, Кафедра математики. – Кемерово : КузГТУ, 2019. – 25 с. – URL: http://library.kuzstu.ru/meto.php?n=9834 (Дата обращения 08.12.2022). – Текст электронный.